Closely Interactive Human Reconstruction with Proxemics and Physics-Guided Adaption

Buzhen Huang[1,2*] Chen Li[1] Chongyang Xu[3] Liang Pan[2] Yangang Wang[2] Gim Hee Lee[1]

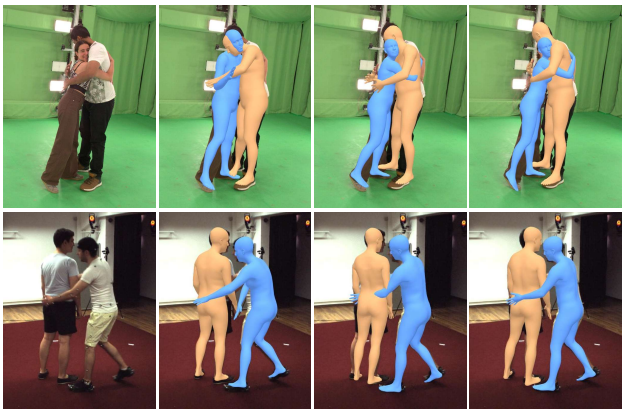
[1]National University of Singapore [2]Southeast University [3]Sichuan University

Abstract

Existing multi-person human reconstruction approaches mainly focus on recovering accurate poses or avoiding penetration, but overlook the modeling of close interactions. In this work, we tackle the task of reconstructing closely interactive humans from a monocular video. The main challenge of this task comes from insufficient visual information caused by depth ambiguity and severe inter-person occlusion. In view of this, we propose to leverage knowledge from proxemic behavior and physics to compensate the lack of visual information. This is based on the observation that human interaction has specific patterns following the social proxemics. Specifically, we first design a latent representation based on Vector Quantised-Variational AutoEncoder (VQ-VAE) to model human interaction. A proxemics and physics guided diffusion model is then introduced to denoise the initial distribution. We design the diffusion model as dual branch with each branch representing one individual such that the interaction can be modeled via cross attention. With the learned priors of VQ-VAE and physical constraint as the additional information, our proposed approach is capable of estimating accurate poses that are also proxemics and physics plausible. Experimental results on Hi4D, 3DPW, and CHI3D demonstrate that our method outperforms existing approaches. The code is available at https://github.com/boycehbz/HumanInteraction.

1. Introduction

In everyday life, people continuously interact with each other to achieve goals or simply to exchange states of mind that enables us to live in groups and share skills and purposes. Analyzing human interactions in 3D space with computer vision techniques may promote the understanding of our social intellect. However, current multi-person human reconstruction methods often focus on pose accuracy [4, 17, 56], penetration avoidance [21, 74] or spatial distribution reasonableness [18, 19, 57, 64], and always ignore the important close interactions.



(a) Input (b) BUDDI (c) Initial Prediction (d) Ours

Figure 1. Our method reconstructs closely interactive humans with plausible body poses, natural proxemic relationships and accurate physical contacts from single-view inputs. To address this challenging task, we formulate the reconstruction as a distribution adaption from the initial prediction (c). Compared to the existing solution, BUDDI [36] (b), our method (d) is more robust to visual ambiguity.

In this paper, we aim to reconstruct humans in close proximity with their physical social interactions from a monocular video. In addition to estimating plausible body poses, the task focuses on modeling natural interactive behaviors and accurate physical contacts. The main challenge of this task comes from insufficient visual information caused by depth ambiguity and severe inter-person occlusion. This results in unsatisfactory prediction with high uncertainty. To circumvent the problems, we propose to leverage knowledge from proxemic behavior and physics to compensate the lack of visual information. This is based on the observation that human interaction has specific patterns following the social proxemics, and human can still infer plausible poses based on their prior knowledge even when there is severe occlusion. Inspired by this observation, we

[1]The work was done while Buzhen Huang is a visiting student at National University of Singapore.

introduce a proxemics and physics guided diffusion model to reconstruct closely interactive humans.

First, we propose a latent representation to model human social interactions and proxemic behaviors. Historically, latent motion priors [32, 46, 50, 54, 78] have shown promising performance in modeling human dynamics. However, human interaction requires not only vivid individual motions but also realistic interrelationships. We found that directly encoding the two-person interactions in a unified latent space with the similar structure as previous priors [32, 78] compromise individual motion fidelity. We therefore design a dual-branch network to learn two discrete codebooks to model the social interactions with Vector Quantised-Variational AutoEncoder (VQ-VAE) [62]. These codebooks represent high-fidelity motions and can share information during the inference phase.

Subsequently, we design a diffusion model to reconstruct 3D human interactions from monocular videos. However, conventional diffusion models always start from the standard Gaussian distribution to predict a clean sample, and the early denoising iterations primarily produce noises with limited information [69, 73]. Therefore, we first regress an initial distribution from image features to draw plausible poses for the denoising process. Although the coarse poses may not be fully consistent with image observations due to the visual ambiguity and occlusion, they are valid signals to be evaluated by the latent interaction prior and physical constraints. We thus use the diffusion model as an adaptor to refine the initial distribution under the guidance of learned interaction prior, physical constraints, and image observations. In each timestep, the current interactive poses are fed into the interaction prior to find the closest pair, which are then used to update the current states. In addition, we further design a metric to evaluate the penetration, which reflects the physical plausibility of the current interaction state. We also project the 3D joint positions to the 2D image plane and calculate the projection loss gradients, and enforce the results to be consistent with the image observation. The physical loss, projection loss gradients, and image features are then concatenated as a condition to guide the diffusion model to achieve the distribution adaption. After several diffusion timesteps, we can obtain the final distribution to draw accurate and realistic 3D interactions. To summarize, the main contributions of this paper are as follows:

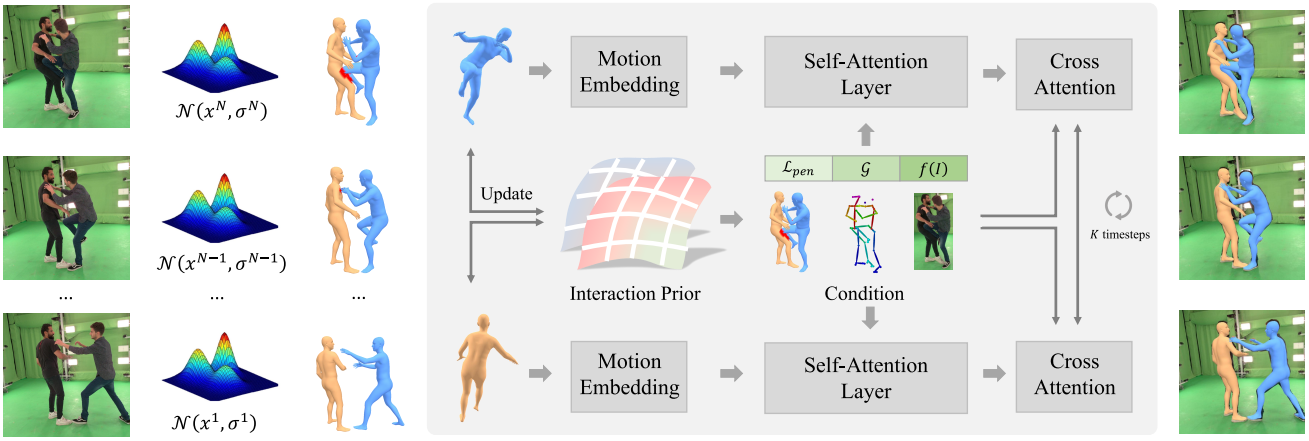
- We formulate the close interaction reconstruction as a distribution adaption process, which can estimate plausible body poses, natural proxemic relationships and accurate physical contacts from a single-view video.
- We design a dual-branch discrete prior to learn human interactive behaviours that can provide additional knowledge for single-view reconstruction.
- We propose a novel diffusion model to incorporate proxemics, physics, and image observations for improving the reconstruction. This model achieves state-of-the-art performance in closely interactive scenarios.

2. Related Work

Single-person shape and pose estimation. Early works in single-person shape and pose estimation directly predict a 3D mesh with optimization [2] or regression [22]. With the development of computer vision techniques, more advanced backbones [7, 12, 55], representations [16, 28], camera models [29], and priors [39] are introduced to improve the estimation. To exploit the temporal context of human motion for better temporal consistency, video-based methods use the recurrent neural network [26] or transformer [80] to process the input. However, these methods encounter global inconsistency, since they can only produce root-relative motion. Since the aforementioned methods do not consider depth ambiguity and occlusion from monocular motion capture, recent diffusion-based works [5, 13, 15] model the uncertainty of 2D-3D lifting for 3D pose estimation. Nevertheless, while single-person methods can achieve better performance in joint accuracy, they do not consider the relationships between humans in the same scene and cannot be used to reconstruct human interactions.

Multi-person shape and pose estimation. Recently, along with the success of deep learning in 3D computer vision, multi-person shape and pose estimation has made tremendous progress. To address the inter-person occlusions, some works [4, 27, 42, 56] design various model architectures to extract valid features and can recover accurate body meshes from monocular images. However, they do not consider absolute positions and cannot model delicate human interactions. To estimate multi-person meshes in a unified 3D space, a few works estimate absolute translations with projection geometry [3, 17, 74, 75], but the strategy is affected by depth and body shape coupling [61]. In addition, it also strongly depends on the quality of 2D poses, which is hard to obtain in interactive cases. Recently, a few works introduce position representation [57, 76], 6D pose estimation [37], and ordering-aware loss [18, 21, 24, 64] to constrain relative positions among humans. Nonetheless, the coarse ordinal relation is inadequate for close interaction reconstruction. Only a few works explicitly consider the close interactions. Fieraru *et al*. designed several losses to prevent self-collision and interpenetration. [10, 11] incorporate additional contact constraints in the optimization. BUDDI [36] further trains a diffusion-based proxemics prior to assist the fitting. However, all these methods cannot utilize temporal information and are still confronted with visual ambiguities.

Closely interactive motion generation. Human motion generation is a hot topic in recent years. Different from sin-



(a) Input (b) Distribution Initialization (c) Distribution Adaption (d) Output

Figure 2. **Overview of our method.** Given a monocular video with interactive humans (a), we first regress an initial distribution for each person (Inter-person penetration is marked in red) (b). To refine the distribution, we design a proxemics and physics guided diffusion model to achieve distribution adaption (c). Specifically, the motions drawn from the initial distributions are updated by a discrete interaction prior. The updated motions are then fed into a dual-branch diffusion model to denoise under the guidance of physics and image observations. The denoised motions are then used as the input for the next timestep. The adaption takes several diffusion timesteps and finally produce accurate results (d).

gle person scenarios [59], the closely interactive behavior is difficult to record, and the insufficient data is the bottleneck of multi-person motion generation for a long time. Early work [71] can only adopt sampling strategy to generate interactions. Due to the scarcity of training data, some recent works [65, 79] use single-person data with reinforcement learning to achieve multi-person motion generation. The trained models cannot be generalized to various interaction types [49, 65]. Other works [41, 58, 67, 68] train the models with motion capture data, but the data contains only a few close interaction cases. Game engines can also be used to generate training data for close interactions [52, 53, 66]. With the development of motion capture techniques, the recent works can use real interaction data to generate human reactions [6, 9, 14, 43]. InterGen [30] builds a large-scale two-person interaction dataset based on a system with massive cameras, which promote the close interaction generation. It also proposes a diffusion-based network to generate realistic interactions. However, all the above methods adopt skeletons to represent the interactive pairs, and only focus on the pose accuracy. Furthermore, although they can generate interactive behaviors, their models cannot be used as a prior for other downstream tasks (*e.g.*, monocular motion capture). In contrast, our interaction prior based on mesh representation can consider the important body contacts.

3. Our Method

In this work, we aim to reconstruct closely interactive human motions from monocular videos. Since it is difficult to obtain sufficient valid information from single-view images, we design a discrete interaction representation to learn interactive human behaviors to assist the motion capture. We then model the uncertainty of human interaction as a set of Gaussian distributions and formulate the interaction reconstruction as a distribution adaption. With a dual-branch diffusion model, we can gradually adapt interaction distributions under the guidance of proxemic behaviors, physical laws and image observations.

3.1. Representation

For an interactive pair, we use two SMPL models [33] to represent the interactions. We also adopt 6D representation [81] to describe joint rotations, and thus the parameters for a single person $x = \{\theta, \beta, \tau\}$ consist of pose $\theta \in \mathbb{R}^{144}$, shape $\beta \in \mathbb{R}^{10}$ and translation $\tau \in \mathbb{R}^3$. An interactive motion with N frames of two people can be denoted as $\mathbf{x}^{1:N} = \{\mathbf{x}^{a,1:N}, \mathbf{x}^{b,1:N}\}$, where $\mathbf{x}^{a,1:N} = \{x^i\}_{i=1}^N$. We use $\hat{\mathbf{x}}$ to represent ground-truth data. Since the single-view reconstruction is an ill-posed problem, we use a set of Gaussian distributions $\mathcal{N}(\mathbf{x}^{1:N}, \sigma^{1:N})$ to model the uncertainty, where σ is the variances. A motion represented with SMPL parameters can be directly drawn from the distributions. Given a video $\mathcal{I} = \{I^i\}_{i=1}^N$ with close interactions of two people, the output of our network are two sets of distributions $\{\mathcal{N}^a, \mathcal{N}^b\}$.

3.2. Discrete Interaction Prior

The way that people react to and interact with the surrounding world is a result of evolution that follows certain rules. With only partial observations from single-view images, humans can also infer the complete 3D poses for an interactive pair. We thus hope to learn the interactive behaviors to provide additional knowledge for the monocular interaction

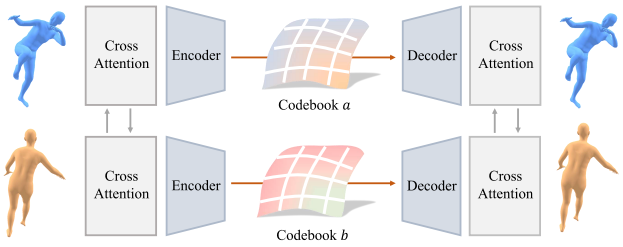


Figure 3. The prior has a dual-branch structure, with each branch representing the motion of a character. Each branch has a codebook learned by VQ-VAE, which models interactive behaviours. In addition to the codebook, the two branches share the same weights, and they can exchange information with the cross-attention module.

reconstruction. Recently, VQ-VAE [62] achieves promising performance in modeling prior knowledge across different modalities [45, 77]. In addition, the discrete representation can be used to measure the certain distance between the predicted results from the prior, which is a significant strength over common VAEs [25, 82] in reconstruction tasks. We therefore design a novel dual-branch VQ-VAE model for modeling the two-person interactions.

Motion state. We found that joint position and velocity are important to learn the interactive behaviour. In addition to the SMPL parameters $\{\theta, \beta, \tau\}$, we further combine the joint positions J, joint velocities $\dot{J}$, and joint positions relative to the counterpart J' in to the motion states for learning the prior. The motion state of a character is denoted as $\mathbf{X} = \left\{\theta, \beta, \tau, J, \dot{J}, J'\right\}$.

Model architecture. Since directly encoding the two-person interactions in a unified latent space with the similar structure as previous priors [32, 78] compromise individual motion fidelity, we thus design a dual-branch network to learn the discrete prior. As shown in Fig. 3, the two branches represent two individuals respectively and can share information with each other. Specifically, the motion states of two characters are first passed through a motion embedding layer to obtain the latent state h_0. We then add a positional encoding on the encoded latent state, and feed it into S transformer blocks to obtain denoised hidden states h_S. Each block consists of two multi-head attention layers followed by one feed-forward network. In the first attention layer, the individual state h_s^a and condition c are first processed with an adaptive layer norm [40], and then the normalized state $\bar{h}_s^a$ go through a multi-head self-attention module:

$$\mathbf{c}_s^a = \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \tag{1}$$

where we set the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ to be:

$$\mathbf{Q} = \bar{h}_s^a \mathbf{W}_{s1}^Q, \quad \mathbf{K} = \bar{h}_s^a \mathbf{W}_{s1}^K, \quad \mathbf{V} = \bar{h}_s^a \mathbf{W}_{s1}^V. \tag{2}$$

$\mathbf{W}^Q$, $\mathbf{W}^K$ and $\mathbf{W}^V$ are projection weights for query, key and value.

The second attention layer has the same structure as the first, but the key and value are the states of the counterpart:

$$\mathbf{Q} = \bar{h}_s^a \mathbf{W}_{s2}^Q, \quad \mathbf{K} = \bar{h}_s^b \mathbf{W}_{s2}^K, \quad \mathbf{V} = \bar{h}_s^b \mathbf{W}_{s2}^V. \tag{3}$$

With this design, one branch can sense the states of the other and generate more plausible interactions. A feed-forward layer is then used to regress the states in the next stage from the output of two attention layers.

With the above encoder, the motion states are mapped to a set of latent codes $\mathbf{Z} = E(\mathbf{X})$. For each latent code $z_i \subset \mathbf{Z}$, we find the closest vectors from corresponding codebook C by calculating the Euclidean distance, *i.e.*:

$$\hat{z}_i = \underset{c_k \in C}{\arg\min} \|z_i - c_k\|_2 . \tag{4}$$

A decoder with the same structure as the encoder is then used to reconstruct the states from sampled latent codes $\hat{\mathbf{Z}}$. In addition to the codebooks, the two branches share the same weights.

Model training. To train the VQ-VAE, we optimize the following objectives:

$$\mathcal{L}_{vq} = \mathcal{L}_{rec} + \alpha \underbrace{\|Z - \text{sg}[\hat{Z}]\|_2}_{\mathcal{L}_{\text{commit}}}, \tag{5}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and α is a hyper-parameter for the commitment loss $\mathcal{L}_{\text{commit}}$. The reconstruction loss is given by:

$$\mathcal{L}_{rec} = \mathcal{L}_{smpl} + \mathcal{L}_{joint} + \mathcal{L}_{vel} + \mathcal{L}_{int}, \tag{6}$$

which is the sum of the supervisions from the SMPL parameters: $\mathcal{L}_{\text{smpl}} = \|[\beta, \theta] - [\hat{\beta}, \hat{\theta}]\|_2^2$, 3D joint positions: $\mathcal{L}_{\text{joint}} = \|J_{3D} - \hat{J_{3D}}\|_2^2$ and velocities: $\mathcal{L}_{\text{vel}} = \|\dot{J}_{3D} - \hat{\dot{J}}_{3D}\|_2^2$ on each character.

We also supervise the relative distance between two characters, *i.e.*:

$$\mathcal{L}_{\text{int}} = \||J_{3D}^a - J_{3D}^b| - |\hat{J}_{3D}^a - \hat{J}_{3D}^b|\|_2^2. \tag{7}$$

To avoid codebook collapse [45], the cookbooks are optimized by exponential moving average (EMA) and codebook reset (Code Reset) operations following [77].

3.3. Initial Interaction Distribution Prediction

To achieve the reconstruction, we use diffusion model as a distribution adaptor, which has shown its powerful performance in generative tasks [47, 73]. However, conventional diffusion model always start from standard Gaussian noises and the early denoising iterations primarily produce noises with limited information [69, 73]. Previous works thus require more diffusion timesteps to generate satisfactory results. Different from text and other high-level signals, images contain more prior knowledge and can relieve the uncertainty.

We thus first regress initial distributions from images for an interactive motion. To break the constraints of limited interaction data and improve the generalization ability, we design the initial distribution prediction network to have a single-person structure. Specifically, we adopt a ViT [7] to extract image features for a single person, and then combine the image features and bounding-box information to regress SMPL parameters with a transformer decoder. To obtain absolute positions for representing interaction, the translation τ of SMPL model is transformed from estimated camera parameters like CLIFF [29]. Subsequently, we train the network on single-person pose estimation datasets with the following loss functions:

$$\mathcal{L}_{init} = \mathcal{L}_{smpl} + \mathcal{L}_{joint} + \mathcal{L}_{reproj}, \tag{8}$$

where $\mathcal{L}_{smpl}$ and $\mathcal{L}_{joint}$ are the SMPL and 3D joint position supervisions defined after Eq. 6. The reprojection loss is given by:

$$\mathcal{L}_{\text{reproj}} = \|\Pi\left(J_{3D} + \tau\right) - \hat{J_{2D}}\|_2^2, \tag{9}$$

where $\Pi(\cdot)$ projects the 3D joints to 2D image with camera parameters, and $\hat{J_{2D}}$ is ground-truth 2D pose.

Although we can use sufficient data to train the single-person network, it still cannot work well on closely interactive cases. The reason is that the current deep learning algorithms cannot identify the belonging of each pixel in complex interaction cases, which induces severe visual ambiguity for the reconstruction. In addition, the inter-person occlusions further increase the uncertainty. We thus use the predicted SMPL parameters x to construct the initial distributions $q\left(x, \sigma\right)$ for the diffusion model. The coarse distribution is then used in the diffusion model to achieve the adaption.

3.4. Interaction Distribution Adaption

With the initial values, we first associate each person across time with predicted poses, positions and 2D bounding-boxes. The procedure is similar to PHALP [44]. The difference is that we do not consider the appearance due to the serious overlapping in the interactive image. After the association, we can obtain two sets of Gaussian distributions $\left\{\mathcal{N}^{a,1:N}, \mathcal{N}^{b,1:N}\right\}$.

Dual-branch diffusion model. We then design a dual-branch diffusion network to refine the initial distributions. As shown in Fig. 2, the diffusion model has the same structure as the interaction prior encoder. After the S transformer blocks, we concatenate the hidden states h_H and bounding-box information to regress pose, shape and translation parameters as in Sec. 3.3. The output parameters are then transformed into distributions for the next diffusion timestep.

Diffusion with initial distributions. To train the diffusion model for pose estimation, previous methods [15, 48] inject time-dependent noises sampled from standard Gaussian distribution to ground-truth motion $\hat{\mathbf{x}}_0$, which can be formulated as:

$$q(\mathbf{x}_t \mid \hat{\mathbf{x}}_0) = \sqrt{\hat{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}\left(0, \mathrm{I}\right), \tag{10}$$

where α_t is a constant hyper-parameter [38], and $\hat{\alpha}_t = \prod_{i=0}^{t} \alpha_i$.

We found that $\mathbf{x}_t$ follows standard Gaussian distribution and the early iterations are meaningless for a human motion, which results in the model to spend more timesteps in the reverse diffusion process. We thus diffuse towards the initial distributions. It should be noted that directly sample noises from the initial distributions will result in a deviant distribution $x_t \sim \mathcal{N}\left(Tx, \sigma\right)$ since x is not equal to zero. Consequently, we change the diffusion process to be:

$$q(\mathbf{x}_t \mid \hat{\mathbf{x}}_0) = \mathbf{x} + \sqrt{\hat{\alpha}_t}(\hat{\mathbf{x}}_0 - \mathbf{x}) + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}\left(0, \sigma\right). \tag{11}$$

The sampled motions that satisfied the discrete interaction prior and physical laws are considered plausible, and the output signals can be used to guide the denoising process. With the above diffusion process, a generative model can be obtained by reversing the process, starting from samples $\mathbf{x}_t \sim \mathcal{N}\left(\mathbf{x}, \sigma\right)$, which is defined as:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\alpha(\mathbf{x}_t, c), \tilde{\beta}_t\sigma), \tag{12}$$

where $\mu_\alpha(\mathbf{x}_t, c)$ is the estimated mean by the diffusion model under the condition of c in timestep $t-1$. $\tilde{\beta}_t$ is the variance calculated by the hyper-parameters β_t, $\hat{\alpha}_t$ and $\hat{\alpha}_{t-1}$.

Proxemics and physics guided denoising. With the diffusion framework, we can gradually refine the initial distributions in each timestep under the condition of image features [13]. However, the image features in interactive cases are always ambiguous, and the model cannot get sufficient information to generate a valid interaction. We design discrete interaction prior to provide additional knowledge for the distribution adaption.

For a sampled interactive motion, we first calculate the character states in each frame and then feed the states to the trained interaction prior. The states of two characters are then encoded to two latent codes by the interaction prior encoder. We can find the closest codes from the codebooks by calculating the Euclidean distance between the encoded latent codes and codebooks. We then recombine the closest latent codes in the codebooks and decode them to a new interaction. By this way, although the initial motion may not be a valid interaction due to the visual ambiguity and occlusion, we can enforce the motion to follow proxemic and interactive behaviors with the trained interaction prior.

The output motions from the prior are then used as the input of diffusion model.

However, the prior only considers the interactive behavior, and cannot guarantee the physical plausibility. Since body contacts are important for human interactions, we further use a physical metric to evaluate the penetration on two-person meshes. Specifically, for the interactive meshes output from the prior, we first detect the set of colliding triangles using bounding volume hierarchies (BVH) [23]. The penetrated vertices are then used to sample the values in local 3D distance fields [60], which reflect the degree of mesh penetrations and collisions. The penetration loss is therefore written as:

$$\mathcal{L}_{pen} = \sum_{(f_a, f_b) \in \mathcal{C}} \Big\{ \sum_{v_a \in f_a} \| -\Psi_{f_b}(v_a)\, n_a \|^2 + \sum_{v_b \in f_b} \| -\Psi_{f_a}(v_b)\, n_b \|^2 \Big\}, \tag{13}$$

where f_a, f_b are two colliding triangles in the detected colliding triangles $\mathcal{C}$. v and n are vertex position and normal, respectively, and $\Psi(\cdot)$ is the distance field. The values are summed as a condition in the reverse diffusion process.

We also require the interaction to be consistent with the image observations. For 3D poses in the current timestep, we calculate the gradient vector of 3D joints to detected 2D keypoints:

$$\mathcal{G} = \frac{\partial \| \Pi(J_{3D}) - p_{2D} \|_2^2}{\partial J_{3D}}. \tag{14}$$

Finally, the condition c of the diffusion model combines the image features, penetration values $\mathcal{L}_{pen}$, and projection loss gradients $\mathcal{G}$.

Model training. We train the diffusion model with the following loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{smpl}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{pen}}, \tag{15}$$

where the loss terms are defined in previous sections. In addition, we randomly mask image features and projection loss gradients in the condition to address the missing detections and visual ambiguities.

4. Experiments

4.1. Datasets

Common datasets. Following previous work, the typical datasets the typical datasets used for training are Human3.6M [20], MPI-INF-3DHP [35], COCO [31], MPII [1]. We use the image and video datasets to train the initial dsitribution prediction network. The video datasets are also used to train the diffusion model by masking the features from the counterpart. **InterHuman [30]** is a large-scale dataset containing diverse two-person interactions. Due to the lack of color images, we use it to train the discrete interaction prior only. **Hi4D [72]** is an accurate multi-view dataset for closely interacting humans. It contains 20 unique pairs of participants with varying body shapes and clothing styles to perform diverse interaction motion sequences. 5 pairs (23, 27, 28, 32, 37) are used as testset, and the rest are used for training. **CHI3D [10]** captures 3 pairs of people in close interaction scenarios with a Vicon MoCap system and 4 additional RGB cameras. We use the standard splits of this dataset. **3DPW [63]** also contains several sequences with two-person interactions. We use these sequences for evaluation only.

4.2. Metrics

We report the Mean Per Joint Position Error (MPJPE), and MPJPE after rigid alignment of the prediction with ground truth using Procrustes Analysis (PA-MPJPE) on these datasets. To measure the mesh quality, the Mean Per Vertex Position Error (MPVPE) is also used. In addition, we further evaluate the interaction error, which is defined as $\frac{1}{K \times K} \sum_{i=1}^{K} \sum_{j=1}^{K} | \|J_i^a - J_j^b\| - \|\hat{J}_i^a - \hat{J}_j^b\| |$, and K is the number of joints.

4.3. Comparison to State-of-the-Art Methods

We conduct several experiments to demonstrate the effectiveness of our method in closely interactive scenarios. We choose the current SOTA single-person and multi-person mesh recovery models as baseline methods. Human4D and CLLIF are designed for single-person scenarios. They can achieve high joint accuracy, but do not consider the interactions. BEV and GroupRec aim to multi-person cases and incorporate spatial constraints into the reconstruction. In addition, BUDDI is the most relevant work to ours, which reconstruct interactive humans from single images with a optimization. Tab. 1 shows a quantitative comparison with these baseline methods on Hi4D dataset. All methods are finetuned on Hi4D training set for a fair comparison. Since Human4D uses a weak-perspective camera model, it cannot obtain accurate absolute positions in camera coordinates. Although Human4D can predict precise root-relative joint positions, it cannot be used to reconstruct interactive pairs. Different from Human4D, CLIFF estimates humans in the original camera coordinates and can produce valid relative positions for interactive humans. However, it processes each human iteratively and results in a high interaction error due to depth ambiguity. We also compare our method with BEV and GroupRec. They constrain the spatial positions of humans, but the constraint can only work well on relative large scenes. Besides, BUDDI trains a proxemic prior to fit human models to interactive images. However, the optimziation is not robust to depth ambiguity and 2D pose noises. In Fig. 4, we find that BUDDI overfits to the noisy

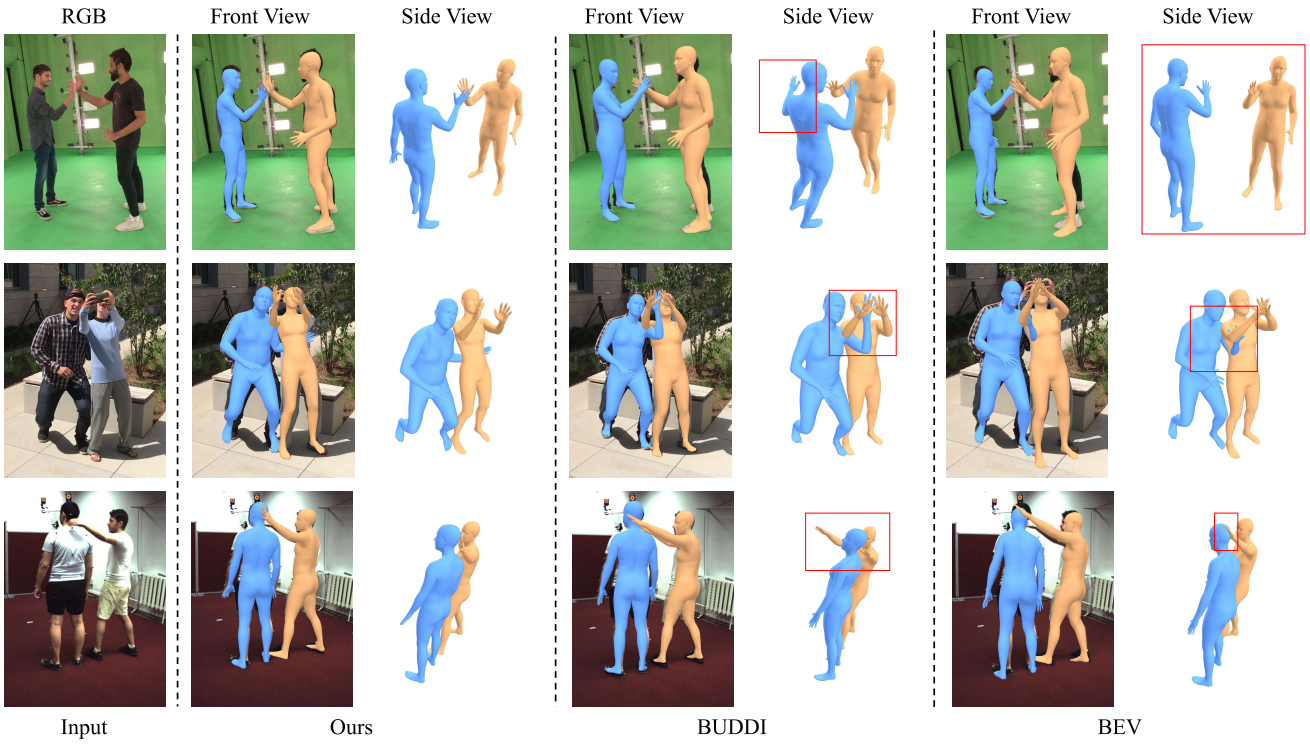


Figure 4. Qualitative comparison with BUDDI [36] and BEV [57]. Our method can produce more accurate body poses and physical contacts.

Method	MPJPE	PA-MPJPE	MPVPE	Interaction
Human4D [12]	72.1	52.4	88.6	–
CLLIF [29]	91.3	53.6	109.6	141.5
BEV [57]	91.8	52.2	101.2	131.0
GroupRec [18]	82.4	51.6	88.6	98.8
BUDDI [36]	96.8	70.6	116.0	102.6
Ours	**63.1**	**47.5**	**76.4**	**81.4**

Table 1. **Comparisons on Hi4D.** Our method can outperform existing single-person and multi-person approaches in terms of joint position accuracy and interaction state. "–" means the results are not available.

Method	MPJPE	PA-MPJPE	MPVPE	Interaction
Human4D [12]	72.9	49.1	**107.0**	–
BEV [57]	78.3	**48.5**	116.9	136.4
GroupRec [18]	73.3	48.7	109.4	110.6
BUDDI [36]	83.6	53.6	126.1	113.1
Ours	**70.6**	51.4	107.9	**100.3**

Table 2. **Comparisons on 3DPW.** Our method can achieve competitive performance in terms of joint accuracy in outdoor scenarios. In addition, we still produce better interactions without training on in-the-wild video data.

2D poses and produce a wrong result. Although it is designed for close interaction reconstruction, its performance is strongly affected by the 2D pose detection. Thus, it is still inferior to other baseline methods. In addition, its optimization framework is time-consuming, which takes $3 \sim 4$ min to reconstruct a pair. In contrast, our method takes the detected 2D poses as one of conditions and is more robust to detection noises due to the random masking strategy. With the assistance of the discrete prior, our method can achieve the SOTA in closely interactive cases.

We also conduct experiments on 3DPW datasets. We select all interactive sequences as a benchmark to evaluate our method. Tab. 2 shows the results of baseline methods and ours. Although the data used for training the diffusion model are captured in indoor scenes, our method can also predict satisfactory results in outdoor scenarios. The results reveal that our method can achieve better human interactions than previous works.

We further conduct a qualitative comparison on CHI3D in Fig. 4. Since BUDDI is a two-stage method, its performance strongly depend on the quality of 2D poses. However, the 2D pose detection always fails in close interactive scenarios. Consequently, it may produce incorrect body poses. In addition, BEV showes inter-person penetrations due to the lack of physical constraints. In contrast, our method can get better performance with the assistance of interaction prior and physical guidance.

4.4. Ablation Study

Discrete interaction prior. Reconstructing closely interactive humans from single-view images is a highly ill-posed problem due to the inter-person occlusions. The image can provide limited valid information for the reconstruc-

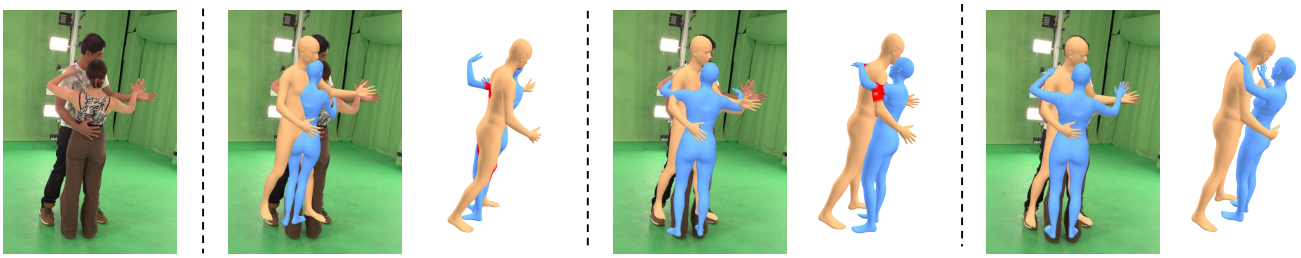


Figure 5. Ablation study. The initial prediction is severely affected by visual ambiguity and cannot reconstruct natural interaction. The interaction prior update the motions from initial prediction with proxemic behaviours. Although the prior can produce better interaction, it still suffers from inter-person penetrations. With the distribution adaption, our method can refine the results, and reconstruct accurate interaction and physical contacts.

Method	MPJPE	PA-MPJPE	MPVPE	Interaction
Initial Prediction	73.7	55.8	88.4	93.7
Single Branch	64.1	48.6	81.5	89.3
w/o prior	67.6	50.3	83.2	90.1
w/o physical guidance	63.6	**47.0**	**76.1**	86.1
Ours	**63.1**	47.5	76.4	**81.4**

Table 3. **Ablations on Hi4D.** "Single Branch" uses a single-person prior, and the two characters cannot share information during the inference. "w/o prior" and "w/o physical guidance" denote our model without the discrete prior and physical guidance.

tion. The interaction prior models the interactive behaviours in discrete codebooks and can incorporate proxemic prior knowledge into the framework. For a noisy initial prediction, which may not follow the interactive behaviours, we find the closet interaction in the codebooks to refine the results. In Fig. 5, we found that the incorrect predicted results can be adjusted with the prior, and the output of the prior can be a valid interaction to be used in the reverse diffusion process. Even for the severely occluded cases, the prior can still produce plausible 3D poses with the interaction relationships. In Tab. 3, we found that the model performance decrease Without the prior.

Physical guidance. Although the prior can provide interactive behaviours to assist the reconstruction, the output results still suffer from penetrations. We thus provide a BVH method to detect colliding triangles and calculate the penetration loss with local distance field. Different from previous works that directly use the constraint as a supervision, we also incorporate the loss values as a condition in the reverse diffusion process. In each timestep, the model can sense the penetration and refine the results in the next step. In Fig. 5, we can observe that the penetration is alleviated with the physical guidance. Although the model inference does not obey physical laws, the model can get feedback of the penetration with this strategy and thus produces better results compared to direct supervision.

Dual-branch diffusion model. To investigate the importance of interaction, we use a single-branch model to replace the dual-branch structure in discrete prior and diffusion model. In the prior, we use a single codebook to model human motions, which is similar to [77]. During the distribution adaption process, the initial motions is refined using the same codebook and not share information between two characters. Although it can also predict accurate relative joint and vertex positions, the interaction error is large due to the lack of proxemics.

5. Limitation and Future Work

Although our method can reconstruct interactive humans from single-view videos, there still exist some limitations. First, the current design can only support two-person interactions. If the setting is extended to reconstruct social interactions with more people, our method has to maintain more codebooks and may result in redundancy. In the future, the network can be improved to aggregate similar motions in the same codebook for more general social interactions. Second, although we use in-the-wild image datasets to train the initial prediction network, our model performance may still deteriorate in outdoor scenarios. As a result, to build a in-the-wild dataset for close interaction is also a promising direction in the future.

6. Conclusion

In this work, we introduce a novel dual-branch diffusion model to incorporate proxemic behavior, physical constraint, and image observation to reconstruct close interactive humans. To alleviate the impact of occlusions and visual ambiguities, the proposed discrete prior encodes human close interactions in two codebooks and can provide additional knowledge for the reconstruction. We further formulate the reconstruction as a distribution adaption to consider the uncertainty of the ill-posed problem. Finally, the model refine the distribution under the guidance of observations in each reverse diffusion timestep, and can produce accurate and realistic interactions.

Acknowledgement. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-024) and China Scholarship Council under Grant Number 202306090192.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 6

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[3] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, pages 660–677, 2022. 2

[4] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, pages 1475–1484, 2022. 1, 2

[5] Jeongjun Choi, Dongseok Shim, and H Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *arXiv preprint arXiv:2212.02796*, 2022. 2

[6] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *TMM*, 2023. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[8] Rong et al. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 2

[9] Yanwen Fang, Chao Li, Jintai Chen, Pengtao Jiang, Yifeng Geng, Xuansong Xie, Eddy KF LAM, and Guodong Li. Pgformer: Proxy-bridged game transformer for multi-person extremely interactive motion prediction. *arXiv preprint arXiv:2306.03374*, 2023. 3

[10] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 2, 6

[11] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023. 2

[12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023. 2, 7

[13] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, pages 13041–13051, 2023. 2, 5

[14] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *CVPR*, pages 13053–13064, 2022. 3

[15] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *ICCV*, pages 15977–15987, 2023. 2, 5

[16] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *TPAMI*, 45(4):5010–5026, 2022. 2

[17] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior. *TIP*, 31:4679–4692, 2022. 1, 2

[18] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning. In *ICCV*, pages 14873–14883, 2023. 1, 2, 7

[19] Buzhen Huang, Jingyi Ju, and Yangang Wang. Crowdrec: 3d crowd reconstruction from single color images. *arXiv preprint arXiv:2310.06332*, 2023. 1

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 6

[21] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 1, 2

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2

[23] Tero Karras. Maximizing parallelism in the construction of bvhs, octrees, and k-d trees. In *Proceedings of the Fourth ACM SIGGRAPH / Eurographics Conference on High-Performance Graphics*, pages 33–37, 2012. 6

[24] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, pages 1715–1725, 2022. 2

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[27] Haoyuan Li, Haoye Dong, Hanchao Jia, Dong Huang, Michael C Kampffmeyer, Liang Lin, and Xiaodan Liang. Coordinate transformer: Achieving single-stage multi-person mesh recovery from videos. In *ICCV*, pages 8744–8753, 2023. 2

[28] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 2

[29] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022. 2, 5, 7

[30] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 3, 6

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[32] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *TOG*, 39 (4):40–1, 2020. 2, 4

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017. 6

[36] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 1, 2, 7

[37] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *CVPR*, pages 14474–14483, 2021. 2

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 5

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 4

[41] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *CVPR*, pages 17121–17130, 2023. 3

[42] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, pages 21254–21263, 2023. 2

[43] Muhammad Rameez Ur Rahman, Luca Scofano, Edoardo De Matteis, Alessandro Flaborea, Alessio Sampieri, and Fabio Galasso. Best practices for 2-body pose forecasting. In *CVPR*, pages 3613–3623, 2023. 3

[44] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *CVPR*, pages 2740–2749, 2022. 5

[45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019. 4, 1

[46] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, pages 11488–11499, 2021. 2

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4

[48] Cédric Rommel, Eduardo Valle, Mickaël Chen, Souhaiel Khalfaoui, Renaud Marlet, Matthieu Cord, and Patrick Pérez. Diffhpe: Robust, coherent 3d human pose lifting with diffusion. In *ICCV*, pages 3220–3229, 2023. 5

[49] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[50] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *ICCV*, pages 14725–14737, 2023. 2

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[52] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *TOG*, 39(4):54–1, 2020. 3

[53] Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. Neural animation layering for synthesizing martial arts movements. *TOG*, 40(4):1–16, 2021. 3

[54] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *TOG*, 41(4):1–13, 2022. 2

[55] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 2

[56] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 1, 2

[57] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 1, 2, 7

[58] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *ICCV*, pages 9601–9611, 2023. 3

[59] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[60] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 6

[61] Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *3DV*, pages 53–63, 2021. 2

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 2, 4

[63] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 6

[64] Hao Wen, Jing Huang, Huili Cui, Haozhe Lin, Yu-Kun Lai, Lu Fang, and Kun Li. Crowd3d: Towards hundreds of people

reconstruction from a single image. In *CVPR*, pages 8937–8946, 2023. 1, 2

[65] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Control strategies for physically simulated characters performing two-player competitive sports. *TOG*, 40(4):1–11, 2021. 3

[66] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, pages 2228–2238, 2023. 3

[67] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. Joint-relation transformer for multi-person motion prediction. In *ICCV*, pages 9816–9826, 2023. 3

[68] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *ICLR*, 2022. 3

[69] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, pages 14928–14940, 2023. 2, 4

[70] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1

[71] Kangxue Yin, Hui Huang, Edmond SL Ho, Hao Wang, Taku Komura, Daniel Cohen-Or, and Hao Zhang. A sampling approach to generating closely interacting 3d pose-pairs from 2d annotations. *TVCG*, 25(6):2217–2227, 2018. 3

[72] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 6

[73] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, pages 16010–16021, 2023. 2, 4

[74] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 1, 2

[75] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *NeurIPS*, 31, 2018. 2

[76] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556, 2021. 2

[77] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 4, 8, 1

[78] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *ICCV*, pages 11343–11353, 2021. 2, 4

[79] Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. Simulation and retargeting of complex multi-character interactions. *arXiv preprint arXiv:2305.20041*, 2023. 3

[80] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021. 2

[81] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 3

[82] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, pages 9054–9064, 2023. 4

Closely Interactive Human Reconstruction with Proxemics and Physics-Guided Adaption

Supplementary Material

In the supplementary material, we first provide the details of VQ-VAE, motion representation, and training procedures to help the reproduction of the experimental results. More comparisons, analyses, and qualitative experiments are also conducted to further demonstrate the superiority of the proposed method. Finally, we show some failure cases to illustrate the limitations of the current method. We also provide a supplementary video for this work.

7. VQ-VAE

A naive training of VQ-VAE suffers from codebook collapse [45]. To avoid the limitation, we adopt exponential moving average (EMA) and codebook reset (Code Reset) to improve the codebook utilization. Specifically, EMA updates the codebook smoothly: $C^t \leftarrow \lambda C^{t-1} + (1-\lambda)C^t$, and C^t is the codebook in the current iteration. $\lambda = 0.99$ is a exponential moving constant. Code Reset finds inactivate codes during the training and reassigns them according to input data.

For the model architecture, the encoder of the discrete interaction prior consists of a motion embedding layer, a positional encoding layer, and 4 transformer blocks. The decoder has a motion decoding layer and 4 transformer blocks. Each codebook has a size of 256×256.

8. Motion representation

Previous works [77] always represent human motion in a canonical space, and the global rotation and translation are obtained by accumulating local angular and linear velocities. This representation cannot be directly applied to multi-person scenarios since it does not maintain the person-to-person spatial relationships. To address this problem, we use the root position of character a in the first frame as origin and transform the interactive motions to the new coordinate. Consequently, the joint positions and velocities are kept in the world frame. In addition, the two-person interactions satisfy commutative property, which means the interaction $\{\mathbf{x}^a, \mathbf{x}^b\}$ and $\{\mathbf{x}^b, \mathbf{x}^a\}$ are equivalent.

9. Implementation details

The diffusion model has the same stracuture as the VQ-VAE encoder, which contains a motion embedding layer, a positional encoding layer, and 4 transformer blocks. The number of diffusion timesteps is set to 100 during the training stage. In the inference, we adopt DDIM sampling strategy [51] with 5 timesteps to achieve the distribution adaption. For the projection loss gradients, we use ViT pose [70] to predict 2D poses. To train the diffusion model, 25% of projection loss gradients and image features are randomly masked. The frame length of interactive motions is 16, and the batch size for VQ-VAE and diffusion model are 256 and 32, respectively. All the models are trained with AdamW [34] optimizer using a learning rate of 1e-4 on a single GPU of NVIDIA GeForce RTX 4090.

Dataset	w/o proj. loss	Ours
3DPW	73.8	70.6
Hi4D	64.2	63.1

Table 4. **Ablation on projection loss gradients.** "w/o proj. loss" denotes our model without the projection loss gradients. The number are MPJPE.

Method	Accel↓	A-PD↓
GroupRec	25.2	1.34
Ours	10.7	1.15

Table 5. Accel and A-PD are acceleration error and average penetration depth, respectively.

Timesteps	step=1	step=3	step=5	step=10
MPJPE	68.2	64.4	63.1	63.0

Table 6. **Ablation on number of duffusion timesteps.** The ablation is conducted on Hi4D.

10. Extended experiments

Projection loss gradients. We also investigate the projection loss gradients, which provide signals to guide the reverse diffusion process and enforce the 3D models to be consistent with image observations. The term can improve joint accuracy for common scenarios. However, it has limited effect on severely occluded cases since the current state-of-the-art 2D pose detectors cannot produce reliable results for ambiguous images. In Tab. 4, Hi4d has more complex interactions than 3DPW, and the performance on 3DPW dataset significantly decreases without the projection loss gradients.

Diffusion guidance. In Fig. 6, we show the intermediate results in each timestep during the inference phase. The inference contains 5 timesteps with the denoising diffusion implicit models [51]. Although the reverse diffusion has the same timesteps, we find that the denoising model without the guidance produces slight changes and the final results still show severe penetrations. In contrast, our model can refine the interactions and alleviate penetrations, which

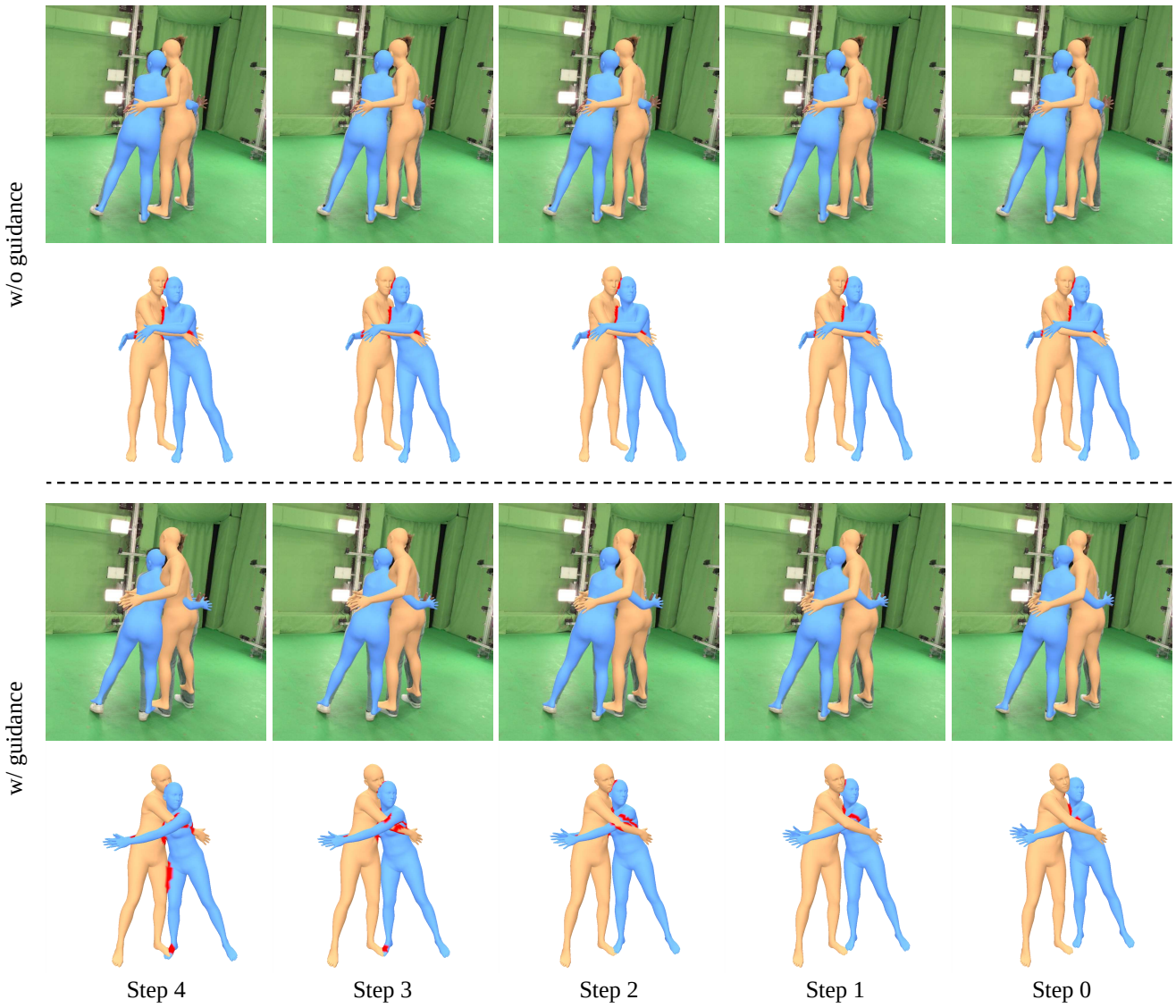


Figure 6. Comparison between the models with and without the guidance. Without the guidance, the reverse diffusion process has limited effect in improving the interations.

demonstrates the importance of proposed guidance.

Diffusion timesteps. We analyze the impact of different timesteps in Tab. 6. The performance increases with more timesteps at first and then becomes stable. To balance the accuracy and efficiency, we use 5 timesteps in the inference phase.

Penetration and acceleration error. In Tab. 5, we further use the average penetration depth (A-PD) [8], which reflects the degrees of inter-penetration, to evaluate the body contact and penetration, and our method can produce better performance. Our method also outperforms GroupRec [18] on the acceleration error due to the proposed velocity loss and temporal architecture.

More results. We also show more qualitative results in Fig. 7. Our method can reconstruct closely interactive humans with plausible body poses, natural proxemic relationships and accurate physical contacts from single-view inputs.

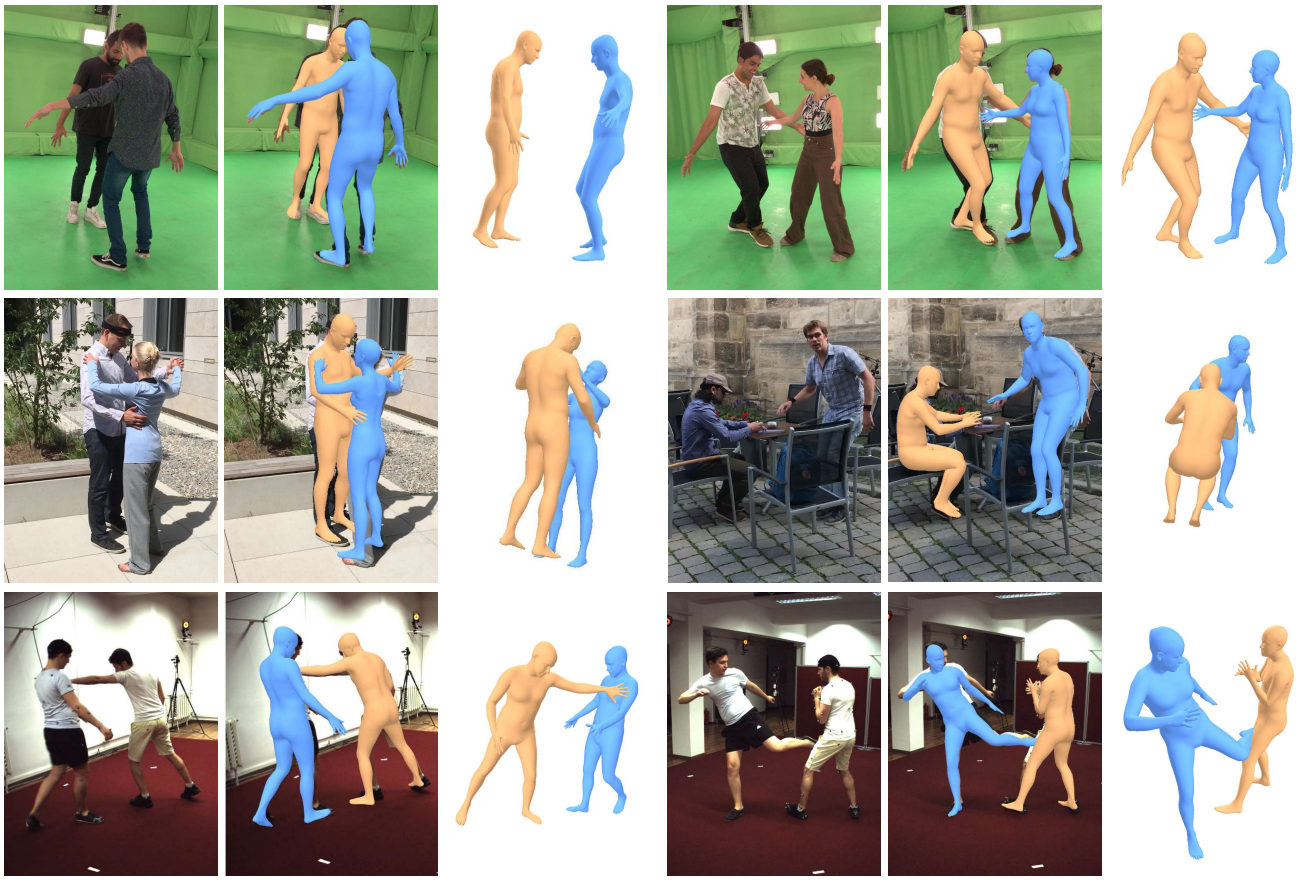


Figure 7. More qualitative results on Hi4D, 3DPW and CHI3D datasets. Our method can reconstruct closely interactive humans with plausible body poses, natural proxemic relationships and accurate physical contacts from single-view inputs