Ponimator: Unfolding Interactive Pose for Versatile Human-human Interaction Animation

Shaowei Liu* Chuan Guo[2]† Bing Zhou[2]† Jian Wang[2]†
[1]University of Illinois Urbana-Champaign [2]Snap Inc.
https://stevenlsw.github.io/ponimator/

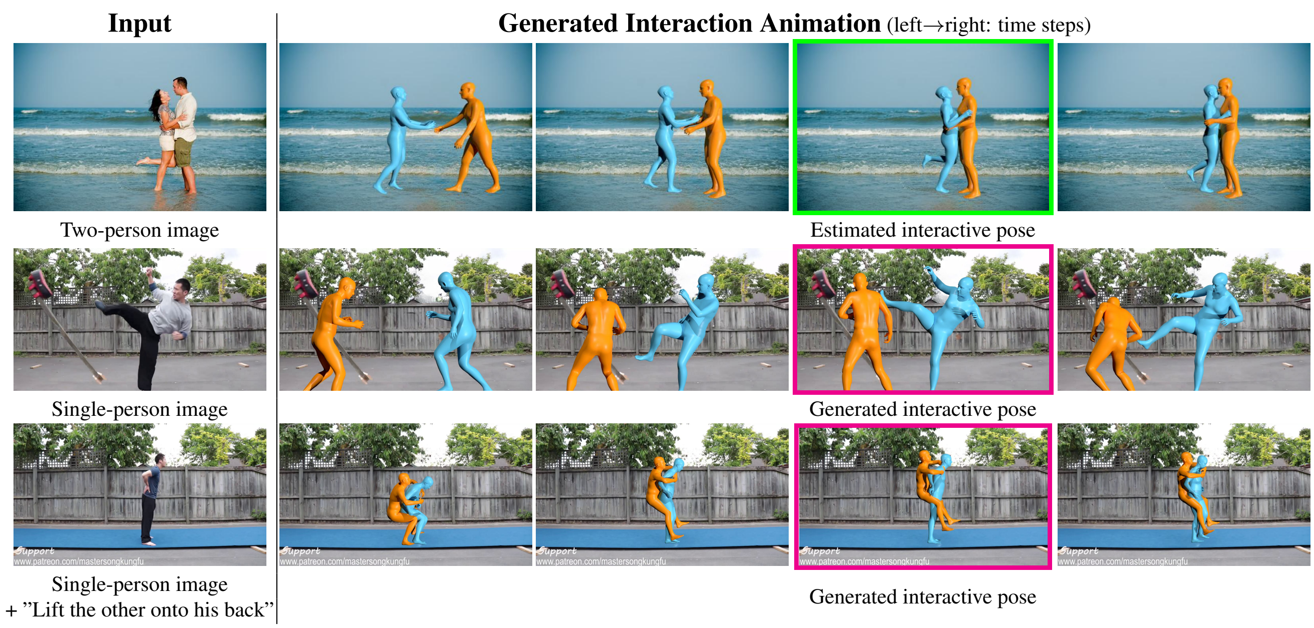

Figure 1. Ponimator enables versatile interaction animation applications anchored on *interactive poses*. For two-person images (top), Ponimator generates contextual dynamics from estimated interactive poses (green box). For single-person images (middle) with optional text prompts (bottom), Ponimator first generates partner interactive poses (magenta box) and then fulfill the interaction dynamics.

Abstract

Close-proximity human-human interactive poses convey rich contextual information about interaction dynamics. Given such poses, humans can intuitively infer the context and anticipate possible past and future dynamics, drawing on strong priors of human behavior. Inspired by this observation, we propose Ponimator, a simple framework anchored on proximal interactive poses for versatile interaction animation. Our training data consists of close-contact two-person poses and their surrounding temporal context from motion-capture interaction datasets. Leveraging interactive pose priors, Ponimator employs two conditional diffusion models: (1) a pose animator that uses the temporal prior to generate dynamic motion sequences from interactive poses, and (2) a pose generator that applies the spatial prior to synthesize interactive poses from a single pose, text, or both when interactive poses are unavailable. Collectively, Ponimator supports diverse tasks, including image-based interaction animation, reaction animation, and text-to-interaction synthesis, facilitating the transfer of interaction knowledge from high-quality mocap data to open-world scenarios. Empirical experiments across diverse datasets and applications demonstrate the universality of the pose prior and the effectiveness and robustness of our framework. Codes and video visualization can be found at https://stevenlsw.github.io/ponimator/

*Work done at an internship at Snap Research NYC, Snap Inc.
†Co-corresponding author

1. Introduction

The interplay between humans plays a crucial role in our daily lives. These interactions convey key social signals that reflect relationships and intentions. For example, a simple hug typically expresses closeness, a handshake serves as a formal greeting, while combat indicates opposing stances. A key observation is that interactive poses in close proximity (e.g., handshake) carry rich prior information about interaction dynamics. Specifically, a pair of such poses reveals contextual cues about spatial relationships, constraints, and intent, often suggesting probable ranges of past and future motions. These interactive poses can act as a bridge for modeling interaction dynamics with reduced complexity while inherently preserving prior knowledge of close interactions.

In this paper, we present *Ponimator*, a novel framework that leverages the dynamics priors embedded in interactive poses through a generative model, demonstrating its versatility across various interaction animation tasks. We develop this interaction prior using a combination of two high-quality human-human interaction datasets: Inter-X [60] and Dual-Human [7]. From these datasets, we construct a collection of two-person poses in close proximity, as shown in Fig. 2, along with their preceding and subsequent interaction motions. Using this collection, we train a conditional diffusion model to generate contextual interaction dynamics given a pair of closely interactive poses.

We first demonstrate the application of our learned pose-to-dynamic interactive priors for open-domain images. Social interactions are frequently depicted in images, yet existing works [7, 9, 10, 36] typically focus only on reconstructing static interactive poses, lacking the temporal dynamics of these interactions. Meanwhile, video diffusion models [3, 16, 18] can animate images over time but often struggle to maintain motion and interaction integrity. In contrast, Ponimator seamlessly transfers learned interaction prior knowledge from high-quality 3D mocap datasets to these in-the-wild images through estimated interactive poses, as shown in Fig. 1 (top). For broader applications, we developed an additional conditional diffusion model that leverages the spatial prior to generate interactive poses from multiple input types, including text descriptions, single poses, or both. Thus, when only a single person appears in an image, Ponimator can first generate a partner pose with an optional text prompt, and then animate the interactive poses over time (see Fig. 1). Furthermore, by anchoring on these interactive poses, Ponimator is able to generate short-clip two-person motions with proximal contact (see Fig. 8) directly from text input.

Our key contributions are summarized as follows: 1) We present Ponimator, a simple framework designed to learn the dynamics prior of interactive poses from motion capture data, particularly focusing on proximal human-human interaction animations; 2) The learned prior is universal and generalizes effectively to poses extracted from open-world images, enabling animation of social interactions in human images; 3) Ponimator can generate interactive poses from a single-person pose, text, or both, combined with interactive pose animation, enabling diverse applications including reaction animation and text-to-interaction synthesis.

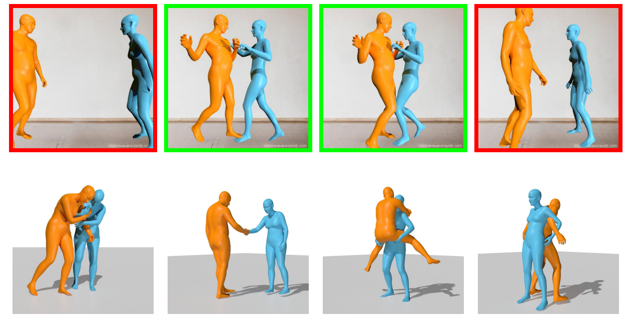

Figure 2. Interactive poses refer to two-person poses in proximity and close contact. The top row displays interactive (green) and non-interactive (red) poses within one sequence. Interactive poses allow observers to intuitively infer the temporal context, while non-interactive poses are more ambiguous and difficult to interpret. The bottom row showcases common daily interactive poses.

2. Related work

Human-human Interactions in Images. Human–human interactions are prevailing in social images. Significant progress has been made in interactive pose estimation [9, 10, 36] and interaction sequence reconstruction [19, 55]. Ugrinovic *et al*. integrate a physical simulator into the human mesh recovery pipeline to capture the physical significance of interactive poses. Huang *et al*. [19] use Vector-Quantised representation learning and specialized losses to learn a discrete interaction prior, but suffer from limited interpretability and generalization. In contrast, our method directly anchors on interactive poses for interaction modeling without relying on additional physical simulators or intricate model designs. Our simple and interpretable prior generalizes well to in-the-wild settings, adhering to the principle that simplicity leads to robustness. The interactive pose prior is also explored in BUDDI [36], which estimates two-person static poses from images but is limited to static pose modeling and overlooks the rich dynamics of interactions. In contrast, our work unlocks interactive motions for both animation and generation in arbitrary open-world images.

Human-human Motion Synthesis. Generating human motion dynamics has been a long-standing task [1, 2, 26, 27, 30, 35]. Utilizing generative models have gained widespread popularity recently [12–14, 24, 25, 28, 40, 41, 53, 54, 58, 59, 65, 66]. With the success of applying generative models in single-person motion synthesis and the release of large-scale two-person interaction datasets, such as InterGen [29] and Inter-X [60], there has been a surge in

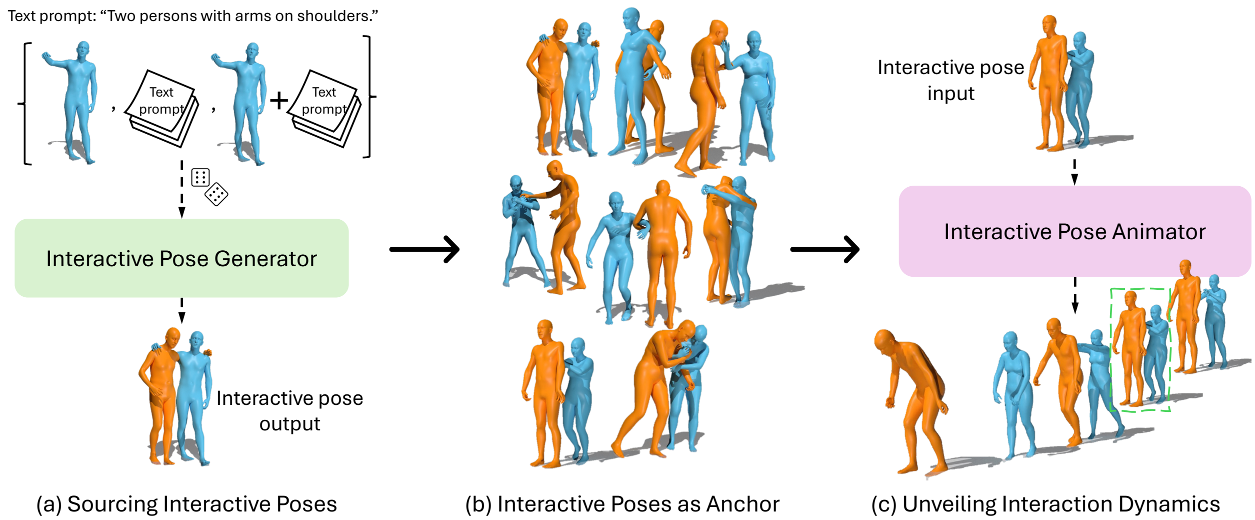

Figure 3. Framework overview. Ponimator consists of a pose generator and animator, bridged by interactive poses. The generator takes a single pose, text, or both as input to produce interactive poses, while the animator unleashes interaction dynamics from static poses.

research [5, 6, 11, 23, 32, 33, 42, 46, 47, 49, 53, 54, 61] focused on multi-person motion generation. However, most existing studies generate two-person motions following input text, but often overlooking close-contact dynamics. For example, Liang *et al.* [29] proposed a diffusion model for two-person motion generation, but it relies on detailed text input and struggles with realistic interaction. In contrast, our framework focuses on short-range interactions by leveraging generalizable interaction priors from static interactive poses, naturally ensuring physical contact between individuals and seamlessly generalizes to open-world scenarios.

Human-human Motion Prediction. A body of work focuses on tracking multi-person motions from videos [21, 22, 48], forecasting future multi-person motions based on past movements [15, 39, 51, 52, 57, 62, 63] and generating reactive motion based on an individual's full motion sequence [5, 8, 11, 33, 45, 49, 61]. However, existing methods rely on long history context or full individual motions while treating interactive poses and human dynamics separately. In contrast, our approach bridges these two modalities by anchoring on interactive poses and leveraging their prior for dynamics forecasting. This integration enables our model to generate both past and future interaction dynamics while supporting flexible inputs with fewer constraints, such as text, single-pose, or both, unlocking diverse applications in animation and generation.

3. Approach

Ponimator leverages interactive pose priors as intermediates for interaction animation, as shown in Fig. 3. We first introduce interactive poses and motion modeling (Sec. 3.1). Then, we present the pose animator (Sec. 3.2), which transforms interactive poses into motion, followed by the pose generator (Sec. 3.3), which generates interactive poses from various inputs. Finally, in Sec. 3.4, we explore Ponimator's applications to real-world images and text.

3.1. Interactive Pose and Motion Modeling.

Interactive pose and motion. Our work defines interactive poses as the poses of two individuals in proximity and close contact. For person a, we use the SMPLX parametric body model [37] to model the pose $\mathbf{x}^a = (\boldsymbol{\phi}^a, \boldsymbol{\theta}^a, \boldsymbol{\gamma}^a)$ and shape $\boldsymbol{\beta}^a \in \mathbb{R}^{10}$. Here, $\boldsymbol{\theta}^a \in \mathbb{R}^{21\times3}$ is the joint rotations, $\boldsymbol{\phi}^a \in \mathbb{R}^{1\times3}$ and $\boldsymbol{\gamma}^a \in \mathbb{R}^{1\times3}$ represents the global orientation and translation. The interactive pose of two individuals a and b is given as $\mathbf{x}_I = (\mathbf{x}_I^a, \mathbf{x}_I^b)$. An interaction motion consists of a short pose sequence $\boldsymbol{\mathcal{X}}$ of length N, centered around an interaction moment, along with shape parameters $\boldsymbol{\theta}$ of both individuals, where $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_i\}_{i=1}^N$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^a, \boldsymbol{\beta}^b)$. $\boldsymbol{\mathcal{X}}$ includes an pair of interactive poses $\mathbf{x}_I$ at interaction moment index I within the sequence, and its nearby past poses $\mathbf{x}_{1:I}$ and future poses $\mathbf{x}_{I+1:N}$. An example of interactive pose and motion is shown in Fig. 2.

Interaction motion modeling. The interactive pose $\mathbf{x}_I$ encodes rich *temporal* and *spatial* priors. As shown Fig. 2, interactive poses convey motion dynamics (top row) and spatial relationships (bottom row) between individuals. The strong prior make it easier for models to learn, whereas non-interactive poses lack clear interaction cues, making learning more challenging. Therefore, we model the interaction motion $(\boldsymbol{\mathcal{X}}, \boldsymbol{\beta})$ by anchoring on its interactive pose $\mathbf{x}_I$.

$$p(\boldsymbol{\mathcal{X}}, \boldsymbol{\beta}) = \underbrace{p(\boldsymbol{\mathcal{X}}; \mathbf{x}_I, \boldsymbol{\beta})}_{\text{temporal prior}} \cdot \underbrace{p(\mathbf{x}_I, \boldsymbol{\beta})}_{\text{spatial prior}} \tag{1}$$

Learning prior from diffusion model. Each prior's distribution in Eq. (1) is captured by a generative diffusion model [17] G, trained on high-quality mocap data. To

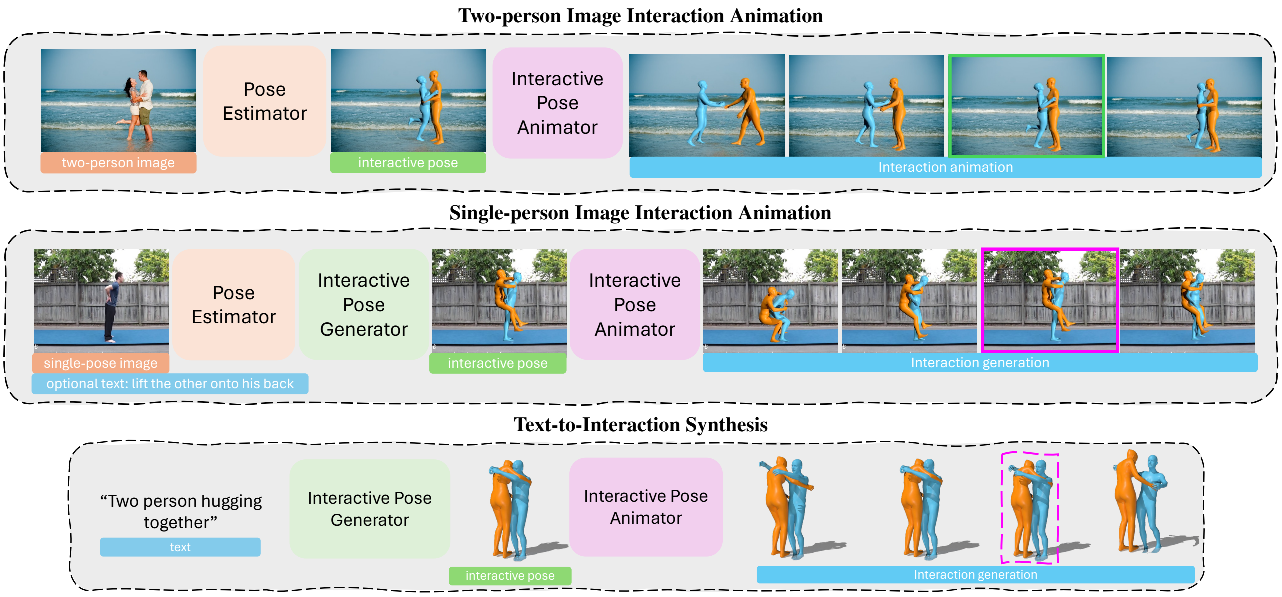

Figure 4. Applications. Our framework enables two-person image animation, single-person interaction generation, and text-to-interaction synthesis. For two-person images, we estimate interactive poses using an off-the-shelf model [36]. For single-person images, we first estimate the pose by [4] and generate its interactive counterpart. For text input, our unified pose generator could synthesize the pose directly. These poses are then fed into our animator to generate human dynamics.

model the underlying distribution of data $\mathbf{z}_0$, the diffusion model introduces noise $\boldsymbol{\epsilon}$ to the clean data $\mathbf{z}_0$ in the forward pass, following $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, where $\alpha_t \in (0,1)$ are constants, t is the diffusion timestep $t \in [0, T_{\text{diffusion}}]$. The model G aims to recover clean input by $\hat{\mathbf{z}}_0 = G(\mathbf{z}_t, t, \mathbf{c})$ from the noisy observations $\mathbf{z}_t$ and condition $\mathbf{c}$, optimizing the objective:

$$\mathcal{L}_D = \mathbf{E}_{\mathbf{z}_0, \mathbf{c}, \boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I}), t}[\|\mathbf{z}_0 - G(\mathbf{z}_t, t, \mathbf{c})\|_2^b] \quad (2)$$

During inference, the model iteratively predicts $G(\mathbf{z}_t, t, \mathbf{c})$ from $t = T_{\text{diffusion}}$ to $t = 0$, gradually denoising the sample until it recovers the original clean data $\hat{\mathbf{z}}_0$.

Close-proximity training data. We collect large-scale training data from public mocap datasets, InterX [60] and DualHuman [7], without requiring contact annotations. Interactive poses are detected by spatial proximity, and if within a threshold, we extract the pose with its past and future frames to form a 3-second interaction clip.

3.2. Unveiling Dynamics from Interactive Poses

The interactive pose animator captures the temporal prior in $p(\boldsymbol{\mathcal{X}}; \mathbf{x}_I, \boldsymbol{\beta})$ given an interactive pose $\mathbf{x}_I$ and two person's shape $\boldsymbol{\beta}$. The objective is to generate the motion sequences $\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\mathbf{x}}_i\}_{i=1}^N$ where $\hat{\mathbf{x}}_I \approx \mathbf{x}_I$, as shown in Fig. 3 (c).

Interactive pose-centered representation. We anchor the entire sequence on the interactive pose $\mathbf{x}_I$ and define the denoising target $\mathbf{z}_0$ as the motion residuals with respect to interactive poses $\mathbf{z}_0 = \{\mathbf{x}_i - \mathbf{x}_I\}_{i=1}^N$ This learning objective enforces model to learn the contextual dynamics strongly shaped by interactive poses. During inference, we recover the predicted pose sequence $\{\hat{\mathbf{x}}_i\}_{i=1}^N$ by $\hat{\mathbf{z}}_0 + \mathbf{x}_I$.

We encode the interactive time index I with a one-hot vector $\mathbf{m}_I \sim \texttt{OneHot}(I) \in \{0,1\}^N$, where $\mathbf{m}_I^i = 1 \texttt{ iff } i = I$. To better preserve the spatial structure of interactive pose at time I in pose sequences, we apply an imputation strategy to the diffusion model, where the noise input $\mathbf{z}_t$ in Eq. (2) is substituted with $\tilde{\mathbf{z}}_t$:

$$\tilde{\mathbf{z}}_t = (1 - \mathbf{m}_I) \odot \mathbf{z}_t + \mathbf{m}_I \odot \mathbf{0}, \quad \mathbf{c} = (\mathbf{m}_I, \mathbf{x}_I, \boldsymbol{\beta}), \quad (3)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{c}$ is the input condition. After imputation, noise is added to interactive poses (i.e., $\tilde{\mathbf{z}}_t + \mathbf{x}_I$) before fed into the network.

Condition encoding. The interaction time condition $\mathbf{m}_I$ is concatenated with the initial model input along the feature dimension. We encode the remaining conditions $(\mathbf{x}_I, \boldsymbol{\beta})$ by leveraging the SMPLX joint forward kinematics (FK) function $\texttt{FK}(\cdot,\cdot)$ to compute joint positions of interactive pose $\mathbf{j}_I = (\texttt{FK}(\mathbf{x}_I^a, \boldsymbol{\beta}_a), \texttt{FK}(\mathbf{x}_I^b, \boldsymbol{\beta}_b))$. Here, $\mathbf{j}_I$ inherently encodes both individuals' poses and shapes. It is further embedded through a single-layer MLP and injected into the model layers via AdaIN [20].

Architecture and training. We adopt the DiT [38] architecture as our diffusion model, built on stacked Transformer blocks [56] that alternate spatial attention for human contact and temporal attention for motion dynamics. To train the model, besides diffusion loss $\mathcal{L}_D$ in Eq. (2), we apply the SMPL loss $\mathcal{L}_{\text{smpl}}$ as the MSE between the denoised pose sequence and the clean input. We also use an interaction loss $\mathcal{L}_{\text{inter}}$ [29] and a velocity loss [54]. $\mathcal{L}_{\text{vel}}$

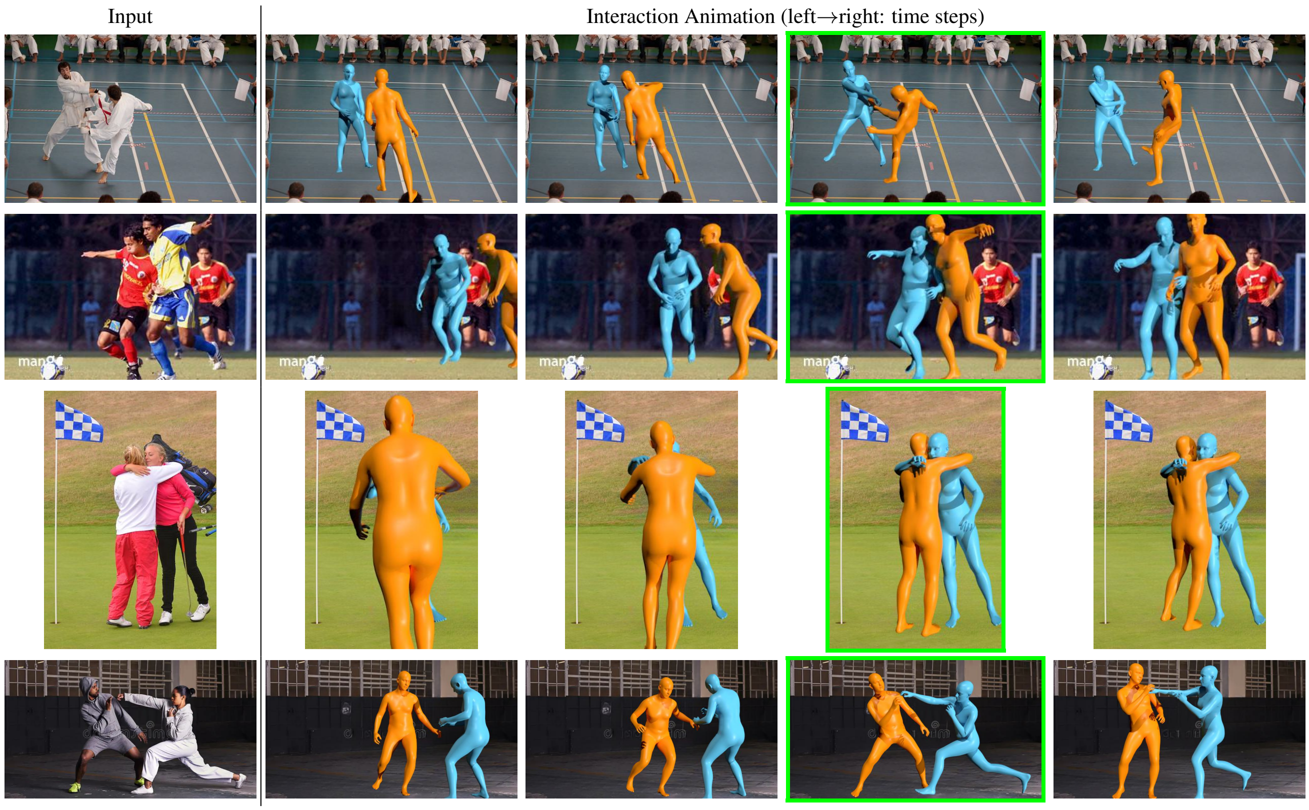

Figure 5. Interactive pose image animation on FlickrCI3D dataset [9]. Left shows the input image, right shows the animated interaction motions. Interactive-pose frame is labeled in green box.

encourages contact between individuals in close proximity, while $\mathcal{L}_{\text{vel}}$ ensures motion coherence. The total loss $\mathcal{L} = \lambda_D \mathcal{L}_D + \lambda_{\text{smpl}} \mathcal{L}_{\text{smpl}} + \lambda_{\text{inter}} \mathcal{L}_{\text{inter}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}}$. To improve robustness and generalization to noisy real-world poses, we apply augmentation by adding random noise to interactive pose $\mathbf{x}_I$. Please refer to Supp. for details.

3.3. Interactive Pose Generator

The interactive pose generator models $p(\mathbf{x}_I, \boldsymbol{\beta})$ in Eq. (1), leveraging the spatial prior to generate $\mathbf{x}_I, \boldsymbol{\beta}$ from various conditions, as shown in Fig. 3(a).

Unified input conditioning. Given various input conditions, including text $\mathbf{c}$, single person pose $(\mathbf{x}_I^a, \boldsymbol{\beta}^a)$, or both, the model generates $\mathbf{z}_0^a = (\mathbf{x}_I^a, \boldsymbol{\beta}^a)$ and $\mathbf{z}_0^b = (\mathbf{x}_I^b, \boldsymbol{\beta}^b)$, which together form the diffusion target $\mathbf{z}_0 = (\mathbf{z}_0^a, \mathbf{z}_0^b)$ in Eq. (2). To integrate these conditions into a unified model, we introduce two masks, $\mathbf{m}_c$ and $\mathbf{m}_a$, to encode the presence of text and pose conditions, respectively. These masks are sampled independently from a Bernoulli distribution with probability $p_{\text{condition}}$ during training. We modify the model input $\mathbf{z}_t$ and text condition $\mathbf{c}$ to $\tilde{\mathbf{c}}$ in Eq. (2) as:

$$\tilde{\mathbf{z}}_t = ((1 - \mathbf{m}_a) \odot \mathbf{z}_t^a + \mathbf{m}_a \odot \mathbf{z}_0^a, \mathbf{z}_t^b), \quad \tilde{\mathbf{c}} = \mathbf{m}_c \odot \mathbf{c}. \tag{4}$$

This design enables the model to accommodate multiple combinations of conditions.

In SMPL, human shapes are coupled with genders $g \in \{\text{male}, \text{female}, \text{neutral}\}$. To enable a more generic shape condition, we instead use the global joint positions of rest pose $\mathbf{j}_{\text{rest}}^{\{a,b\}}$, which inherently capture both shape and gender information, and define the diffusion target as $\mathbf{z}_0 = (\mathbf{x}_I^{\{a,b\}}, \mathbf{j}_{\text{rest}}^{\{a,b\}})$. After generation, we can recover $\boldsymbol{\beta}^{\{a,b\}}$ from $\mathbf{j}_{\text{rest}}^{\{a,b\}}$ using inverse kinematics (IK).

Architecture and training. We use the same architecture as pose animator with modifications below. (1) The text condition $\mathbf{c}$ is encoded via CLIP [44], processed by two trainable Transformer layers, and injected by AdaLN [20]. (2) We retain spatial attention layers and remove temporal attentions. The model is trained with standard diffusion loss $\mathcal{L}_D$ in Eq. (2), SMPL loss $\mathcal{L}_{\text{smpl}}$, and bone length loss $\mathcal{L}_{\text{bone}}$ minimizes the MSE with ground-truth lengths in the SMPLX [37] kinematic tree. Total loss $\mathcal{L} = \lambda_D \mathcal{L}_D + \lambda_{\text{smpl}} \mathcal{L}_{\text{smpl}} + \lambda_{\text{bone}} \mathcal{L}_{\text{bone}}$. Please see Supp. for details.

3.4. Applications

Our framework supports two-person interactive pose image animation, single-person pose interaction generation, and text-to-interaction synthesis, as shown in Fig. 4.

Interactive pose image animation. As shown in 1st row of Fig. 4, given a two-person image, we estimate the interactive pose $\hat{\mathbf{x}}_I$ using an off-the-shelf model [36]. The

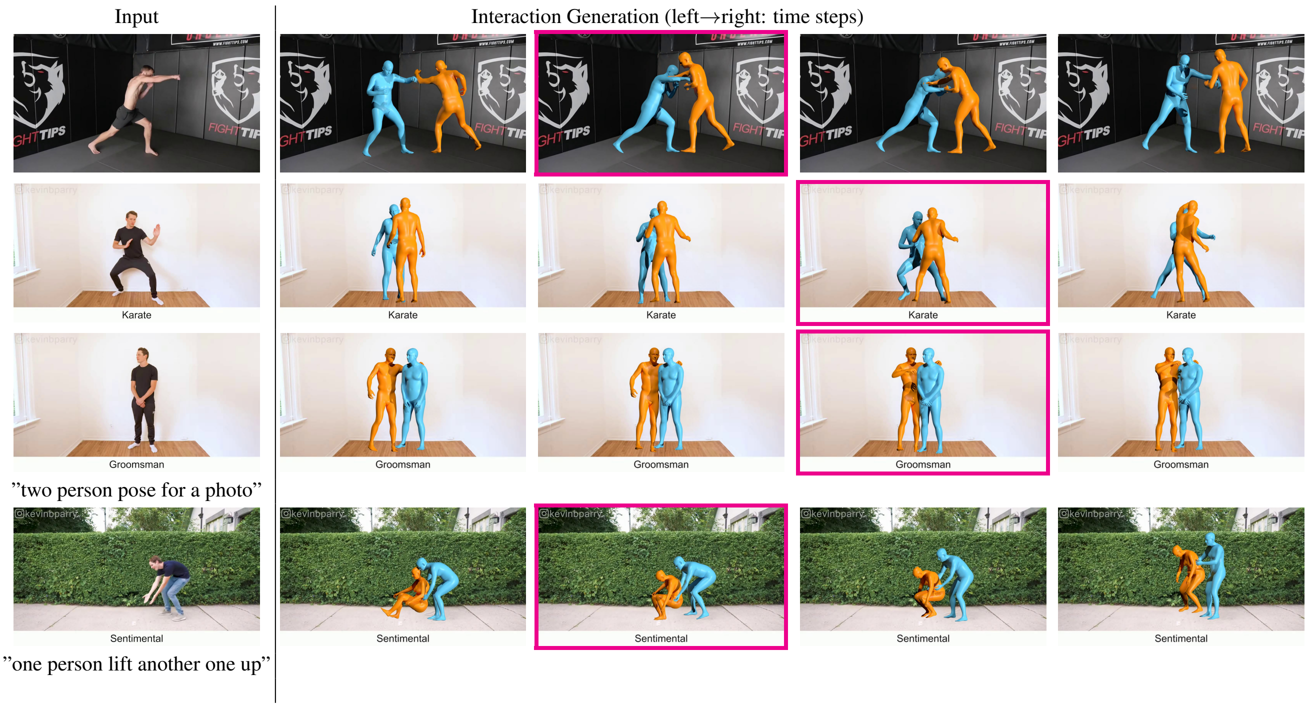

Figure 6. Single-person image interaction generation on Motion-X [31] dataset. Left shows the single person image input, right shows the generated two-person interaction dynamics. The generated interactive pose frame is labeled in magenta box. Top two rows display single-person pose inputs, while the bottom two show the same with accompanying text below the input image.

estimated pose is fed into our interactive pose animator (Sec. 3.2) to generate motions guided by the temporal prior in interactive poses. Our model provides flexible interaction timing control by adjusting I in Eq. (3), where $I = 0$ predicts future motion, $I = N$ reconstructs the past, and generally, $n = \frac{N}{2}$ enables symmetric animation. Open-world animation results are shown in Fig. 5.

Single-person pose interaction generation. As shown in the 2nd row of Fig. 4, given a single-person image, we estimate the pose $\hat{\mathbf{x}}_I^a$ using off-the-shelf model such as [4] and feed it into our interactive pose generator (Sec. 3.3). We set $\mathbf{m}_a = \mathbf{0}, \mathbf{m}_c = 0$ in Eq. (4) as model input, disabling text input and allowing $\hat{\mathbf{x}}_I^a$ to generate its interactive counterpart $\mathbf{x}_I^b$ using the spatial prior in interactive poses. Alternatively, setting $\mathbf{m}_c = 1$ enables additional text conditioning. Once the interactive pose $\hat{\mathbf{x}}_I = (\hat{\mathbf{x}}_I^a, \hat{\mathbf{x}}_I^b)$ is obtained, it is fed into the interactive pose animator (Sec. 3.2) to synthesize motion dynamics. Open-world results are presented in Fig. 6.

Text-to-interaction synthesis. As shown in 3rd row of Fig. 4, given a short phrase, we generate the interactive pose $\hat{\mathbf{x}}_I$ by setting $\mathbf{m}_a = \mathbf{0}, \mathbf{m}_c = 1$ in Eq. (3). The generated $\hat{\mathbf{x}}_I$ is then passed to the pose animator to produce the corresponding motion. Examples for "two-person hugging together" and "push" are presented in Figs. 4 and 8.

4. Experiments

Implementation details. We extract interactive poses by detecting SMPL-X vertices contacts [36] below a threshold in each mocap dataset within a 3s window. The interactive pose animator has 8 layers (latent dim 1024) and is trained using AdamW [34] (LR $1e$-4). All loss weights are 1 except $\lambda_{\text{inter}} = 0.5$.To handle real-world noise, we augment training by adding Gaussian noise (scale 0.02) to interactive poses. At inference, DDIM [50] samples 50 steps, generating $3s$ motions at 10fps in $0.24s$ on an A100. The interactive pose generator follows a similar setup with $p_{\text{text}} = 0.8$, $p_{\text{pose}} = 0.2$, and a frozen CLIP-ViTL/14 [44] text encoder. The pose generation take $0.21s$. Models are trained for 4000 epochs with batch sizes of 256 (pose animator) and 512 (pose generator). Please see Supp. for details.

Datasets. We train and test our model on two large-scale datasets: Inter-X [60] (11k sequences) and Dual-Human [7] (2k sequences). We follow the official split for Inter-X and use a 3:1 training-testing split for Dual-Human, excluding non-interactive motion sequences.

Metrics. We follow the evaluation metrics in [43, 46, 54]: **Frechet Inception Distance (FID)**, the feature distribution against ground truth (GT). We compute it by training a motion autoencoder to encode motion into features for each task; **Precision (Pre.)**, the likelihood that generated motions fall within the real distribution; **Recall (Rec.)**, the likelihood that real motions fall within the generated distribution; **Diversity**, the variance of generated motions. We also evaluate the physics plausibility via **Contact Frame Ratio (CR., %)**—proportion of frames with two-person contact—and averaged **Inter-person Penetration (Pene.**, cm).

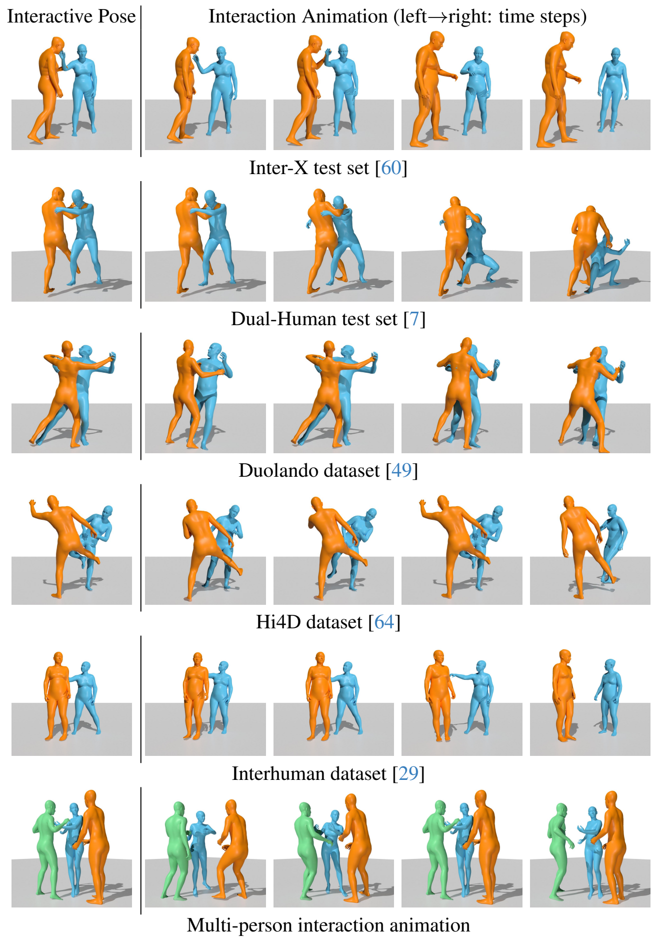

Figure 7. Interactive pose animation on in-domain datasets (Inter-X[60], Dual-Human [7]), out-of-domain dataset (Duolando [49], Hi4D [64], Interhuman [29]), and random composed multi-person pose. Each row: left—interactive pose, right—animation sequence. Our learned interactive pose prior is universal, generalizing across datasets and enabling multi-person interactions (6th row) without modification or retraining.

4.1. Effectiveness of Anchoring on Interactive Poses

Previous works model human-human interaction dynamics either by finetuning on single-person motion priors with interaction data (e.g., ComMDM [46], RIG [53]) or by learning interaction dynamics from scratch (e.g., InterGen [29]). In this work, we model interaction dynamics by anchoring on proximal interactive poses. To evaluate the effectiveness of these approaches, we employ a simple task—unconstrained generation. We further adapt MDM [54] to accommodate two-person motions in our setting. Ponimator seamlessly supports unconstrained generation by setting $\mathbf{m}_a = 0$ and $\mathbf{m}_c = 0$. Experimental results on our dataset collection from Inter-X [60] are shown in Tab. 1. We observe that previous methods [29, 46, 53] struggle to synthesize close-contact interactions, while the adapted MDM* [54] exhibits lower interaction motion quality. In contrast, by simply anchoring on interactive poses, our model achieves superior motion realism (FID of 22.6) and physical contact (contact ratio of 68.1).

Method	FID ↓	Pre. ↑	Rec.↑	Div. →	CR.→	Pene.↓
GT	0.3	1.0	1.0	10.1	70.6	3.8
MDM* [54]	62.6	**0.79**	0.20	9.8	66.4	5.3
ComMDM [46]	88.8	0.37	0.49	10.9	44.3	4.7
RIG [53]	65.2	0.46	0.65	10.6	44.3	**4.3**
InterGen [29]	56.6	0.57	0.46	**10.1**	50.9	**4.3**
Ours	**22.6**	0.58	**0.72**	10.2	**68.1**	5.0

Table 1. Unconstrained interaction synthesis comparison on Inter-X [60] dataset. → means the closer to ground truth the better the result. Method in $*$ is adapted from ours for two-person interaction. Our method largely outperforms others in motion quality and contact ratio, naturally ensuring physical contact and motion realism by anchoring on interactive poses.

	Inter-X				Dual-Human			
Method	FID↓	Div.→	CR. →	Pene.↓	FID↓	Div.→	CR. →	Pene.↓
GT	0.3	10.1	70.6	3.8	2.1	12.0	70.4	3.4
InterGen*	18.9	10.6	44.4	**4.3**	88.8	**11.9**	44.3	**4.1**
w/o anchor	7.1	9.8	67.3	5.1	36.9	11.6	70.7	4.5
- time	6.3	10.3	66.9	5.2	30.3	12.6	67.3	5.1
- joints	5.6	10.0	67.6	5.1	29.9	12.3	70.2	4.4
random-pose	5.8	**10.1**	67.4	5.1	30.1	12.3	69.3	4.5
ours	**5.0**	9.9	**68.5**	5.1	**24.2**	11.8	**70.4**	4.5

Table 2. Interactive pose animation comparison on Inter-X [60] and Dual-Human [7] dataset. InterGen* is adapted to take interactive poses input but lacks explicit interaction modeling, limiting its use of pose priors. Interactive pose anchoring, condition encoding, and interactive frames are crucial for the performance.

4.2. Interactive Pose Animation

To evaluate the interactive pose animator, we compare against baselines and key ablations on Inter-X [60] and Dual-Human [7] datasets in Tab. 2. We ablate key components of pose animator: **w/o anchor** removes interactive pose anchoring, replacing the denoising target $\mathbf{z}_0$ with $\{\mathbf{x}_i\}_{i=1}^N$; **- time** removes the interaction time encoding $\mathbf{m}_I$; **- joints** removes joints condition encoding; **InterGen*** replaces text conditions with interactive pose condition while keeping all other settings unchanged; **random-pose** uses random instead of interactive frames as anchor. All baselines are trained under the same setting. Tab. 2 highlights the importance of interactive pose anchoring and interaction conditioning. InterGen* overlooks input poses, resulting in poorer performance. In contrast, our method explicitly

Method	FID↓	Div.→	MModality↑	CR. →	Pene.↓
GT	0.06	6.78	-	70.6	3.8
InterGen	2.87	6.76	1.42	39.8	**3.9**
w/o anchor	2.74	**6.78**	1.41	39.0	4.0
Ours	**1.82**	**6.78**	**1.46**	**45.9**	4.3

Table 3. Text-to-interaction synthesis results on Inter-X [60] dataset. Our unified pipeline outperforms end-to-end w/o interactive pose as anchor method in short-term interaction synthesis.

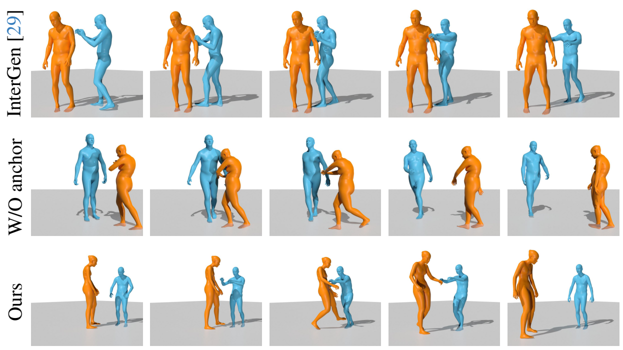

Figure 8. Text-to-interaction comparison for "push". Anchored on interactive poses, our method achieves better contact and more realistic dynamics than InterGen [29] and the end-to-end baseline.

models interaction and contact and achieves better results.

Universal interactive pose prior. We visualize the animated motion in Fig. 7 on in-domain datasets (Inter-X[60], Dual-Human [7]) and out-of-domain datasets (Duolando [49], Hi4D [64], Interhuman [29]). Our approach generalizes to unseen subjects and interactions using the universal interactive pose prior. Our model is surprisingly capable of generating interactions beyond two persons without modification or retraining (see last row in Fig. 7).

Open-world two-person image animation. Our model generalizes to open-world images by extracting interactive poses from FlickrCI3D [9] dataset using [36]. As shown in Fig. 5, it transforms static poses into realistic motion. Please visit our project page for video visualization.

4.3. Interaction Motion Generation

We evaluate interaction motion generation on the Inter-X dataset [60] using text and single-person poses.

Text-to-interaction synthesis We focus on 3s interaction generation, evaluating FID, Diversity, and **MModality**—the ability to generate diverse interactions from the same text [29, 54]. We compare with InterGen [29] and an end-to-end w/o interactive pose baseline, both trained and tested on the same data. As shown in Tab. 3 and Fig. 8, they struggle with contact modeling, while ours excels in short-term interaction generation using interactive pose priors.

Interaction synthesis from single pose We evaluate single

Method	FID↓	Pre.↑	Rec.↑	Div.→	CR.→	Pene.↓
GT	0.3	1.0	1.0	10.1	70.6	3.8
w/o anchor	40.0	0.87	0.43	9.6	67.5	**5.0**
Ours	**27.8**	**0.91**	**0.48**	**9.7**	**73.3**	5.2

Table 4. Single pose-to-interaction synthesis results on Inter-X [60] dataset. Compared to without anchor baseline, our method uses interactive poses for more effective interaction modeling.

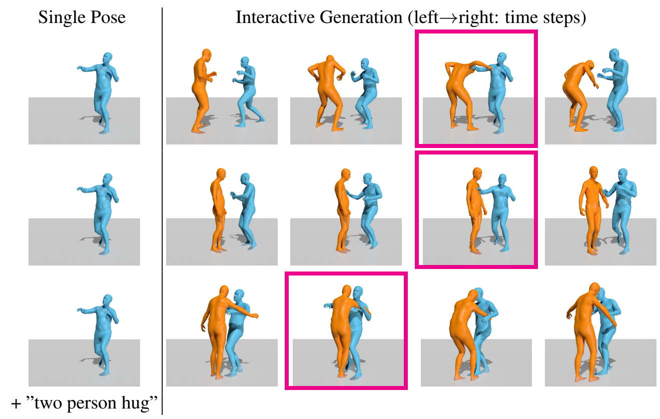

Figure 9. Diverse interactive motion generation. From a single pose, our framework generates varied interactive poses (magenta box) and motions (1st, 2nd rows) and text-driven ones (3rd row).

pose-to-interaction synthesis on Inter-X [60] dataset, comparing our method with an end-to-end without interactive pose baseline, which struggles in the large motion space (Tab. 4). Our method leverages interactive poses to generate diverse motions under varying input conditions in Fig. 9.

Open-world single-person image animation. Our model generalizes to open-world single-person images by estimating poses [4], generating interactive counterparts, and animating motion. Fig. 6 shows results on Motion-X [60] dataset. Please visit our project page for videos.

4.4. Limitations

Our method has few limitations: (1) model short interaction segments; (2) ignore scene context; (3) pose errors may cause contact errors or foot sliding; (4) close interactions may lead to penetration. Please see Supp. for details.

5. Conclusion

We introduce Ponimator, which integrates a pose animator and generator for interactive pose animation and generation using conditional diffusion models. The animator leverages temporal priors for dynamic motion generation, the generator uses spatial priors to create interactive poses from a single pose, text, or both. Ponimator enables open-world image interaction animation, single-pose interaction generation, and text-to-interaction synthesis, exhibiting strong generalization and realism across datasets and applications.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*. IEEE, 2019. 2

[2] Okan Arikan and David A Forsyth. Interactive motion generation from examples. *TOG*, 2002. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[4] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 4, 6, 8

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

[6] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. *arXiv preprint arXiv:2405.15763*, 2024. 3

[7] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *CVPR*, 2024. 2, 4, 6, 7, 8

[8] Yanwen Fang, Jintai Chen, Peng-Tao Jiang, Chao Li, Yifeng Geng, Eddy KF Lam, and Guodong Li. Pgformer: Proxy-bridged game transformer for multi-person highly interactive extreme motion prediction. *arXiv preprint arXiv:2306.03374*, 2023. 3

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 2, 5, 8

[10] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 2021. 2

[11] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: Reactive 3d motion synthesis for two-person interactions. *arXiv preprint arXiv:2311.17057*, 2023. 3

[12] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Multimedia*, 2020. 2

[13] 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.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*. Springer, 2022. 2

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

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 2

[19] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *CVPR*, 2024. 2

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4, 5

[21] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *CVPR*, 2017. 3

[22] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *CVPR*, 2017. 3

[23] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modelling. *arXiv preprint arXiv:2410.10010*, 2024. 3

[24] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2023. 2

[25] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023. 2

[26] Lucas Kovar and Michael Gleicher. Flexible automatic motion blending with registration curves. In *Symposium on Computer Animation*. San Diego, CA, USA, 2003. 2

[27] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *SIGGRAPH*, 2008. 2

[28] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *TOG*, 2022. 2

[29] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, 2024. 2, 3, 4, 7, 8

[30] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Learning*, 2018. 2

[31] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2024. 6

[32] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *ICCV*, pages 20609–20620, 2023. 3

[33] Yunze Liu, Changxi Chen, and Li Yi. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983*, 2023. 3

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[35] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 2

[36] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, 2024. 2, 4, 5, 6, 8

[37] 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*, 2019. 3, 5

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4

[39] Xiaogang Peng, Yaodi Shen, Haoran Wang, Binling Nie, Yigang Wang, and Zizhao Wu. Somoformer: Social-aware motion transformer for multi-person motion prediction. *arXiv preprint arXiv:2208.09224*, 2022. 3

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

[41] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*. Springer, 2022. 2

[42] Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2in: Leveraging individual information to generate human interactions. *arXiv preprint arXiv:2404.09988*, 2024. 3

[43] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *CVPR*, 2023. 6

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 5, 6

[45] 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*, 2023. 3

[46] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024. 3, 6, 7

[47] Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open domain text-driven synthesis of multi-person motions. *arXiv preprint arXiv:2405.18483*, 2024. 3

[48] Bing Shuai, Alessandro Bergamo, Uta Buechler, Andrew Berneshawi, Alyssa Boden, and Joseph Tighe. Large scale real-world multi-person tracking. In *ECCV*. Springer, 2022. 3

[49] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024. 3, 7, 8

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[51] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *TOG*, 2020. 3

[52] Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. Neural animation layering for synthesizing martial arts movements. *TOG*, 2021. 3

[53] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *ICCV*, 2023. 2, 3, 7

[54] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3, 4, 6, 7, 8

[55] Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. Multiphys: multi-person physics-aware 3d motion estimation. In *CVPR*, 2024. 2

[56] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 4

[57] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *NeurIPS*, 2021. 3

[58] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021. 2

[59] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, 2022. 2

[60] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, 2024. 2, 4, 6, 7, 8

[61] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *CVPR*, 2024. 3

[62] 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*, 2023. 3

[63] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *ICLR*, 2023. 3

[64] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, 2023. 7, 8

[65] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2

[66] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2