ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2026
D. Levin and M. Chu
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 8

GNOCHI: Generative Neural mOdel for Close Human-Human Interactions

Gonzalo Gómez-Nogales[†1] Marc Comino-Trinidad[†1] Andrés Casado-Elvira[1] Dan Casas[‡1]

[1]Universidad Rey Juan Carlos, Madrid, Spain.

Figure 1: *Given an input 3D pose (i.e., the conditioning pose, in purple), our generative model infers a 3D pose of a human in close interaction (i.e., the reaction pose, in green). This enables the conditional generation of 3D humans in close interaction, which can be used fine-grain control signal for image generative methods (right).*

Abstract

Creating realistic 3D human-human interactions in virtual environments is challenging due to the high degrees of freedom in human body and the need for physically accurate poses that do not collide with each other. Traditional methods for human-human interaction are based on motion tracking or 3D body reconstruction, but lack generative capabilities. Recent generative methods enable the synthesis of individual or interacting motions via text or image input, but generally fall short in modeling close interactions. This paper introduces a novel generative model for close 3D human-human interactions using a conditional variational autoencoder (cVAE), which generates poses for one human conditioned on the pose of another, allowing for controlled and diverse interaction synthesis. To train our model, we address two underlying long-standing challenges in the field of human-human interaction: data scarcity, for which we propose an automated supervised data augmentation strategy that generates synthetic yet realistic interaction poses; and collision awareness in generative approaches, for which we propose a self-supervised loss based on a collision resolution technique using volumetric proxies to ensure physically correct interactions. We extensively evaluate the capabilities of our model, and demonstrate a wide variety of plausible and physically correct interactions, not possible to generate with current state-of-the-art methods.

CCS Concepts

• ***Computing methodologies*** → ***Physical simulation; Collision detection; Mesh models;***

1. Introduction

Creating life-like virtual 3D scenes is central to Computer Vision, Graphics, and VR, impacting dataset generation, human tracking, 3D scene understanding, and character animation. However, modeling and reconstructing how humans interact in everyday 3D scenes is a highly complex task due to the large number of degrees of freedom involved. Additionally, as humans, we are very sensitive to non-physically-correct virtual scene arrangements (*e.g.*, interpenetrations, mesh collisions, or impossible configurations), hence,

† Equal contribution
‡ Work done prior to joining Amazon

errors in modeling human interactions automatically produce significant visual discomfort.

Existing research on 3D interaction generally falls into three categories: human-scene methods [JZL*24, HGT*21, GCM*24, ZZW*23,ZWZ*22,HCV*21,MCZ*25], for large-scale navigation; human-object methods [JLC*23, XBPM22, CKA*22, HVT*19, YGKT24, TCBT22], for manipulation and grasping; and human-human models [MYP*24, YPMK23, FZO*20, FZS*21, MMR*24]. We focus on the latter, specifically addressing the synthesis of natural, close-contact interactions between two highly articulated bodies.

Unfortunately, most existing human-human interaction methods prioritize tracking or reconstruction [FZO*20, FZS*21, YPMK23, GBAPM22, YGK*23, SBL*21, SLB*22]. These approaches yield complex optimization or learning-based strategies for pose estimation, yet they lack generative capabilities and cannot sample from a learned distribution to synthesize unseen human-human interactions. Recent works [TF23, CTO*23, FTC*25, SDH*25, JGCL25, LZL*24] proposed generative models for human interaction, usually from text prompts, but they do not tackle the specific case of close interactions, nor do they incorporate robust collision awareness in their formulations. BUDDI [MYP*24] is the notable exception, and it proposes a generative model capable of synthesizing realistic poses of two humans in close interaction. Despite that BUDDI models both poses in a joint latent space, we show that conditional sampling can be achieved by incorporating the conditioning pose information throughout the diffusion process, similar to the in-painting strategies for conditional image synthesis. However, our experiments show that this often compounds errors in close interactions, leading to physically non-plausible root placements or orientations, or unresolved colliding 3D meshes.

To mitigate this limitation, we propose a novel generative model for close 3D human-human interaction that is explicitly conditioned on one pose. Implemented as a conditional variational autoencoder (cVAE) , our model uses the SMPL [LMR*15] parameters of a conditioning human to synthesize the body parameters of a reactive agent. Since our model is probabilistic, we can generate many different poses given the same conditioning pose, which further increases the applicability of the method. Additionally, we demonstrate that our model yields a continuous latent space that allows to seamlessly interpolate between reactive poses while keeping the same conditioning pose.

Manual rigging of two-person interactions is challenging; adjusting a single joint in a conditioning pose typically requires the artist to manually re-balance dozens of degrees of freedom in the reacting pose to maintain contact and avoid penetration. Our model acts as a 3D pose copilot, offering a variety of semantically plausible reactive candidates that serve as a high-quality starting point for artists, effectively shortening the manual iterative loop.

To train our model, we tackle two long-standing problems in data-driven human-human interaction: data scarcity and collision awareness in data-driven models. Regarding the first problem, despite the few recently introduced datasets for this task [FZO*20, YGK*23, XLY*24, ZSZ*21, GBAPM22], the coverage and diversity of *close interactions* are still limited because these datasets rely on cumbersome motion capture, manual annotations, or expensive post-processing pipelines that do not scale well. To mitigate this, we introduce a novel automated strategy to, for the first time, apply data augmentation techniques to 3D human-human interaction data, enabling to create synthetic and physically-plausible poses without requiring any manual work or new recordings. Our key contribution is an individual-level stochastic process to manipulate the degrees of freedom of captured human-human motions while guaranteeing the correctness of the resulting interaction. We achieve this by combining a learned human pose subspace [PCG*19] that guarantees natural (yet unseen) individual poses, with an efficient collision resolve strategy that guarantees physically correct human-human poses in interactions without mesh intersections at the vertex level.

Our second contribution addresses a fundamental limitation in data-driven models: while our augmented dataset provides potentially unlimited collision-free human interaction samples, trained data-driven models (*e.g.*, cVAE) cannot inherently guarantee collision-free states for unseen samples. We tackle this limitation by incorporating an additional decoder trained with a collision-aware novel self-supervised loss based on highly-efficient distance computation across volumetric proxies, ensuring physically plausible interactions during generation.

We demonstrate that our generative approach can synthesize pose-conditioned human-human interactions for a wide variety of scenarios, including sports, dancing, fighting, and social communications. Additionally, as a forward generative model, our approach can be used to synthesize reactive interaction partners or refine tracking results from current human mesh recovery methods [SLB*22, MYP*24], providing a richer variety of plausible responses and enhanced controllability of the final output.

2. Related Work

2.1. Datasets

Obtaining accurate human-human body data is inherently challenging. Various methods exist to transform real-world data into parameters suitable for computational processing [FZO*20, YGK*23, XLY*24, GBAPM22, ZSZ*21]. Fieraru *et al.* [FZO*20] use 2D annotated data to predict 3D contact, resulting in two datasets: CHI3D based on captured data, and FlickrCI3D based on Flickr images. Using an alternating optimization scheme, Yin *et al.* [YGK*23] segment the 3D surface of two avatars and fit SMPL [LMR*15] model parameters to them. Inter-X [XLY*24] is the largest labeled human interaction dataset, featuring around 11,000 sequences and over 8.1 million frames. It provides detailed descriptions, action categories, and annotations for interaction order. Zheng *et al.* [ZSZ*21] addresses self-occlusions through multiview capture. Their MultiHuman dataset consists of 150 static scenes with different levels of occlusions and ground truth 3D human models. To predict subsequent dance moves, Guo *et al.* [GBAPM22] created a dataset of dancers in extreme poses, capturing dynamic and complex movements for motion prediction tasks. In this work, we leverage the Hi4D [YGK*23] dataset which, despite the impressive quality of the reconstructed human interactions, still contains a limited amount of in-contact interactions. We apply our novel

pose augmentation technique to the subset of in-contact interactions from Hi4D, generating 10× samples that we use to train our model.

2.2. Multi-Character Animation

Generating human-human interaction is a classic problem in computer animation. Early works use optimization-based frameworks [LHP06] and spatial relationship descriptors [HKT10] to adapt captured motions to new interactions. Later, patch-based methods [KHHL12] tiled or concatenated pre-recorded interactions to generate variety, while data-driven generate-and-rank systems [WLO*14] composed and ranked scenes from motion-capture libraries. While effective, these **data-driven** approaches required dense motion libraries.

More recently, there is increasing use of physics-based simulation [ZGY*23], alongside strategic planning methods that explicitly model competitive and collaborative goals [SKY07, SKY12]. Other recent works aim at learning time-dependent dynamic human interactions [GZD*25, LGZW25]. Instead, we focus our efforts on learning a continuous space generative manifold of static interaction, which is a fundamental problem in keyframe animation.

2.3. Reconstruction of People in Close Interaction

Many previous works [MYP*24, MOT*21, YPMK23, SYZ*23, JGK*24, ZSZ*21] aim to accurately reconstruct interacting humans. SLAHMR [YPMK23] reconstructs sequences involving multiple people by accurately decoupling human and camera motion. The underlying model uses an optimization method to achieve this decoupling, allowing for the reconstruction of complex scenes. However, while the method is effective in handling multiple people, it falls short in accurately recovering close interactions between them, highlighting an area for potential improvement in future research. Shuai et al. [SYZ*23] tackle the challenge of reconstructing close human interactions from multiple views. Their approach integrates a learning-based pose estimation component that uses multi-view 2D keypoint heatmaps as input and reconstructs the pose of each individual using a 3D conditional volumetric network. This method significantly improves the accuracy of pose reconstruction in crowded scenes. Other works [MOT*21, MSB*22] address the challenge of modeling single avatars with accurate self-contact using collision-specific terms. Rather than focusing on interactions between multiple avatars, these approaches emphasize the importance of self-contact, ensuring that the reconstructed avatars maintain realistic and physically plausible poses.

Closest to our work, BUDDI [MYP*24] focuses on modeling two interactive avatars using a diffusion model, which learns the joint distribution over the poses of two people in close social interaction. The model is particularly notable for its ability to generate pairs of 3D avatars that exhibit realistic and close interactions, making it highly relevant for applications requiring detailed social dynamics. Our objective, however, is to explicitly generate an interacting avatar given the pose of another one. Despite that BUDDI can be adapted to pose-conditional sampling, our method yields more accurate contact for close interactions.

2.4. Collision Aware Reconstruction

Some works [FZS*21, UPP*24, MYP*24] put a special effort into avoiding collisions and mesh intersections between avatars. MultiPhys [UPP*24] integrates a physics simulator in an autoregressive manner to refine kinematic estimates and ensure physical compliance. The simulator captures coherent spatial placement between individuals and eliminates penetration issues. The pipeline involves feeding motion estimated by a kinematic-based method into the physics simulator, which then corrects any collisions or penetrations by adjusting the avatars' positions and poses. BUDDI [MYP*24] employs a collision-checking strategy that primarily relies on a coarse approximation of the avatar mesh. This simplified representation facilitates efficient collision detection and resolution by reducing computational complexity. The model uses ground-truth contact annotations to fit SMPL-X to images, ensuring that avatars maintain realistic social distances and avoid collisions. Similarly, REMIPS [FZS*21] also uses a coarse mesh approximation by applying decimation operators to detect and solve self- and interpenetration-collisions. This model incorporates self-contact and interaction-contact losses directly into the learning process, ensuring that the reconstructed avatars do not intersect with each other or the environment. The use of self-supervised losses allows the model to generalize well to in-the-wild scenarios, maintaining physical consistency in multi-person 3D reconstructions. Instead of using a coarse version of SMPL, we resolve collisions at the vertex level to create a large yet vertex-accurate dataset. At train time, we use highly-efficient volumetric capsules to compute potential intersections.

3. Close Contact Dataset

Capturing and reconstructing 3D humans in close interaction is a tedious task due to the unavoidable occlusions that prevent observing the full human surface, even in multi-camera environments. Recent methods [YGK*23] have proposed complex pipelines for accurately reconstructing close interactions in 3D; however, they are expensive and do not scale well.

In data-driven methods, to circumvent the scarcity of training data, data augmentation is a common strategy to increase the number of dataset samples, and it plays a fundamental part in many image-based machine learning methods [MBS*18]. However, augmenting 3D human-human pose data is challenging due to the many nuances that make human interaction realistic.

To this end, we propose a novel approach to build a pseudo-synthetic dataset of human-human interactions. Starting from a subset of the Hi4D [YGK*23] dataset, we first apply an individual-level data-augmentation strategy (Section 3.1), and then resolve vertex-level human-human collisions using a physics-based simulation strategy (Section 3.2). This enables us to obtain a dataset of unseen yet physically-correct *close* (*i.e.* in-contact) human-human interaction poses, which we later use to train a pose-conditioned generative model in Section 4.

3.1. Individual-level Pose Augmentation

We begin our dataset creation by selecting a subset of frames from the Hi4D [YGK*23] dataset that contain contact interactions,

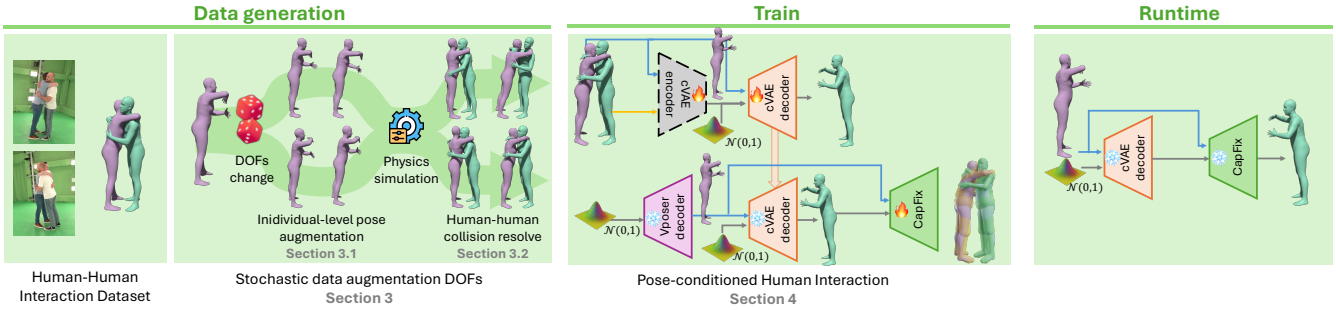

Figure 2: *From an existing dataset [YGK*23] (left), we apply individual-level noise at each body joint (Section 3.1), and jointly resolve the collisions using a physical simulator (Section 3.2) to create an augmented dataset. At train time, our model learns to infer a reacting avatar given a conditioning avatar (center top, Section 4.2). Then, a second model is trained to fix collisions between two interacting avatars (center bottom, Section 4.3). At runtime, we can sample a reacting collision-free avatar given a conditioning pose.*

namely poses where both interacting meshes are at less than 1 cm from each other. For each body pose of each selected frame, we first apply random noise to the rotation of specific body joints, and then project the resulting pose into a subspace of feasible poses using the VPoser [PCG*19] autoencoder, ensuring the final pose is anatomically plausible and introducing additional noise.

To generate random rotations, we employ the axis-angle representation for its intuitive and mathematically simple nature, the full details of which are provided in the Supplementary Material. We carefully define a specific rotation distribution for each joint. For example, knees only rotate around the horizontal axis, while for other joints, such as hips, the distribution is weighted for all three spatial axes. For each joint, we define a set of weights w_x, w_y, w_z and a maximum angle α_M. We then sample a vector

$$\overrightarrow{v_r} = \frac{w_x r_x + w_y r_y + w_z r_z}{\|w_x r_x + w_y r_y + w_z r_z\|}, \tag{1}$$

where $r_x, r_y, r_z \sim \mathcal{U}_{[-0.5,0.5]}$. Finally, the angle α is similarly obtained

$$\alpha = \alpha_M w_\alpha, \tag{2}$$

where $w_\alpha \sim \mathcal{U}_{[-1,1]}$. This symmetric uniform distribution effectively places the rest pose at the center of each joint's maximum sweep angle. The final random axis-angle is obtained by multiplying $\alpha \cdot \overrightarrow{v_r}$. To apply the noise to the joint, we convert both the joint and the random axis-angle rotation into matrix rotation representation and multiply them. Setting two weights w_i to zero means selecting only one axis for rotation. If one or none of the weights are zero, controlling the random rotation becomes more challenging. Because our perturbation is based strictly on these maximum sweep bounds, the naive random rotation can potentially result in an anatomically unfeasible configuration, such as joints bending backward. To explicitly address this lack of joint constraints, we pass the randomized pose through the VPoser autoencoder [PCG*19]. VPoser projects the perturbed pose back into a learned subspace of human poses, ensuring that the final output is always anatomically valid regardless of the initial random noise.

3.2. Human-Human Collision Resolve

Our individual-level pose augmentation described in Section 3.1 is a naive strategy that can lead to new, albeit physically incorrect, poses that may intersect with nearby individuals. To mitigate this, we employ a highly efficient strategy to accurately resolve collisions between interacting humans, which is crucial in building our large dataset.

Approximating humans with coarse volumetric proxies is a common technique to resolve collisions because it allows for fast inter-proxy distance computation. However, in the specific case of close human-human contact, this strategy produces unrealistic interactions due to the coarse approximation of the human surface. To circumvent this problem, we leverage a pre-existing solver that implements a method based on accurate Signed Distance Fields (SDFs) to resolve human-human intersections.

It is important to note that the initial SMPL fits of the Hi4D data are not inherently penetration-free. For each pair of humans interacting in the original dataset, the solver computes the SDF of each human in the scene using OpenVDB. This effectively creates two grids (one for each human) that can be efficiently used to query distances to human surfaces from any 3D world coordinate. The solver then checks the distance of all surface points of one human (i.e., all the vertices) against the other human in the frame, and vice versa. For each surface vertex that is *inside* the other body or its own body, a penalty force $E_{\text{col}}(\delta) = \frac{1}{2} k_\delta \delta^2$ is applied in the direction of the closest surface point normal. Here, δ represents the distance from the vertex position to the surface of the penetrated body, and this force is applied to the bones that influence the vertex.

We integrate the penalty equations using the popular optimization formulation of backward Euler [GSS*15, KMOW00]. To solve it, we use Newton's method with analytical computation of gradients and Hessians. Just one Newton iteration works well for our disentanglement simulations. Each Newton iteration yields a sparse linear system of size $6 \times 24 \times 2$, where 2 is the number of humans interacting, 24 is the number of bones per body, and 6 is the number of degrees of freedom per bone. We solve this linear system using the conjugate gradient method.

We present results of our collision resolution approach in Figure 3. Notably, our method effectively eliminates mesh interpenetrations while preserving close contact between the avatars.

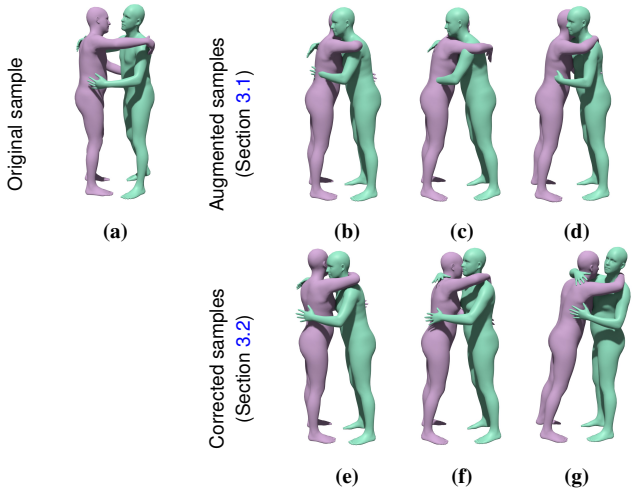

Figure 3: *Data augmentation steps. We add noise to the original poses (3a), yielding augmented poses (3b, 3c, and 3d) which may contain collisions. Our physics solver fixes these, yielding physically-correct interactions (3e, 3f, and 3g).*

3.3. Final Dataset Details

For each avatar pair in the Hi4D dataset, we generated 10 variations. Because the random noise introduced during pose augmentation can cause the avatars to drift apart, we filtered out any samples that no longer met our contact threshold, resulting in a final dataset comprising 55,897 examples. We later merged this dataset with the 5,744 poses from Hi4D interaction scans, yielding a full augmented dataset of 61,641 poses. To prevent any information leakage between sets, we implemented a sequence-heldout train-test split rather than a random per-sample partition. Augmented variants are strictly assigned to the same split as their base sequence. Detailed sample distributions for the training and testing sets are provided in Section 5.1.

4. Pose-Conditioned Human Interactions

Our goal is to model physically-correct and realistic 3D human poses in close interaction. Specifically, we address the problem of *pose-conditioned pose* inference, where we are given a pose (the conditioning pose) and aim to infer poses (the reaction poses) that realistically interact with the input.

To this end, we first introduce the representation used to describe poses (Section 4.1). We then present our compact yet expressive generative latent space tailored for humans in interaction (Section 4.2), which we train using a conditional Variational Autoencoder (cVAE) that leverages the pose-augmented human-human dataset introduced in Section 3. Next, we describe a self-supervised strategy (Section 4.3) to learn to resolve the residual human-human collisions in our latent space. At inference time (Section 4.4), our method enables the generation of human-human 3D pose interactions conditioned on the pose of one of the avatars. Figure 2 depicts the detailed architecture of this pipeline. The full architecture details of the model can be found in the Supplementary Material.

4.1. Pose Representation

We define the pose of an avatar X as

$$\mathbf{p}_X = \{\theta_X, \mathbf{r}_X, \mathbf{t}_X\}, \tag{3}$$

where θ_X are the joint rotations, $\mathbf{r}_X$ is the global rotation, and $\mathbf{t}_X$ is the global translation. We use the original SMPL [LMR*15] 23 joints but, to facilitate training, we leverage the continuous 6D representation for each joint rotation by Zhou *et al.* [ZBL*19]. Consequently, the joint rotations θ_X is a tensor in $\mathbb{R}^{23\times3\times2}$ and the global rotation matrix $\mathbf{r}_X$ is $\mathbb{R}^{3\times2}$. The global translation $\mathbf{t}_X$ is represented by a 3D position vector $\mathbf{t}_X$ in $\mathbb{R}^3$.

Each sample consists of two poses: a conditioning pose $\mathbf{p}_C$ and a reaction pose $\mathbf{p}_R$. We use pose coordinates relative to the conditioning avatar $\mathbf{p}_C$, therefore, translation $\mathbf{t}_C$ and the rotation $\mathbf{r}_C$ of avatar C are the origin of coordinates and a unit matrix, respectively, and $\mathbf{p}_C$ can be defined by the θ_C alone.

4.2. Latent Space for Human Interaction

To learn a generative model for human-human interaction, we use a conditional Variational Autoencoder (cVAE) [SYL15] consisting of an encoder $\mathcal{E}$ and decoder $\mathcal{D}$ implemented as multilayer perceptron (MLP) networks. We use a custom cVAE architecture and train the encoder-decoder network end to end. The encoder $\mathcal{E} : (\mathbb{R}^{23\times3\times2}, \mathbb{R}^{23\times3\times2}, \mathbb{R}^3, \mathbb{R}^{3\times2}) \rightarrow (\mathbb{R}^{128}, \mathbb{R}^{128})$ takes as input $\{\theta_C, \theta_R, \mathbf{t}_R, \mathbf{r}_R\}$ and predicts the latent variables that define the normal distribution (*i.e.*, mean μ and variance parameters σ^2) of the latent space. The decoder $\mathcal{D} : (\mathbb{R}^{23\times3\times2}, \mathbb{R}^{128}) \rightarrow (\mathbb{R}^{23\times3\times2}, \mathbb{R}^{3\times2}, \mathbb{R}^3)$ takes as input a concatenation of a sample $\mathbf{z} = \mu + \mathbf{z}_0 \cdot \sigma \quad | \quad \mathbf{z}_0 \sim \mathcal{N}(0,1)$ and the conditioning pose θ_C, and it decodes the reaction pose $\hat{\mathbf{p}}_R = \{\hat{\theta}_R, \hat{\mathbf{r}}_R, \hat{\mathbf{t}}_R\}$. As depicted in Figure 4, the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are implemented using shared and specialized components. For example, the decoder $\mathcal{D}_S : \mathbb{R}^{128} \rightarrow \mathbb{R}^{128}$ first converts our latent representation $\mathbf{z}$ into a shared decoded pose representation. Then, we extract the joint rotations $\hat{\theta}_R$, global translation $\hat{\mathbf{t}}_R$, and global rotation $\hat{\mathbf{r}}_R$ using task-specific decoder networks $\mathcal{D}_\theta$, $\mathcal{D}_\mathbf{t}$, and $\mathcal{D}_\mathbf{r}$ as follows:

$$\hat{\theta}_R = \mathcal{D}_\theta(\mathcal{D}_S(\{\theta_C, \mathbf{z}\})) \tag{4}$$

$$\hat{\mathbf{r}}_R = \mathcal{D}_\mathbf{r}(\mathcal{D}_S(\{\theta_C, \mathbf{z}\})) \tag{5}$$

$$\hat{\mathbf{t}}_R = \mathcal{D}_\mathbf{t}(\mathcal{D}_S(\{\theta_C, \mathbf{z}\})) \tag{6}$$

where $\{\theta_C, \mathbf{z}\} \in \mathbb{R}^{266}$ denotes the concatenation and flattening of the tensors.

We train our cVAE network end-to-end using the loss

$$\mathcal{L}_{\text{VAE}} = \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \tag{7}$$

where $\mathcal{L}_{\text{KL}}$ is the Kullback–Leibler divergence term that enforces the latent spaces to follow a Gaussian distribution, and $\mathcal{L}_{\text{rec}}$ is a body surface term defined as

$$\mathcal{L}_{\text{rec}} = \left\| f_{\text{SMPL}}(\mathbf{p}_R) - f_{\text{SMPL}}(\hat{\mathbf{p}}_R) \right\|_2^2 \tag{8}$$

that enforces the output reaction pose body parameters $\hat{\mathbf{p}}_R$ to generate SMPL body surface vertices close to ground truth vertices. We empirically set the loss weights to $\lambda_{\text{KL}} = 1 \times 10^{-3}$ and $\lambda_{\text{rec}} = 7.5$.

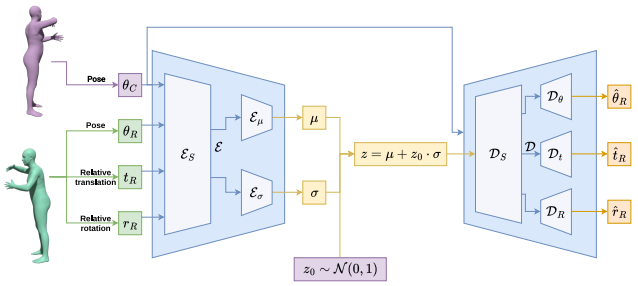

Figure 4: *Detailed cVAE model architecture. The input consists of the conditioning pose θ_C and the reaction pose $\theta_R, \mathbf{t}_R, \mathbf{r}_R$, where $\mathbf{t}_R$ and $\mathbf{r}_R$ are the relative translation and rotation of the reacting avatar with respect to the conditioning avatar. These four tensors are fed into the encoder, which has shared weights $\mathcal{E}_S$ and specialized weights $\mathcal{E}_\mu, \mathcal{E}_\sigma$. The interaction is then encoded into two tensors, μ and σ, that define a normal distribution. To sample a reacting pose from the distribution, we first sample $\mathbf{z}_0$ from the standard normal distribution and then apply the reparameterization trick to obtain the sample $\mathbf{z}$. To decode this sample, the decoding process uses shared weights $\mathcal{D}_S$ and specialized weights $\mathcal{D}_\theta, \mathcal{D}_\mathbf{t}, \mathcal{D}_\mathbf{r}$, which return the corresponding body parameters. Note that the conditioning pose θ_C is also an input to the decoder. Different $\mathbf{z}$ values will generate different reactions for a given pose θ_C. Similarly, the same $\mathbf{z}$ value will generate different avatars depending on the pose θ_C it has to react to.*

The network is trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 50.

Once our cVAE is trained, the decoder $\mathcal{D}$ is be able to generate realistic reactive poses $\mathbf{p}_R$ for any arbitrary conditioning pose $\mathbf{p}_C$.

4.3. Learning to Resolve Human-Human Contact

Despite training on collision-free data, our cVAE cannot guarantee intersection-free samples at test-time. To mitigate this, we propose an additional module $\mathcal{CF}$, which is trained in a self-supervised manner. Our main idea is to infer joint rotation offsets that slightly modify the reaction pose to resolve collisions with the conditioning pose. We do not modify the global translation and rotation to avoid the trivial solution of moving one avatar far from the other. This module is illustrated in Figure 5.

More specifically, the $\mathcal{CF}$ module is a MLP network that takes θ_{Cv} and $\{\hat{\theta}_{Rv}, \hat{\mathbf{r}}_{Rv}, \hat{\mathbf{t}}_{Rv}\}$ as input and predicts joint rotation offsets $\Delta\hat{\theta}_{Rv}$ such that the SMPL mesh rigged using $\{\hat{\theta}_{Rv} + \Delta\hat{\theta}_{Rv}, \hat{\mathbf{r}}_{Rv}, \hat{\mathbf{t}}_{Rv}\}$ does not intersect with the SMPL mesh rigged using θ_{Cv}:

$$\begin{aligned} \mathcal{CF} : (\mathbb{R}^{23\times3\times2}, \mathbb{R}^{23\times3\times2}, \mathbb{R}^{3\times2}, \mathbb{R}^{3}) &\to \mathbb{R}^{23\times3\times2} \\ (\theta_{Cv}, \hat{\theta}_{Rv}, \hat{\mathbf{r}}_{Rv}, \hat{\mathbf{t}}_{Rv}) &\mapsto \Delta\hat{\theta}_{Rv} \end{aligned} \tag{9}$$

To train $\mathcal{CF}$ in a self-supervised manner, we first generate random SMPL poses θ_{Cv} by sampling the VPoser [PCG*19] decoder. These poses, together with a random latent space vector $\mathbf{z} \sim \mathcal{N}(0,1)$, are fed into our decoder $\mathcal{D}$ generating the predicted poses $\mathcal{D}(\{\theta_{Cv}, \mathbf{z}\}) = \{\hat{\theta}_{Rv}, \hat{\mathbf{r}}_{Rv}, \hat{\mathbf{t}}_{Rv}\}$.

We train the $\mathcal{CF}$ network end-to-end, while keeping the $\mathcal{D}$ parameters frozen, using the loss:

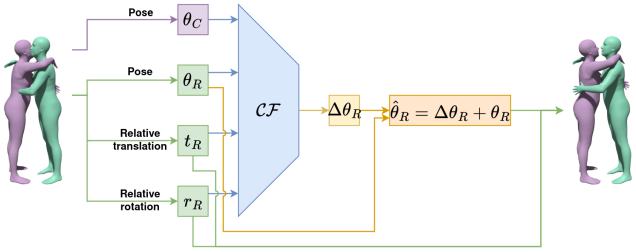

Figure 5: *Detailed architecture collision resolve module. The output is an offset $\Delta\theta_R$ that, once applied to the original reacting pose θ_R, results in a pose $\hat{\theta}_R$ that avoids colliding with the conditioning avatar with pose θ_C.*

$$\mathcal{L}_{\mathcal{CF}} = \lambda_{\text{col}}\mathcal{L}_{\text{col}} + \lambda_{\Delta\hat{\theta}}\mathcal{L}_{\Delta\hat{\theta}} + \lambda_{\text{joints}}\mathcal{L}_{\text{joints}} \tag{10}$$

where $\mathcal{L}_{\Delta\hat{\theta}} = ||\Delta\hat{\theta}||_2^2$ is a regularizer to penalize the joint rotation offsets from growing too large, $\mathcal{L}_{\text{joints}}$ is a regularizer to penalize the 3D joint positions from moving too far from the original ones, and $\mathcal{L}_{\text{col}}$ is a collision penalty between two bodies, explained next.

To efficiently detect and prevent mesh intersections, we approximate each human body using a set of 24 capsules rigidly attached to the skeleton joints (see Supplementary Material for rigging details). Using these capsules, for a pair of avatars, C and R, we compute a distance matrix $\mathbf{D}_{C\times R} \in \mathbb{R}^{24\times24}$, representing the penetration distance between a pair of capsules across both bodies. Positive values indicate no intersection, while negative values indicate capsule overlap. Since capsules provide a loose approximation of the avatar geometry, we establish a tolerance threshold $\mathbf{D_T}$ from our training data. This threshold is computed as the 95th percentile of distance values across all training poses, allowing for natural proximity between avatars without triggering the collision penalty. The collision loss $\mathcal{L}_{\text{col}}$ is then defined as

$$\mathcal{L}_{\text{col}} = \|\min(\mathbf{D}_{C\times R} - \mathbf{D_T}, 0)\| \tag{11}$$

This formulation penalizes only the distances below the learned tolerance threshold $\mathbf{D_T}$, effectively preventing unrealistic intersections while allowing natural close interactions.

We empirically set the loss weights to $\lambda_{\text{col}} = 30$, $\lambda_{\text{joints}} = 0.5$, and $\lambda_{\Delta\hat{\theta}} = 0.1$. Similar to the cVAE, this module is optimized using Adam with a learning rate of 1×10^{-4} and a batch size of 50.

4.4. Inference

As hinted in Figure 2, at inference time, only a conditioning avatar is needed. More specifically, given a conditioning body pose θ_C and after sampling a latent space vector $\mathbf{z} \sim \mathcal{N}(0,1)$, the generated reacting avatar can be computed as:

$$\mathbf{p}_R = \mathcal{CF}(\mathcal{D}(\{\theta_C, \mathbf{z}\})). \tag{12}$$

Method	Data Aug.	In-Distribution (ID)						Out-of-Distribution (OOD)		
		$\mathcal{L}_{\text{rec}}\downarrow$	$\mathcal{E}_{\mathbf{r}_R}\downarrow$	$\mathcal{E}_{\mathbf{t}_R}\downarrow$	**Inter**$\downarrow$	**IoU**$\downarrow$	**VPoser**$\downarrow$	**Inter**$\downarrow$	**IoU**$\downarrow$	**VPoser**$\downarrow$
cVAE ($\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}$)	No	32.731	151.168	15.428	299.790	2.135	54.070	815.900	6.191	60.887
cVAE ($\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}$)	Yes	**7.462**	**10.132**	**3.127**	206.810	1.441	**30.836**	205.057	1.390	**26.391**
cVAE ($\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}, \mathcal{L}_{\text{col}}$)	Yes	66.725	431.798	41.218	**26.090**	**0.229**	239.468	**15.571**	**0.138**	233.418
cVAE ($\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}$) + $\mathcal{CF}$	Yes	9.679	**10.132**	**3.127**	121.690	0.846	31.499	120.629	0.833	28.970

Table 1: *In-Distribution (ID) and Out-of-Distribution (OOD) metrics for the different cVAE variants evaluated on the baseline heldout test set and standalone Mixamo keyframes respectively. Values $\mathcal{L}_{rec}$, $\mathcal{E}_{\mathbf{r}_R}$, $\mathcal{E}_{\mathbf{t}_R}$ and IoU have been scaled by 10^3. The average volume intersection Inter is represented in cm^3. Best results are highlighted in **bold**, and second-best results are underlined.*

5. Results and Evaluation

5.1. Ablation Study

In this section, we present several metrics to justify our architectural choices for conditional human pose generation. Specifically, we compare four distinct model variants. The first is a baseline vanilla cVAE trained exclusively with $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{KL}}$ losses (as detailed in Section 4.2) on a subset of Hi4D [YGK*23] poses identified as containing close-contact interactions. The second variant maintains this vanilla architecture but is trained on our augmented dataset. The third model builds upon this augmented version by incorporating the additional $\mathcal{L}_{\text{col}}$ loss during training. Finally, we evaluate our complete proposed architecture, which applies the $\mathcal{CF}$ module to the augmented-trained vanilla cVAE.

Training: We train our models using a sequence-heldout train-test split, keeping all frames from a given base interaction sequence strictly within either the training or testing set. Specifically, the data is organized into 100 sequence groups. The baseline split contains 5,744 poses from Hi4D interaction scans, divided into 76 training groups (4,616 poses) and 24 testing groups (1,128 poses). The augmented setting merges these baseline poses with 55,897 augmented poses, totaling 61,641 poses partitioned into 49,202 for training and 12,439 for testing.

Metrics: We utilize six key metrics to thoroughly evaluate our models. The first three assess pose reconstruction accuracy: the test-split reconstruction loss $\mathcal{L}_{\text{rec}}$, the mean error on the generated reaction avatar global orientation $\mathcal{E}_{\mathbf{r}_R}$, and the mean error on the generated reaction avatar global translation $\mathcal{E}_{\mathbf{t}_R}$. The next two metrics evaluate physical plausibility by measuring collisions. Specifically, we compute the mean intersected volume (**Inter**) $V(C \cap R)$ between each pair of conditioning avatar C and reaction avatar R, and the mean intersection over union (**IoU**), where the individual $\text{IoU}_{C,R}$ is defined as:

$$\text{IoU}_{C,R} = \frac{V(C \cap R)}{V(C) + V(R) - V(C \cap R)}. \tag{13}$$

Finally, to assess the overall naturalness of the generated poses, we compute the **VPoser** plausibility score. We take the generated body pose, isolate the first 21 body joints, encode this pose using VPoser, and calculate the latent prior energy as $\frac{1}{2}\sum(\exp(\text{logvar}) + \mu^2 - 1 - \text{logvar})$. Poses that lie closer to the learned human-pose prior yield lower energy values, indicating that lower scores reflect more plausible and realistic human poses.

Evaluation: We evaluate our models on in-distribution (ID) and out-of-distribution (OOD) benchmarks, sampling 10 generations per conditioning pose for both. For the ID setup, we use 100 conditioning poses from the baseline test split (yielding 1,000 generations). Because these samples possess ground-truth interacting partners, we evaluate them using the complete suite of reconstruction, collision, and plausibility metrics defined above. For the OOD setting, we select 70 keyframes from 35 Mixamo animations (yielding 700 generations). Since we lack ground-truth reaction poses for this out-of-distribution set, our evaluation relies strictly on measuring the physical plausibility (VPoser) and collision (IoU, Inter) metrics of the generated reaction poses.

Results As shown in Table 1, the performance trends remain remarkably consistent across both the In-Distribution (ID) and Out-of-Distribution (OOD) evaluations. First, we observe a clear improvement when training the baseline cVAE on our augmented dataset (second row) compared to the non-augmented baseline (first row), resulting in a drastic reduction across all the metrics. The augmented baseline establishes a strong foundation, yielding the best overall reconstruction loss ($\mathcal{L}_{\text{rec}}$) as well as the lowest global orientation ($\mathcal{E}_{\mathbf{r}_R}$) and translation ($\mathcal{E}_{\mathbf{t}_R}$) errors.

To address the remaining mesh intersections, incorporating the additional capsule loss term $\mathcal{L}_{\text{col}}$ (third row) proves highly effective at avoiding collisions, yielding the absolute lowest Inter and IoU values. However, this aggressive collision resolution comes at a severe cost. Because the loss penalty artificially pushes bodies apart to clear intersections, the reconstruction loss ($\mathcal{L}_{\text{rec}}$) degrades significantly, and the human pose plausibility prior is broken, as evidenced by a drastic spike in the VPoser energy.

Finally, applying our proposed $\mathcal{CF}$ module (fourth row) offers the most balanced trade-off, consistently securing first or second best across all evaluated metrics. Because $\mathcal{CF}$ resolves collisions by exclusively modifying the local body pose, the global trajectory errors ($\mathcal{E}_{\mathbf{r}_R}$ and $\mathcal{E}_{\mathbf{t}_R}$) remain entirely unaffected, maintaining the best-in-class performance of the augmented baseline. While moving vertices to prevent collisions inevitably incurs a slight penalty to the reconstruction loss compared to the unconstrained augmented model, $\mathcal{CF}$ preserves a significantly better $\mathcal{L}_{\text{rec}}$ than the capsule loss. Ultimately, our module successfully minimizes collisions while maintaining natural and plausible body structures, yielding a VPoser energy that is nearly identical to the augmented-only approach.

5.2. Quantitative evaluation

To further quantitatively evaluate our method, we compare our results with BUDDI [MYP*24] in terms of mesh intersections. To enable this comparison, we adapt BUDDI to generate reaction poses conditioned on a conditioning pose using an in-painting strategy: at each step of the BUDDI sampling process, we pass the conditioning pose prior to the diffusion and denoising steps.

Table 2 presents two metrics (**Inter** and **IoU**) as described in the previous section. To ensure a fair comparison and avoid dataset bias, we compute these metrics not on randomly sampled outputs from our model, but on a specific split of the Hi4D dataset. While both our model and BUDDI were trained on instances from the full Hi4D dataset, we ensure that the selected split contains only poses that were unseen by our model during training. For additional fairness and to obtain a more informative evaluation, we only evaluate test samples that generate reaction poses that are in-contact with the conditioning pose. This is to avoid naive samples that are in interaction but are not in contact, which, of course, do not suffer from mesh intersections.

Our results in Table 2 show that BUDDI samples exhibit significantly more mesh intersections, which is reflected in the higher **Inter** and **IoU** values. In contrast, our method maintains a marginal intersection error, showcasing that it is able to generate interacting meshes that are in close contact but do not intersect with each other.

Method	Inter↓	IoU↓
Ours	**182.45**	**1.25**
BUDDI [MYP*24]	2283.90	16.87

Table 2: *Quantitative comparison between our method and BUDDI [MYP*24]. IoU values have been scaled by* 10^3. *The average volume intersection Inter is represented in* cm^3

We further illustrate these differences in the qualitative evaluation Section 5.3.

5.3. Qualitative evaluation

In this section, we present qualitative results of our approach and compare with BUDDI [MYP*24]. To this end, we generated conditioned reaction avatars with BUDDI as explained in the previous section.

We first evaluate the smoothness of the latent space of each model by interpolating between two random latent vectors z, using a fixed conditioning pose. For a fair comparison and to avoid any dataset bias, we specifically select conditioning poses that are originally generated with BUDDI, hence never seen at train time by our model. In Figure 7, we present for both Ours and BUDDI, 8 equally spaced steps of interpolation between two random samples. Our results demonstrate that the interpolations generated by our method exhibit smoother transitions, a greater diversity of poses, and fewer mesh collisions compared to those produced by BUDDI. For further insights, please refer to the Supplementary Video and Material.

In Figure 8, we compare the expressivity of our model and BUDDI by sampling diverse reaction poses conditioned on a conditioning pose from the Hi4D *dance* category (See the Supplementary Material for more examples). Results demonstrate that our model generates rich and semantically consistent responses. In contrast, BUDDI often lacks contextual coherence and shows interpenetrations.

Figure 6: *Our results (i.e., renders with conditioning and reaction poses) can be used as accurate driving signal to control image-to-image models [Lab24]. This enables fine-control synthesis of photorealistic images with complex human-human interactions.*

Figures 1 and 6 showcase applications of our method, highlighting how our increased controllability enables easy control for text-conditioned image generative tools like FLUX.1-Depth-dev [Lab24], which yields accurate pose control human synthesis. Please refer to the Supplementary Material for more examples.

5.4. User Study

To evaluate the perceptual plausibility of our generated interactions, we conducted a user study with 35 participants (19 female, 16 male), aged 18 to 58 years ($\mu = 32.2$ years, $\sigma = 11.4$ years), rating 3D dyadic scenes on a 5-point Likert scale (1 = Very Implausible, 5 = Very Plausible).

We designed the study around 10 base interactions, each presented under three conditions: ground truth, augmented (i.e. samples from Section 3), and generated with our method. Critically, the conditioning avatar is held identical across all three conditions for a given interaction, ensuring that any difference in plausibility ratings is driven solely by the reacting avatar. Each participant evaluated 30 scenes, distributed across three randomized blocks to prevent immediate cross-condition comparison. Participants interacted with a WebGL viewer that allowed free camera rotation, panning, and zooming to resolve spatial ambiguities before rating.

Mean plausibility scores were 3.97 ± 0.76 for ground truth, 3.91 ± 0.78 for augmented poses, and 4.02 ± 0.69 for generated poses. To test whether generated poses are perceived as equivalent to ground truth, rather than merely not different, we applied equivalence testing (TOST) with a margin of $\delta = 0.5$ points. The results confirm perceptual equivalence between generated and ground truth poses ($p < 0.0001$), between augmented and ground truth poses ($p < 0.0001$), and between generated and augmented poses ($p < 0.0001$). These findings demonstrate that our model produces interactions that are indistinguishable in perceived plausibility from real captured data.

Figure 7: *Two examples (Interpolation 1 and Interpolation 2) of the reacting poses yielded by interpolating two latent vectors (left and right columns) for a constant conditioning pose. The conditioning poses were originally generated with BUDDI, hence were never seen at train time by our model. Our model generates smoother and more expressive poses, notice seamless and smooth changes of the reacting poses along the horizontal axis. In contrast, BUDDI reacting poses are less expressive and exhibit pose discontinuities.*

6. Conclusions

This work presented a novel approach for human-human interaction generation. Our model enables the 3D collision-aware synthesis of humans in interaction, conditioned on one pose. We introduced a new data augmentation strategy that leverages existing datasets to reduce the need for extensive data collection, thus reducing the dependency on costly motion capture systems and post-processing pipelines. This democratizes the creation of high-quality human-human interaction data for a broader range of users.

We also demonstrate that traversing our latent space results in smooth transitions across diverse poses, which can become a powerful tool for artists to fine-tune their creations. Furthermore, our generated reaction poses consistently exhibit greater contextual coherence and fewer interpenetrations compared to those produced by BUDDI [MYP*24].

A potential open path for exploration is integrating our approach as a post-processing step in existing methods for tracking and reconstructing 3D humans [SLB*22,MYP*24], which would provide a greater variety and enhanced controllability of the output pose. Additionally, incorporating our model into text-to-motion generative methods would enable the synthesis of realistic animations of humans in contact.

Despite improving 3D interaction synthesis, our model currently handles only static poses. Since the trajectory leading to a pose often defines its semantic meaning, future work involves extending our data augmentation and architecture to incorporate time-dependent information for dynamic interactions. Furthermore, since each augmented pose is treated as a quasi-static configuration, we currently do not employ Continuous Collision Detection (CCD) in our data generation stage. While our SDF-based approach effectively resolves static interpenetrations, interpreting the augmentation as a continuous deformation and applying CCD could better preserve first-contact configurations, particularly for thin structures such as fingers. Beyond physical and temporal constraints, our stochastic augmentation lacks explicit semantic limits, which could theoretically generate socially awkward interactions. However, our user studies did not identify any such problematic cases. Explicitly enforcing social plausibility remains an interesting area for further exploration.

Additionally, while our $\mathcal{CF}$ module resolves geometric interpenetration via volumetric proxies, it is not a full physics simulation. It does not account for secondary dynamics like muscle deformation, soft-body contact, or gravity-induced balance. In rare cases of extreme mesh entanglement, the module may converge to local minima. We thus view $\mathcal{CF}$ as a geometric prerequi-

Figure 8: *Examples of conditional sampling from our model and BUDDI [MYP*24] for a set of conditioning pose, labeled as dance, from the Hi4D dataset. These conditioning poses belong to the in-distribution benchmark. We show two views and four different samples, organized in four columns. Our generated reaction poses respond naturally to the input, producing realistic human-human 3D interactions that are semantically consistent with the input poses. In contrast, BUDDI reaction poses lack contextual coherence and frequently display noticeable interpenetrations.*

site for realism, rather than a replacement for high-fidelity physics solvers [ZGY*23]. Finally, this module is explicitly designed to reduce inter-body collisions and does not include a separate self-contact loss, meaning the matrix is computed across the two bodies rather than within a single body. However, because our data augmentation stage successfully resolves self-penetrations using SDFs (as detailed in Section 3.2), the model does not observe self-collisions during training. Consequently, the network rarely produces outputs with noticeable self-penetrations in practice. Incorporating an explicit self-collision term during this refinement stage remains a potential extension for future work.

Acknowledgments

This work has been partially funded by the Comunidad de Madrid through two initiatives. First, in the framework of the Multiannual Agreement with the Universidad Rey Juan Carlos in line of Action 1, "Estímulo a la investigación de jóvenes doctores", for the project "Captura de humanos restringida por sus entornos" (Acronym: CaptHuRe) with reference M2736. Second, the presented results are also part of the project "Inteligencia artificial para la industria 4.0: generación de datos, modelado avanzado optimización e interpretabilidad" (Acronym: IDEA-CM), with reference TEC-2024/COM-89, funded through the call for grants for collaborative R&D projects under the modality of "Programas de Actividades de I+D en Tecnologías 2024", according to Order 3177/2024.

References

[CKA*22] CHRISTEN S., KOCABAS M., AKSAN E., HWANGBO J., SONG J., HILLIGES O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022), IEEE, pp. 20545–20554. doi:10.1109/CVPR52688.2022.01992. 2

[CTO*23] CHOPIN B., TANG H., OTBERDOUT N., DAOUDI M., SEBE N.: Interaction Transformer for Human Reaction Generation. *IEEE Transactions on Multimedia 25* (2023), 8842–8854. doi:10.1109/TMM.2023.3242152. 2

[FTC*25] FAN K., TANG J., CAO W., YI R., LI M., GONG J., ZHANG J., WANG Y., WANG C., MA L.: Freemotion: A unified framework for number-free text-to-motion synthesis. In *Computer Vision – ECCV 2024* (Cham, 2025), Leonardis A., Ricci E., Roth S., Russakovsky O., Sattler T., Varol G., (Eds.), Springer Nature Switzerland, pp. 93–109. 2

[FZO*20] FIERARU M., ZANFIR M., ONEATA E., POPA A.-I., OLARU V., SMINCHISESCU C.: Three-dimensional reconstruction of human interactions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2020), IEEE, pp. 7212–7221. doi:10.1109/CVPR42600.2020.00724. 2

[FZS*21] FIERARU M., ZANFIR M., SZENTE T. A., BAZAVAN E. G., OLARU V., SMINCHISESCU C.: Remips: physically consistent 3d reconstruction of multiple interacting people under weak supervision. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2021), NIPS '21, Curran Associates Inc. 2, 3

[GBAPM22] GUO W., BIE X., ALAMEDA-PINEDA X., MORENO–NOGUER F.: Multi-person extreme motion prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022), IEEE, pp. 13043–13054. doi:10.1109/CVPR52688.2022.01271. 2

[GCM*24] GUZOV V., CHIBANE J., MARIN R., HE Y., SARACOGLU Y., SATTLER T., PONS-MOLL G.: Interaction replica: Tracking human–object interaction and scene changes from human motion. In *2024 International Conference on 3D Vision (3DV)* (Davos, Switzerland, 2024), IEEE, pp. 1006–1016. doi:10.1109/3DV62453.2024.00072. 2

[GSS*15] GAST T. F., SCHROEDER C., STOMAKHIN A., JIANG C., TERAN J. M.: Optimization integrator for large time steps. *IEEE transactions on visualization and computer graphics 21*, 10 (2015), 1103–1115. 4

[GZD*25] GHOSH A., ZHOU B., DABRAL R., WANG J., GOLYANIK V., THEOBALT C., SLUSALLEK P., GUO C.: DuetGen: Music Driven Two-Person Dance Generation via Hierarchical Masked Modeling. In *ACM SIGGRAPH Conference Proceedings* (2025). 3

[HCV*21] HASSAN M., CEYLAN D., VILLEGAS R., SAITO J., YANG J., ZHOU Y., BLACK M.: Stochastic scene-aware motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, 2021), IEEE, pp. 11354–11364. doi:10.1109/ICCV48922.2021.01118. 2

[HGT*21] HASSAN M., GHOSH P., TESCH J., TZIONAS D., BLACK M. J.: Populating 3d scenes by learning human-scene interaction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, 2021), IEEE, pp. 14703–14713. doi:10.1109/CVPR46437.2021.01447. 2

[HKT10] HO E. S. L., KOMURA T., TAI C.-L.: Spatial relationship preserving character motion adaptation. In *ACM SIGGRAPH 2010 Papers* (New York, NY, USA, 2010), SIGGRAPH '10, Association for Computing Machinery. URL: https://doi.org/10.1145/1833349.1778770, doi:10.1145/1833349.1778770. 3

[HVT*19] HASSON Y., VAROL G., TZIONAS D., KALEVATYKH I., BLACK M. J., LAPTEV I., SCHMID C.: Learning joint reconstruction of hands and manipulated objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA, USA, 2019), IEEE, pp. 11799–11808. doi:10.1109/CVPR.2019.01208. 2

[JGCL25] JAVED M. G., GUO C., CHENG L., LI X.: InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling. In *The Thirteenth International Conference on Learning Representations* (Singapore, 2025), OpenReview.net, pp. 1–22. URL: https://openreview.net/forum?id=ZAyuwJYN8N. 2

[JGK*24] JIANG Z., GUO C., KAUFMANN M., JIANG T., VALENTIN J., HILLIGES O., SONG J.: Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024). 3

[JLC*23] JIANG N., LIU T., CAO Z., CUI J., ZHANG Z., CHEN Y., WANG H., ZHU Y., HUANG S.: Full-body articulated human-object interaction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, 2023), IEEE, pp. 9331–9342. doi:10.1109/ICCV51070.2023.00859. 2

[JZL*24] JIANG N., ZHANG Z., LI H., MA X., WANG Z., CHEN Y., LIU T., ZHU Y., HUANG S.: Scaling up dynamic human-scene interaction modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024), IEEE, pp. 1737–1747. doi:10.1109/CVPR52733.2024.00171. 2

[KHHL12] KIM M., HWANG Y., HYUN K., LEE J.: Tiling motion patches. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2012), SCA '12, Eurographics Association, p. 117–126. 3

[KMOW00] KANE C., MARSDEN J. E., ORTIZ M., WEST M.: Variational integrators and the newmark algorithm for conservative and dissipative mechanical systems. *International Journal for Numerical Methods in Engineering 49*, 10 (2000), 1295–1325. doi:https://doi.org/10.1002/1097-0207(20001210)49:10<1295::AID-NME993>3.0.CO;2-W. 4

[Lab24] LABS B. F.: Flux. https://github.com/black-forest-labs/flux, 2024. 8

[LGZW25] LIU S., GUO C., ZHOU B., WANG J.: Ponimator: Unfolding interactive pose for versatile human-human interaction animation. In *ICCV* (2025). 3

[LHP06] LIU C. K., HERTZMANN A., POPOVIĆ Z.: Composition of complex optimal multi-character motions. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2006), SCA '06, Eurographics Association, p. 215–222. 3

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: Smpl: a skinned multi-person linear model. *ACM Trans. Graph. 34*, 6 (Oct. 2015). URL: https://doi.org/10.1145/2816795.2818013, doi:10.1145/2816795.2818013. 2, 5

[LZL*24] LIANG H., ZHANG W., LI W., YU J., XU L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. *Int. J. Comput. Vision 132*, 9 (Mar. 2024), 3463–3483. URL: https://doi.org/10.1007/s11263-024-02042-6, doi:10.1007/s11263-024-02042-6. 2

[MBS*18] MUELLER F., BERNARD F., SOTNYCHENKO O., MEHTA D., SRIDHAR S., CASAS D., THEOBALT C.: Ganerated hands for real-time 3d hand tracking from monocular rgb. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA, 2018), IEEE, pp. 49–59. doi:10.1109/CVPR.2018.00013. 3

[MCZ*25] MÜLLER L., CHOI H., ZHANG A., YI B., MALIK J., KANAZAWA A.: Reconstructing people, places, and cameras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, 2025), IEEE, pp. 21948–21958. doi:10.1109/CVPR52734.2025.02044. 2

[MMR*24] MALULEKE V. H., MÜLLER L., RAJASEGARAN J., PAVLAKOS G., GINOSAR S., KANAZAWA A., MALIK J.: Synergy and synchrony in couple dances. *arXiv preprint arXiv:2409.04440 abs/2409.04440* (2024), 1–11. URL: https://arxiv.org/abs/2409.04440, doi:10.48550/arXiv.2409.04440. 2

[MOT*21] MÜLLER L., OSMAN A. A. A., TANG S., HUANG C.-H. P., BLACK M. J.: On self-contact and human pose. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, 2021), IEEE, pp. 9985–9994. doi:10.1109/CVPR46437.2021.00986. 3

[MSB*22] MIHAJLOVIC M., SAITO S., BANSAL A., ZOLLHOEFER M., TANG S.: Coap: Compositional articulated occupancy of people. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022), IEEE, pp. 13191–13200. doi:10.1109/CVPR52688.2022.01285. 3

[MYP*24] MÜLLER L., YE V., PAVLAKOS G., BLACK M., KANAZAWA A.: Generative proxemics: A prior for 3d social interaction from images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024), IEEE, pp. 9687–9697. doi:10.1109/CVPR52733.2024.00925. 2, 3, 8, 9, 10

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA, USA, 2019), IEEE, pp. 10967–10977. doi:10.1109/CVPR.2019.01123. 2, 4, 6

[SBL*21] SUN Y., BAO Q., LIU W., FU Y., BLACK M. J., MEI T.: Monocular, one-stage, regression of multiple 3d people. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, 2021), IEEE, pp. 11159–11168. doi:10.1109/ICCV48922.2021.01099. 2

[SDH*25] SHAN M., DONG L., HAN Y., YAO Y., LIU T., NWOGU I., QI G.-J., HILL M.: Towards open domain text-driven synthesis of multi-person motions. In *Computer Vision – ECCV 2024* (Cham, 2025), Leonardis A., Ricci E., Roth S., Russakovsky O., Sattler T., Varol G., (Eds.), Springer Nature Switzerland, pp. 67–86. 2

[SKY07] SHUM H. P. H., KOMURA T., YAMAZAKI S.: Simulating competitive interactions using singly captured motions. In *Proceedings of the 13th ACM Symposium on Virtual Reality Software and Technology (VRST)* (New York, NY, USA, 2007), VRST '07, Association for Computing Machinery, p. 65–72. URL: https://doi.org/10.1145/1315184.1315194, doi:10.1145/1315184.1315194. 3

[SKY12] SHUM H. P., KOMURA T., YAMAZAKI S.: Simulating multiple character interactions with collaborative and adversarial goals. *IEEE Transactions on Visualization and Computer Graphics 18*, 5 (2012), 741–752. doi:10.1109/TVCG.2010.257. 3

[SLB*22] SUN Y., LIU W., BAO Q., FU Y., MEI T., BLACK M. J.: Putting people in their place: Monocular regression of 3d people in depth. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022), IEEE, pp. 13233–13242. doi:10.1109/CVPR52688.2022.01289. 2, 9

[SYL15] SOHN K., YAN X., LEE H.: Learning structured output representation using deep conditional generative models. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA, 2015), NIPS'15, MIT Press, p. 3483–3491. 5

[SYZ*23] SHUAI Q., YU Z., ZHOU Z., FAN L., YANG H., YANG C., ZHOU X.: Reconstructing close human interactions from multiple views. *ACM Trans. Graph. 42*, 6 (Dec. 2023). URL: https://doi.org/10.1145/3618336, doi:10.1145/3618336. 3

[TCBT22] TAHERI O., CHOUTAS V., BLACK M. J., TZIONAS D.: Goal: Generating 4d whole-body motion for hand-object grasping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022), IEEE, pp. 13253–13263. doi:10.1109/CVPR52688.2022.01291. 2

[TF23] TANAKA M., FUJIWARA K.: Role-aware interaction generation from textual description. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, 2023), IEEE, pp. 15953–15963. doi:10.1109/ICCV51070.2023.01466. 2

[UPP*24] UGRINOVIC N., PAN B., PAVLAKOS G., PASCHALIDOU D., SHEN B., SANCHEZ-RIERA J., MORENO-NOGUER F., GUIBAS L.: Multiphys: Multi-person physics-aware 3d motion estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024), IEEE, pp. 2331–2340. doi:10.1109/CVPR52733.2024.00226. 3

[WLO*14] WON J., LEE K., O'SULLIVAN C., HODGINS J. K., LEE J.: Generating and ranking diverse multi-character interactions. *ACM Trans. Graph. 33*, 6 (Nov. 2014). URL: https://doi.org/10.1145/2661229.2661271, doi:10.1145/2661229.2661271. 3

[XBPM22] XIE X., BHATNAGAR B. L., PONS-MOLL G.: Chore: Contact, human and object reconstruction from a single rgb image. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 125–145. URL: https://doi.org/10.1007/978-3-031-20086-1_8, doi:10.1007/978-3-031-20086-1_8. 2

[XLY*24] XU L., LV X., YAN Y., JIN X., WU S., XU C., LIU Y., ZHOU Y., RAO F., SHENG X., LIU Y., ZENG W., YANG X.: Inter-x: Towards versatile human-human interaction analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024), IEEE, pp. 22260–22271. doi:10.1109/CVPR52733.2024.02101. 2

[YGK*23] YIN Y., GUO C., KAUFMANN M., ZARATE J. J., SONG J., HILLIGES O.: Hi4d: 4d instance segmentation of close human interaction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada, 2023), IEEE, pp. 17016–17027. doi:10.1109/CVPR52729.2023.01632. 2, 3, 4, 7

[YGKT24] YE Y., GUPTA A., KITANI K., TULSIANI S.: G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA, 2024), IEEE, pp. 1911–1920. doi:10.1109/CVPR52733.2024.00187. 2

[YPMK23] YE V., PAVLAKOS G., MALIK J., KANAZAWA A.: Decoupling human and camera motion from videos in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC, Canada, 2023), IEEE, pp. 21222–21232. doi:10.1109/CVPR52729.2023.02033. 2, 3

[ZBL*19] ZHOU Y., BARNES C., LU J., YANG J., LI H.: On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, USA, 2019), IEEE, pp. 5738–5746. doi:10.1109/CVPR.2019.00589. 5

[ZGY*23] ZHANG Y., GOPINATH D., YE Y., HODGINS J., TURK G., WON J.: Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 Conference Proceedings* (New York, NY, USA, 2023), SIGGRAPH '23, Association for Computing Machinery. URL: https://doi.org/10.1145/3588432.3591491, doi:10.1145/3588432.3591491. 3, 10

[ZSZ*21] ZHENG Y., SHAO R., ZHANG Y., YU T., ZHENG Z., DAI Q., LIU Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, 2021), IEEE, pp. 6219–6229. doi:10.1109/ICCV48922.2021.00618. 2, 3

[ZWZ*22] ZHAO K., WANG S., ZHANG Y., BEELER T., TANG S.: Compositional human-scene interaction synthesis with semantic control. In *Computer Vision – ECCV 2022* (Cham, 2022), Avidan S., Brostow G., Cissé M., Farinella G. M., Hassner T., (Eds.), Springer Nature Switzerland, pp. 311–327. 2

[ZZW*23] ZHAO K., ZHANG Y., WANG S., BEELER T., TANG S.: Synthesizing diverse human motions in 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France, 2023), IEEE, pp. 14692–14703. doi:10.1109/ICCV51070.2023.01354. 2