Supplementary Material:
Capturing Closely Interacted Two-Person Motions with Reaction Priors

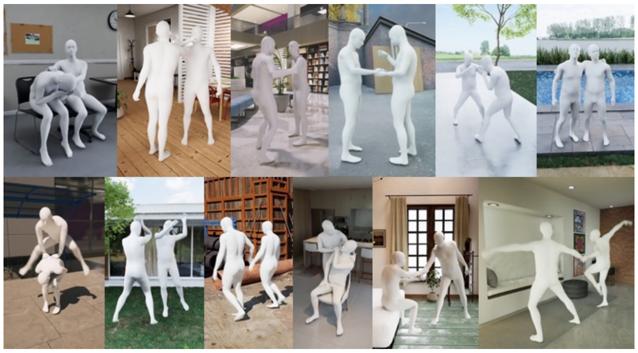

Figure 1. **Sample Poses in Dual-Human.**

In this supplementary material, we provide the statistics and the data processing of Dual-Human (Sec. 1). We also describe the network architectures and running time in details (Sec. 2). What's more, we discuss the social ethics and impact of our dataset (Sec. 3). Additionally, we show some qualitative results in Fig. 4 and the **supplementary video**.

1. Dual-Human Dataset

The preview of our Dual-Human dataset can be found in the supplementary video and Fig. 1. We describe the statistics as well as the details of model fitting here.

1.1. Statistics

The statistics of our dataset is shown in Table 1. Dual-Human has around $2k$ two-person motions with a total duration of 3.05 hours, covering 3 major categories (daily motions, dance, sports, Fig. 2) and nearly 70 subcategories. Specifically, daily motions include 'hug', 'push/pull', 'handshake', 'comfort others', *etc*. Dance mainly includes motions in various ballroom dances. Sports include 'basketball', 'soccer', 'boxing', 'sit-ups', *etc*. There are 118 scenes for the rendering of Dual-Human, including 20 3D scenes and 98 HDRI scenes, which guarantees the variety of backgrounds. We use some assets (*e.g.*, textures, cloths) from BEDLAM [1]. We simulate cloths with Marvelous Designer and render images with Unreal Engine 5.3.

Motions	Scenes	Subjects	Skins	Cloths	Textures
2019	118	40	100	52	1076

Table 1. **Statistics of Dual-Human**.

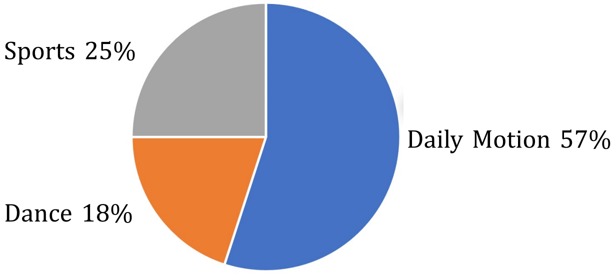

Figure 2. **Distribution of Motions in Dual-Human.**

1.2. Model Fitting

We fit the parametric human model SMPL-X [7] to Xsens 3D joints and extract contact annotations automatically.

We use the common 3D joint loss and regularization losses for poses and shapes as follows:

$$\mathcal{L}_{3d} = \sum_t \rho(\mathcal{J}(\boldsymbol{\theta}_t, \boldsymbol{\beta}) - \boldsymbol{J}_t), \tag{1}$$

$$\mathcal{L}_{preg} = \sum_t ||\boldsymbol{\theta}_t||_2^2, \tag{2}$$

$$\mathcal{L}_{sreg} = ||\boldsymbol{\beta}||_2^2. \tag{3}$$

Besides, the velocities of adjacent frames should be close:

$$\mathcal{L}_{smooth} = \sum_t ||\dot{\boldsymbol{J}}_t - \dot{\boldsymbol{J}}_{t-1}||_2^2. \tag{4}$$

To deal with penetration, we adopt the interpenetration loss in [6], whose effect can be validated from Fig. 3. Specifically, a modified Signed Distance Field (SDF) is defined as follows:

$$\phi(x, y, z) = -\min(SDF(x, y, z), 0), \tag{5}$$

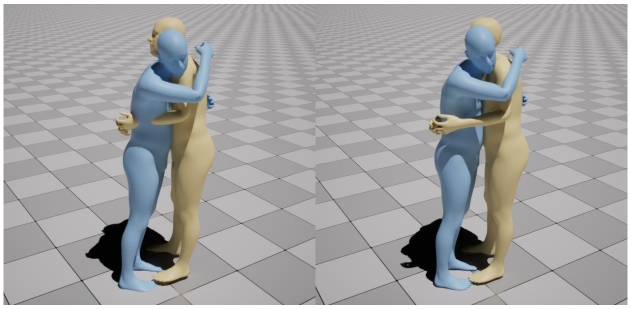

Figure 3. **Comparison**: (a) w/o interpenetration loss; (b) w/ interpenetration loss.

which is positive inside the human and 0 outside the human. ϕ is defined on voxelized human meshes M and prevents the collision through the following loss:

$$\mathcal{L}_{sdf} = \rho(\sum_{v \in M_a} \phi_r(v)) + \rho(\sum_{v \in M_r} \phi_a(v)). \tag{6}$$

The final loss is:

$$\begin{aligned}\mathcal{L}_{fit} =& \lambda_{3d}\mathcal{L}_{3d} + \lambda_{preg}\mathcal{L}_{preg} + \lambda_{sreg}\mathcal{L}_{sreg} + \quad (7)\\ & \lambda_{smooth}\mathcal{L}_{smooth} + \lambda_{sdf}\mathcal{L}_{sdf}, \quad (8)\end{aligned}$$

where $\lambda_{3d} = 5$, $\lambda_{preg} = \lambda_{sreg} = 0.01$, $\lambda_{smooth} = 5$, $\lambda_{sdf} = 0.0001$.

Furthermore, as the annotation process for our large motion data is tedious, we obtain the contact vertex indices based on the SDF value.

2. Network Architecture

In this section, we detail the motion representation, as well as the network structures of motion VAE, reaction INN, pose estimator, and SmoothNet. We also report the running time of each module.

Motion Representation: Before illustrating the network architectures, we first introduce the motion representation in reaction priors. Following [5], we convert 3D human joints to 263-dim hybrid representations as follows:

$$\boldsymbol{x} = [\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \boldsymbol{j}^p, \boldsymbol{j}^v, \boldsymbol{j}^r, \boldsymbol{c}^f], \tag{9}$$

where $\dot{r}^a \in \mathbb{R}$ is the root angular velocity along Y-axis; $\dot{r}^x \in \mathbb{R}$ and $\dot{r}^z \in \mathbb{R}$ are the root linear velocities along X-axis and Z-axis, respectively; $r^y \in \mathbb{R}$ is the root height; $\boldsymbol{j}^p \in \mathbb{R}^{3(N_j-1)}$, $\boldsymbol{j}^v \in \mathbb{R}^{3N_j}$, and $\boldsymbol{j}^r \in \mathbb{R}^{6(N_j-1)}$ are local joint positions, velocities, and local rotations; $\boldsymbol{c}^f \in \mathbb{R}^4$ is the binary foot-ground contact. Here $N_j = 22$.

Additionally, different from single-person motions that are in canonical coordinates, two-person motions should preserve the relative direction and translation of two people. Therefore, we expand the original 263-dim to 269-dim with the initial direction and translation. Note that for motion VAE of a single person, we only use the first 263-dim, and the last 6-dim is used to recover the two-person motions.

Motion VAE: We follow the Transformer VAE structure in [2, 8], as shown in the Fig. 3 in the main manuscript. The linear layer in the encoder converts the motion representation to the latent representation of dimension 1×256. The distribution tokens have the dimension of 512 and can be divided into respective 256-dimensional mean and variance of the latent distribution. The number of layers and heads are 9 and 4 for both the encoder and decoder, respectively. The feed forward networks are 1024-dimensional.

Reaction INN: We first introduce the preliminary knowledge of INN. In order to explain it easier, we borrow some concepts from the normalizing flow, which is one of the popular INNs.

Given a data variable $x \in X$, a prior probability distribution p_Z about a latent variable $z \in Z$, and a bijection $f : X \rightarrow Z$, the model distribution on X can be defined via the change of variable formula as follows:

$$p_X(x) = p_Z(f(x))|\det(\frac{\partial f(x)}{\partial x^T})|, \tag{10}$$

where $\frac{\partial f(x)}{\partial x^T}$ is the Jacobian of f at x. Therefore, as long as the Jacobian of f is invertible and well-designed, a bijective model that is both tractable and extremely flexible can be constructed.

Our reaction INN is constructed following RealNVP [3]. We adopt 8 affine coupling layers and each includes 2-layer MLPs with Leaky ReLU for both the scale network s and the translation network t. Specifically, the feature is first divided into $[\boldsymbol{u}_1, \boldsymbol{u}_2]$, and the operations in the affine coupling layer are as follows to obtain the transformed feature $[\boldsymbol{u}'_1, \boldsymbol{u}'_2]$:

$$\boldsymbol{u}'_1 = \boldsymbol{u}_1 \tag{11}$$

$$\boldsymbol{u}'_2 = \boldsymbol{u}_2 \odot \exp(s(\boldsymbol{u}_1)) + t(\boldsymbol{u}_1). \tag{12}$$

After one layer, $[\boldsymbol{u}'_1, \boldsymbol{u}'_2]$ is swapped to $[\boldsymbol{u}'_2, \boldsymbol{u}'_1]$ and fed to the next layer for more powerful capability.

The Jacobian of the affine coupling layer is:

$$\frac{\partial \boldsymbol{u}'}{\partial \boldsymbol{u}^T} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \frac{\partial \boldsymbol{u}'_2}{\partial \boldsymbol{u}_1} & \mathrm{diag}(\exp(s(\boldsymbol{u}_1))) \end{bmatrix}, \tag{13}$$

from which we know its determinant is $\exp(\sum s(\boldsymbol{u}_1))$ and is irrelevant to the Jacobian of s and t. Therefore we can choose s and t with arbitrary complexity.

In our implementation, the input and output (mean and variance) to the reaction INN both have the dimension of 512. The structure is simple and lightweight.

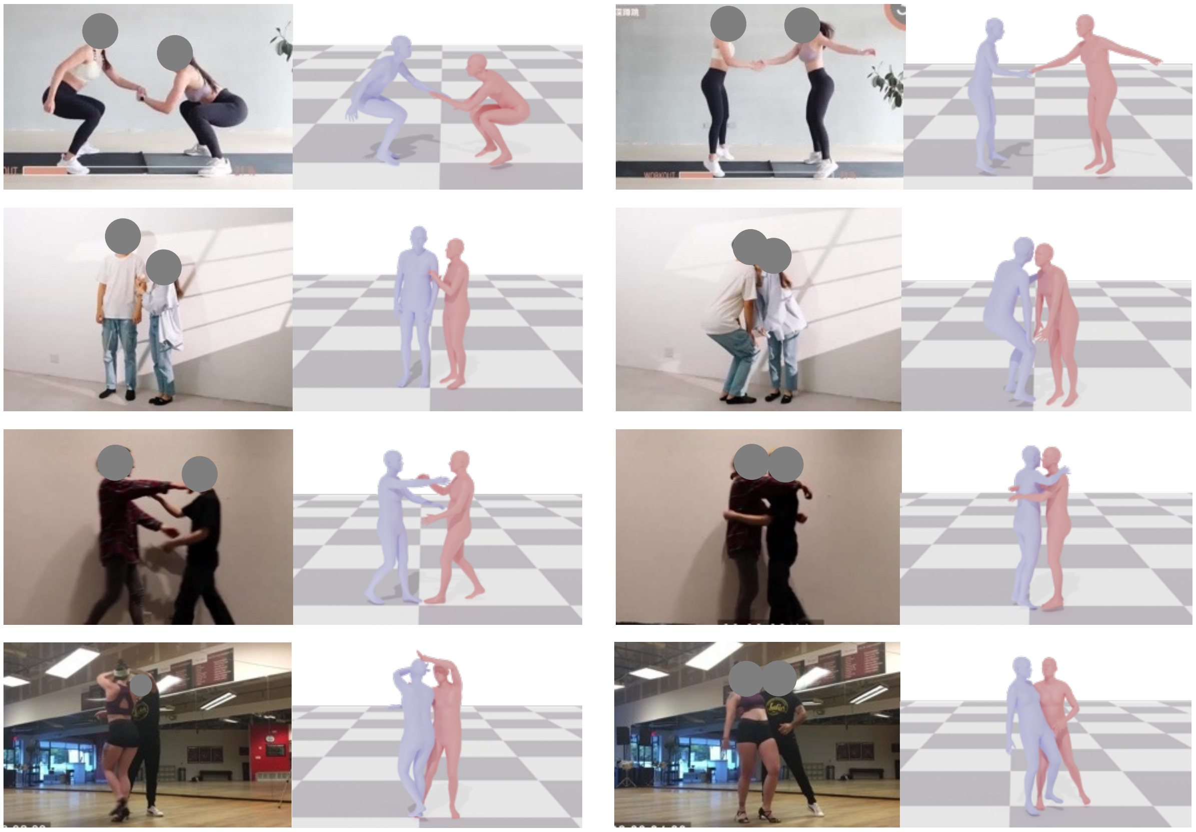

Figure 4. **Qualitative Results on in-the-wild Images.**

Pose Estimator: The pose estimator includes a feature extractor, a Transformer decoder, and final MLPs, which is similar to 4D-Humans [4]. However, instead of cropping the single person, we use the whole image (512×512) as input. HRNet-W48 [9] is adopted to extract the image features ($2048 \times 16 \times 16$). After that, the Transformer decoder (6 layers, 8 heads, 1024 feed forward dimension) with interaction-aware self-attention queries the information of human poses, translations and probability from the image features. The final MLPs map the decoded features to the corresponding variables (63-dim poses, 3-dim translations and 1-dim probability). The translation estimation from the previous frame as the 'Track' query is concatenated with other queries of the current frame to guide the attention.

SmoothNet: We adopt SmoothNet [10] to impose temporal constraints. SmoothNet has an encoder layer, a decoder layer and 5 middle blocks. The encoder and decoder both have 1 layer. Each block contains two linear layers with LeakyReLU, dropout and skip connections. The input window length is 64 and the hidden size is 512. The output has the same dimension as the input.

Running Time: The running times on an RTX 2080ti GPU (batch size 1) of the pose estimator, the SmoothNet, the motion VAE, and the reaction INN are 69ms, 1ms, 20ms, and 3ms, respectively.

3. Social Ethics and Impact

The performers freely volunteer to participate in the motion collection. Since we only capture the motions and use synthetic images, the personal privacy is fully considered.

As for the social impact, our dataset is promising to promote further researches in interacted human motion capture and generation.

References

[1] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 1

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023. 2

[3] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2016. 2

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 3

[5] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 2

[6] 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

[7] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA 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. 1

[8] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 2

[9] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3

[10] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *ECCV*, pages 625–642. Springer, 2022. 3