Motion In-Betweening for Densely Interacting Characters

XIAOTANG ZHANG, Durham University, United Kingdom
ZIYI CHANG, Durham University, United Kingdom
QIANHUI MEN, University of Bristol, United Kingdom
HUBERT P. H. SHUM*, Durham University, United Kingdom

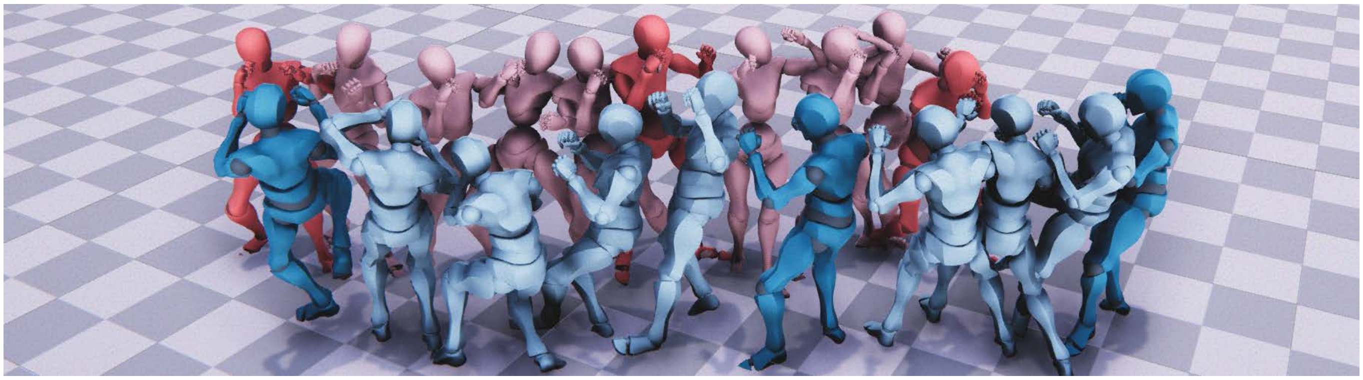


Fig. 1. Our approach is capable of producing interactive in-between motions (light blue and pink characters) for multiple keyposes (blue and red characters) while maintaining high quality over long transitions.

Motion in-betweening is the problem to synthesize movement between keyposes. Traditional research focused primarily on single characters. Extending them to densely interacting characters is highly challenging, as it demands precise spatial-temporal correspondence between the characters to maintain the interaction, while creating natural transitions towards predefined keyposes. In this research, we present a method for long-horizon interaction in-betweening that enables two characters to engage and respond to one another naturally. To effectively represent and synthesize interactions, we propose a novel solution called Cross-Space In-Betweening, which models the interactions of each character across different conditioning representation spaces. We further observe that the significantly increased constraints in interacting characters heavily limit the solution space, leading to degraded motion quality and diminished interaction over time. To enable long-horizon synthesis, we present two solutions to maintain long-term interaction and motion quality, thereby keeping synthesis in the stable region of the solution space. We first sustain interaction quality by identifying periodic interaction patterns through adversarial learning. We further maintain the motion quality by learning to refine the drifted latent space and prevent pose error accumulation. We demonstrate that our approach produces realistic, controllable, and long-horizon in-between motions of two characters with dynamic boxing and dancing actions across multiple keyposes, supported by extensive quantitative evaluations and user studies.

CCS Concepts: • **Computing methodologies → Motion capture**; *Neural networks.*

Additional Key Words and Phrases: Animation, Motion Synthesis, Motion In-betweening, Deep Learning

ACM Reference Format:
Xiaotang Zhang, Ziyi Chang, Qianhui Men, and Hubert P. H. Shum. 2025. Motion In-Betweening for Densely Interacting Characters. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3757377.3763950

1 INTRODUCTION

Motion in-betweening involves synthesizing realistic character movement between predefined keyposes. It enables animators to efficiently create controllable motions by specifying only keyposes. Prior studies [Chai and Qin 2024; Chu and Yang 2024; Harvey and Pal 2018; Harvey et al. 2020; Hong et al. 2024; Oreshkin et al. 2023; Qin et al. 2022; Studer et al. 2024; Tang et al. 2022] have explored various architectures for the motion in-betweening task, under diverse conditioning schemes such as text [Cohan et al. 2024; Pinyoanuntapong et al. 2024], style [Tang et al. 2023], skeletal topology [Gat et al. 2025; Yun et al. 2025] and keyframe timing [Goel et al. 2025; Starke et al. 2023]. Despite extensive research, existing methods primarily focus on motion in-betweening for a single character, and it is non-trivial to extend these methods to multiple densely interacting characters.

Dense interactions such as boxing or dancing are characterized by precise movements as well as precise timing. Extending interaction synthesis to in-betweening requires generation to fulfill three constraints: (1) The two-character motion should be spatio-temporally aligned such that it's semantically interactive. (2) The two-character motion should end at the predefined keypose at the same time. (3) In-betweening should generalize to user-customized keypose

*Corresponding Author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
SA Conference Papers '25, December 15–18, 2025, Hong Kong, Hong Kong
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2137-3/2025/12.
https://doi.org/10.1145/3757377.3763950

which typically lies outside the spatio-temporal distribution learned from dataset. These precise requirements introduce significantly more constraints to the solution space than single-character in-betweening.

Fundamentally, the core challenge of interaction in-betweening is two-fold. First, it requires explicitly modeling interactions to capture precise movement and timing that define the interaction, while dynamically satisfying keyposes for each character. Second, enforcing dense interaction introduces a substantial number of constraints. To fulfill them would easily lead to unnatural motion and cause keyposes unreachable. This degradation compounds over time, ultimately making long-horizon interaction synthesis infeasible.

In this paper, we present a novel solution for long-horizon, densely interacting in-betweening that enables two characters to engage and respond to one another naturally. To represent interaction effectively, we introduce *Cross-Space In-Betweening* to model and synthesize reactive in-between motions for two characters. Our approach first represents motions as spatial offsets relative to keyposes and synthesizes transitions for each character individually. The transitions are then transformed relative to the other character, with interaction conditions integrated via an affine transformation learned by Feature-wise Linear Modulation (FiLM) [Perez et al. 2018]. This two-stage synthesis ensures stable, responsive motion transitions conditioned on the relative positions to both the keyposes and the counterpart character.

To address the challenge of overly restrictive constraints in this task, we present two solutions that help to preserve interaction consistency and motion quality over time, enabling long-horizon interaction in-betweening. We first sustain interaction quality by identifying periodic interaction patterns through adversarial learning, which distinguishes real temporal structures from synthetic, inconsistent interactions. This design is inspired by the observation that dense interactions like boxing and dancing often follow periodic and repetitive spatial distance. Second, we maintain individual motion quality by sampling from a refined latent space during inference—a simple yet effective strategy to mitigate error accumulation and distribution drift common in auto-regressive methods. Together, these two strategies foster a robust latent space informed by interaction periodicity and individual motion correction, supporting our goal of high-quality long-horizon synthesis.

We showcase long-horizon interaction in-betweening across the Boxing [Shum et al. 2010], ReMoCap [Ghosh et al. 2025] and InterHuman [Liang et al. 2024] datasets. Our system enables users to interactively select, translate, and rotate keyposes for two characters, with valid interactions automatically generated in response (see Fig.10). Ablation studies, quantitative evaluations, and user studies demonstrate that our method outperforms prior work on interaction in-betweening.

The main contributions of this work can be summarized as:

- We propose Cross-Space In-betweening that enables stable and responsive interaction modeling for two-character motion in-betweening.
- We maintain long-term interaction quality by identifying periodic interaction patterns through adversarial learning.
- We preserve long-term motion quality by learning to refine the drifted latent space and prevent pose error accumulation.
- We demonstrate that our system is robust to produce responsive in-between interactions for user-defined keyposes.

2 RELATED WORKS

2.1 Multi-Character Interaction Synthesis

Modeling interactions between virtual characters has been widely explored in computer graphics and vision. Early works relied on handcrafted patches or probabilistic models to capture interaction dynamics [Ho and Komura 2009; Kwon et al. 2008; Park et al. 2004; Shen et al. 2019; Shum et al. 2008; Yun et al. 2012], but lacked flexibility and scalability. Deep neural networks have become the dominant paradigm in recent years. A number of works focus on predicting the short-term future of interacting characters [Chopin et al. 2023; Guo et al. 2022; Katircioglu et al. 2021; Tanke et al. 2023]. These methods typically extrapolate trajectories from recent motion history. However, their scope is limited to short temporal horizons, and they generally lack mechanisms for conditional generation. Another line of work generates the motion of one character conditioned on the observed trajectory of another. Recent approaches [Cen et al. 2025; Ghosh et al. 2025; Xu et al. 2024] demonstrate reactive motion synthesis that align with the given input character's ground-truth motion. However, this setting requires complete observation of motion as input, which makes it unsuitable for interaction in-betweening, where only sparse keyposes are available and both characters must be synthesized jointly.

Recently, diffusion-based models have shown promising results in multi-character generation [Liang et al. 2024; Shafir et al. 2023; Tanaka and Fujiwara 2023; Xu et al. 2023, 2024]. These approaches typically employ text prompts or action labels as control signals, which provides flexibility for generating diverse interaction scenarios. However, such high-level conditions do not allow precise control over the spatial-temporal details of the desired interaction. [Zhang et al. 2023] demonstrated interaction synthesis conditioned on character morphology (e.g., body height), and [Cen et al. 2025] generated reactions based on sparse joint positions, but neither method allows explicit control over the exact spatial-temporal interaction between characters. To bridge this gap, we propose a framework that allows two characters to perform natural interactions with guaranteed alignment to user-specified keyposes.

2.2 Motion In-betweening

Motion in-betweening has been extensively explored for single-character animation, ranging from early space-time optimization and probabilistic models [Lehrmann et al. 2014; Ngo and Marks 1993; Rose et al. 1996; Wang et al. 2007; Witkin and Kass 1988] to recent deep learning approaches using recurrent neural networks [Harvey and Pal 2018; Harvey et al. 2020], Transformers [Chai and Qin 2024; Kim et al. 2022; Oreshkin et al. 2023], mixture-of-experts [Starke et al. 2023; Tang et al. 2022, 2023], and diffusion models [Cohan et al. 2024; Studer et al. 2024]. These methods achieve strong results for smooth, controllable single-character transitions under diverse conditioning schemes.

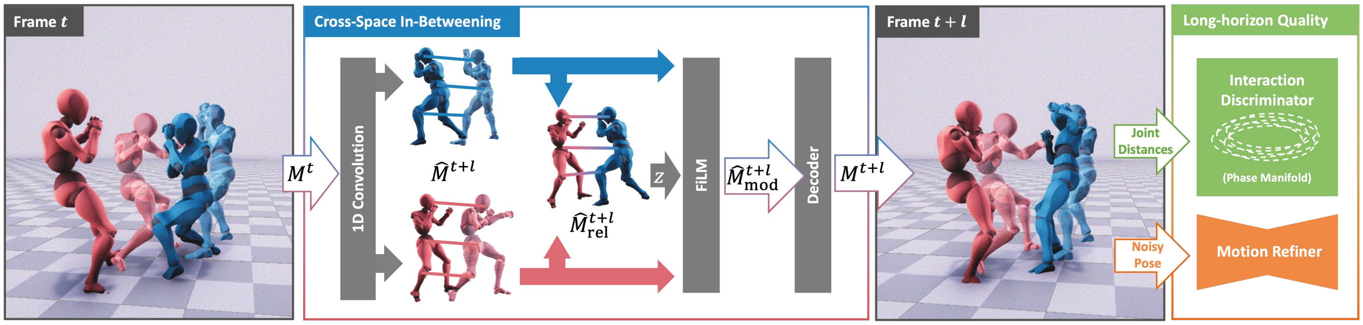


Fig. 2. An overview of our framework. The system first generates an initial prediction for individual character which minimizes the distance to keypose. Then, it extracts relative pose representations as conditions to refine the initial prediction and generates interactive motions. Pairwise joint distances and the outcomes of main network are fed into an interaction discriminator and a motion refiner to model interaction periodicity and to reduce pose error, respectively.

However, directly extending these approaches to two characters is non-trivial. Beyond matching individual keyposes, two-character in-betweening must explicitly model spatial relationships and timing alignment to maintain realistic interactions. Moreover, motion errors compound more rapidly in multi-character settings, where small deviations in one character can destabilize the interaction dynamics. Our method directly addresses these challenges by conditioning motion generation on cross-character spatial representations to preserve interaction fidelity and incorporating two solutions to maintain stable generation.

3 PROBLEM DEFINITION

Our system is designed to auto-regressively predict future motion sequences for both characters based on an input observation sequence. Specifically, the poses of both character $m_1, m_2 \in \mathbb{R}^{J \times C}$ are represented by J joints and C dimensions that capture joint positions and rotations. Given a motion dataset $\mathcal{M}$, we define the motion sequence of two characters starting at frame t and spanning $T = 20$ frames as $M^{t:t+T} = \{(m_1^t, m_2^t), (m_1^{t+1}, m_2^{t+1}), \ldots, (m_1^{t+T}, m_2^{t+T})\}$, which we denote as M^t for simplicity. For each step of prediction, our framework predicts the short sequence of motion in the next $l = 10$ frame based on past observation: $M^{t+l} = f(M^t)$. During inference, the in-between motion sequence is predicted in auto-regressive fashion until reaching the keypose.

4 METHODOLOGY

4.1 Cross-Space In-betweening

It is challenging to predict in-between motions in a two-character scenario, as the transitions pursue natural movements towards keyposes and simultaneously maintain high interaction quality. Thus, our approach strategically decomposes the problem into two key stages: *individual in-betweening* and *interaction modeling*.

In the first stage, individual motions are firstly represented in the coordinate space relative to the keypose (M^t) to obtain the spatial offset following the representation in [Starke et al. 2023]. An initial in-betweening prediction ($\hat{M}^{t+l}$) is then generated by minimizing the distance to the keypose. In the second stage, the motion is then transformed into the coordinate space of its counterpart to obtain their relative spatial information ($\hat{M}^{t+l}_{\text{rel}}$) and to generate the final motion sequence (M^{t+l}). The architecture is shown in Fig. 2.

4.1.1 Individual In-Betweening. Specifically, we convert the input motion into frequencies using Discrete Cosine Transform (DCT) to effectively capture the temporal body dynamics and reduce modeling complexity. An encoder (denoted as *Enc*), composed of 1D convolutional layers followed by a Graph Convolutional Network (GCN) [Kipf and Welling 2016] to capture the spatial dependencies among the joints, is then employed to predict the in-between motions for a single character. This step is formulated as:

$$\hat{M}^{t+l} = Enc(DCT(M^t)). \tag{1}$$

4.1.2 Interaction Modeling. In this stage, we convert the initial coarse prediction ($\hat{M}^{t+l}$) to the root space of the other character to obtain a new representation ($\hat{M}^{t+l}_{\text{rel}}$). These features indicate relative joints positions and rotations between the characters.

$$\hat{M}^{t+l}_{\text{rel}} = \mathcal{T}(\hat{M}^{t+l}), \tag{2}$$

where $\mathcal{T}$ denotes the spatial transformation applied to the motion representation to extract relative pose information.

Learning dependencies between motion features represented in different coordinate spaces is challenging due to they have significantly different distributions. We thus incorporate Feature-wise Linear Modulation (FiLM) layers [Perez et al. 2018] to adaptively bridge this feature gap. FiLM enables the network to modulate motion features through learnable affine transformations, conditioning each character's motion on spatial cues derived from its counterpart. This design is inspired by prior works in style conditioning [Mason et al. 2022; Tang et al. 2023], where FiLM is used for feature modulation for different styles.

Specifically, we first project the relative-space motion features $\hat{M}^{t+l}_{\text{rel}}$ into a normally distributed latent space to regularize the interaction representation and stabilize FiLM conditioning [Kingma 2013]. Then, a FiLM layer is trained to condition the original keypose-space motions on the relative spatial information by modulating the motion features through learned affine parameters. This process enables the model to align feature distributions across coordinate spaces while preserving interaction relevance. The process can be formulated as:

$$z = \mu(\hat{M}_{\text{rel}}^{t+l}) + \epsilon \cdot \sigma(\hat{M}_{\text{rel}}^{t+l}), \quad \gamma, \beta = FiLM(z), \tag{3}$$

where μ, σ and z are mean, log variance and re-parameterized latent variable following normal distribution, respectively. γ and β are learnable scale and shift parameters for motion feature modulation.

The FiLM layer integrates spatial conditions into the interaction modeling process by adaptively modulating the motion features. This modulation yields intermediate features, denoted as $\hat{M}_{\text{mod}}^{t+l}$, which is embedded with relative spatial information between characters. Subsequently, the decoder (Dec)—which shares the encoder's architecture—reconstructs the complete in-between motion sequence based on $\hat{M}_{\text{mod}}^{t+l}$, followed by Inverse Discrete Cosine Transform (IDCT) to recover the final motion frames in the original pose space:

$$\hat{M}_{\text{mod}}^{t+l} = \hat{M}^{t+l} \cdot \gamma + \beta, \quad M^{t+l} = IDCT(Dec(\hat{M}_{\text{mod}}^{t+l})). \tag{4}$$

4.1.3 Training. During training, the outputs are fed back into the network as inputs in auto-regressive manner to enable sequential generation, with scheduled sampling [Bengio et al. 2015] adopted to further improve model robustness. The network is optimized through minimizing mean squared error $\mathcal{L}_{\text{mse}}$ between prediction M^{t+l} and ground truth M_{gt}^{t+l}, as well as a KL divergence loss $\mathcal{L}_{\text{kl}}$:

$$\mathcal{L}_{\text{mse}} = \frac{1}{P} \sum^{P} (M^{t+l} - M_{\text{gt}}^{t+l})^2, \tag{5}$$

$$\mathcal{L}_{\text{kl}} = -0.5 \cdot \left(1 + \sigma - \mu^2 - e^{\sigma}\right), \tag{6}$$

where $P = 3$ denotes the number of auto-regressive prediction steps performed during training. The overall loss function for the Cross-Space In-betweening module is:

$$\mathcal{L}_{\text{inbetween}} = \lambda_{\text{mse}} \mathcal{L}_{\text{mse}} + \lambda_{\text{kl}} \mathcal{L}_{\text{kl}}. \tag{7}$$

Here, λ_{mse}, and λ_{kl} are the corresponding weights that balance different losses.

4.2 Long-horizon Quality

We aim to generate interactive in-between motions that can extend over long horizon. However, the interaction in-betweening task introduces a substantial number of constraints. These constraints can distort the structure of the learned latent space where valid trajectories become sparse and nonlinear. Inference-time distribution quickly drifts from the training one and pose error (e.g., deformed bones in Fig. 15) accumulates after a few prediction steps. Such disruptions in one character's motion propagate abnormal features into the interaction modeling process, causing the network to generate diminished or overly smoothed interactions. (e.g., drifting to keypose without interactive motion, see Fig. 14). To address this challenge, we design two modules modeling interaction periodicity and enhancing motion quality, respectively, which helps to foster a robust latent space at both the interaction and single-character levels to enhance robustness of long-horizon synthesis.

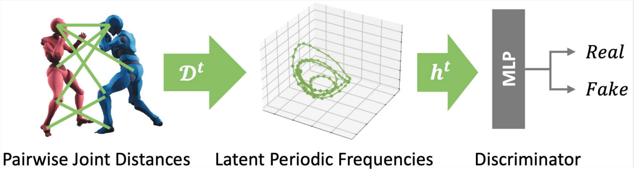


Fig. 3. Details of the interaction periodicity modeling. We first extract the pairwise joint distances $\mathcal{D}^t$ from generated motions (green lines between pairs of joints). We then use Periodic Autoencoder to encode the dynamics as periodic latent frequencies h^t, illustrated as three principal components in the middle via principal component analysis. Our discriminator then learns to identify the interactions from periodic patterns between characters.

4.2.1 Interaction Periodicity. Motivated by the observation that interactions like boxing and dancing exhibit inherent periodic patterns, we leveraged this property to enhance interaction quality and synchronization when generating in-betweening movements.

Previous methods have modeled the periodic patterns of single character with the phase feature [Starke et al. 2023, 2022; Tang et al. 2023], which enhance spatial-temporal alignment in the latent space in predicting subsequent poses. Similarly, we adopt Periodic Autoencoder (PAE) [Starke et al. 2022] to learn the phase feature and generalize it to two characters to capture recurrent patterns in interactive actions by encoding their relative dynamics as periodic frequencies. A discriminator is then deployed to evaluate the interaction quality of predicted motion sequences based on the learned periodicity. The procedure is shown in Fig. 3.

In two-character interactions, periodicity often emerges in the relational patterns between characters, rather than in the motions of each character independently. For example, repeated patterns occur as the punch approaches with decreasing distance between the characters, and then retracts with increasing distance. We thus model the interaction periodicity using pairwise joint distance (PJD) [Tang et al. 2008], which represents the geometry relationship between two characters. Here, we encode PJD dynamics as latent frequencies to represent the periodic patterns. In particular, PJD is formulated as the per-frame offset of Euclidean distance of pairwise joints:

$$d^t = \{\|x_i^t - y_j^t\|_2^2 - \|x_i^{t-1} - y_j^{t-1}\|_2^2 \mid i, j \in J\}, \tag{8}$$

$$\mathcal{D}^t = (d^t, d^{t+1}, \ldots, d^{t+N}) \in \mathbb{R}^{J \times N}, \tag{9}$$

where x_i^t,$y_i^t \in \mathbb{R}^3$ denote the joint positions of two characters in the world space at time step t, d^t is the PJD at time t for all joints, and $\mathcal{D}^t$ is PJD dynamics of length N starting at time t.

The latent frequencies representation $h^t \in \mathbb{R}^{N \times C_\phi}$ is parameterized by sinusoidal functions:

$$h^t = PAE(\mathcal{D}^t) = a^t sin(2\pi(f^t + \phi^t)) + b^t, \tag{10}$$

where ϕ^t is the phase vector predicted by a fully connected layer, and f^t, a^t, b^t are the frequency, amplitude, and bias vector through Fast Fourier Transform (FFT) that models the cyclical dynamics of PJD between two characters. Phase channel C_ϕ is set to be 15.

To further increase the interaction realism, we train the main network (i.e., Cross-Space In-betweening) adversarially that learns to differentiate between realistic and unrealistic interactive patterns

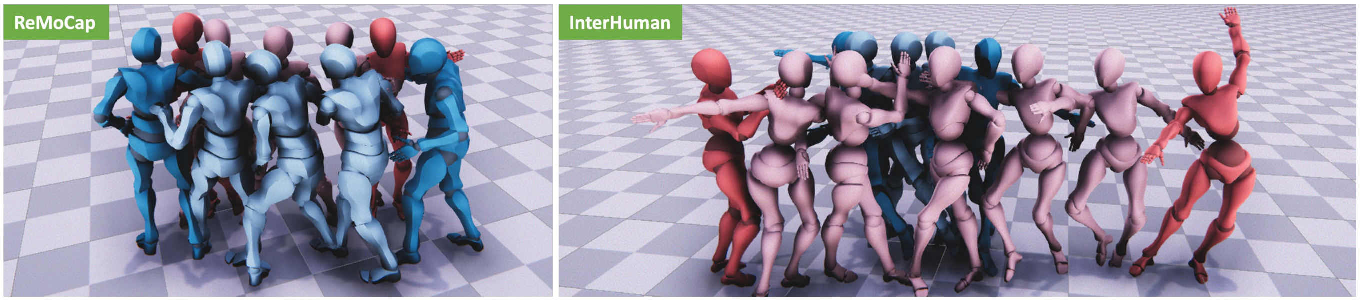


Fig. 4. Qualitative results on ReMoCap and InterHuman dataset. Our method produces smooth and seamless turning motions (light blue and pink) in between keyposes (blue and red).

through the latent frequencies h^t. While the main network autoregressively generates short-term motions, the discriminator offers global supervision by evaluating the periodic quality of the longer predictions (i.e., $N = 30$ in Equation 9). By identifying sampled sequences that could lead to erroneous interactions, it facilitates the generation of more realistic interactive motion patterns.

The adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \mathbb{E}_{M_{\text{gt}}^{t+l:t+l+N} \sim \mathcal{M}}[\log D(M_{\text{gt}}^{t+l:t+l+N})]+ \\ \mathbb{E}_{M^{t:t+N} \sim p_G}[\log(1 - D(G(M^{t:t+N})))]\end{aligned} \tag{11}$$

where D, G are the discriminator and generator (i.e., Cross-Space In-betweening), $\mathcal{M}$ is the ground truth dataset and p_G denotes the distribution of motion sequences generated by the generator G.

4.2.2 Motion Quality. Given that individual pose error will disrupt interaction consistency, we introduce a Motion Refiner to mitigate error accumulation and address motion degradation at the single-character level. It learns to adjust the latent space derived from the drifted distribution generated by Cross-Space In-betweening and correct the deviations during inference so as to avoid sampling erroneous motions.

Specifically, the input to this module is the motion clip M^{t+l} predicted by the main network, which may contain minor pose error (e.g. invalid joint rotations in Fig. 15). Similar to Cross-Space In-betweening, the module consists of an auto-encoder and a GCN for spatial-temporal feature extraction and reconstruction. It preserves the high-level motion semantics of the input while refining joint-level features. The refined output is denoted as M_{refine}^{t+l}. Empirically, we refine motion in non-overlapping segments of 10 frames that is sufficient for real-time inference while preserving motion variability.

During inference, the Motion Refiner is integrated into the pipeline to correct motion predictions step-by-step. This reduces the risk of accumulating pose errors or drifting into unrealistic latent spaces, and helps reshape the latent space to align with valid motion features. In doing so, the refiner preserves long-term motion quality by avoiding dependence on the degraded distribution produced by the main network. Note that the whole generation system is trained end-to-end, with the Motion Refiner updated independently:

$$M_{\text{refine}}^{t+l} = Refiner(M^{t+l}) \tag{12}$$

$$\mathcal{L}_{\text{refine}} = \frac{1}{P}\sum^{P}(M_{\text{refine}}^{t+l} - M_{\text{gt}}^{t+l})^2. \tag{13}$$

5 EXPERIMENT

5.1 Implementation Details

Simulation. Our rendering system is implemented on top of an open-source motion animation framework [Starke et al. 2020] developed in Unity3D. We do not adopt physics-based simulation as it introduces significant training overhead and requires extensive fine-tuning of torque and action smoothness [Liao et al. 2025], which are not essential for demonstrating interactive motions in our setting.

Pose Representation. To achieve efficient rendering, our system avoids using forward kinematics for joint position computation. Instead, each joint is represented independently in Cartesian 3D space, following [Starke et al. 2020]. Each character consists of $J = 52$ joints and $C = 9$ dimensions (3 for position and 6 for rotation) in total. Joint rotation is defined as a pair of forward and upward Cartesian vectors to avoid ambiguous rotations [Zhang et al. 2018].

Datasets. We trained our model using a Boxing dataset as well as two public dancing motion datasets: ReMoCap [Ghosh et al. 2025] and InterHuman [Liang et al. 2024]. Models are trained separately for each dataset. The Boxing dataset is simulated and collected from a previous motion animation system [Shum et al. 2010], resulting in a total of 23,889 frames of intense boxing actions (e.g., punching, kicking and dodging). For the ReMoCap and InterHuman datasets, we specifically select Lindy Hop and Latin dance sequences containing at least 500 frames and discard segments that lack clear interactions. This result in 39,557 and 18,224 total frames, respectively. All datasets are re-targeted to Mixamo character [Mixamo 2025] using Autodesk Motion Builder [Autodesk 2025] and augmented by mirroring along the X-axis. The dataset is divided into 90% for training, 5% for validation, and 5% for testing.

Keyframe Sampling. Each training sample consists of 50 frames: the first 20 are used as reference input, and the remaining 30 for auto-regressive prediction. During preprocessing, a keyframe is randomly selected between frames 50 and 70 (relative to the first frame). All 50 frames in the sample are then transformed into the coordinate space of this keyframe. This approach ensures that the sampled keyframes cover both nearby and distant positions.

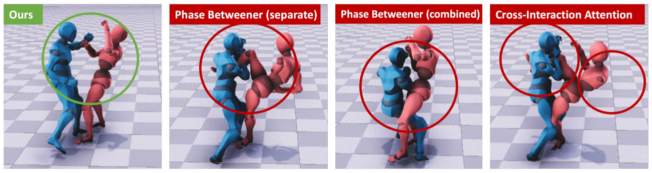

Ours
Phase Betweener (separate)
Phase Betweener (combined)
Cross-Interaction Attention

Fig. 5. Qualitative results compared with baseline methods. Cross-Interaction Attention exhibits severe pose error accumulation issue.

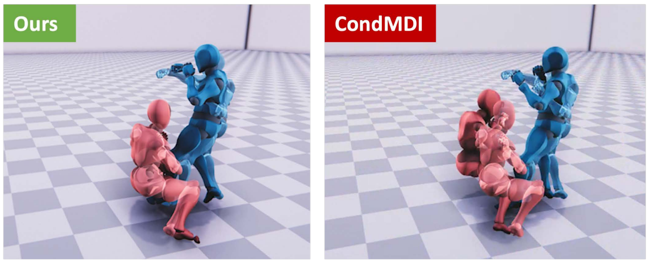


Fig. 6. Keypose alignment performance compared with CondMDI.

Training. We conduct our experiments with an NVIDIA RTX 3080 graphics card, an AMD Ryzen 9 5900X CPU and 32G memory. We train all modules using Adam Optimiser [Kingma and Ba 2014] with a learning rate $\alpha = 0.0001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Network is trained for 100 epochs and cost 40 to 50 hours for each dataset. The motion in-betweening runtime system is adapted from the open-source Unity3D motion animation system [Starke et al. 2020].

5.2 Comparison Methods

We aim to evaluate performance of the two objectives in our task, i.e., interaction synthesis and motion in-betweening, which were addressed separately in previous works. As there is no prior open-source work specifically targeting interaction in-betweening for direct comparison, we evaluate our system against the following representative baselines:

- Phase Betweener [Starke et al. 2023], a single-character in-betweening method with periodicity modeling. To evaluate its generalizability in a two-character setting, we input each character’s motion data either separately or concatenated, referred to as ‘separate’ and ‘combined’ in the following section.
- Cross-Interaction Attention [Guo et al. 2022], a cross attention-based interaction synthesis method with explicit interaction modeling.
- CondMDI [Cohan et al. 2024], a diffusion-based single-character in-betweening framework.

Modification details for these baselines can be found in supplementary materials.

5.3 Qualitative Results

Interaction In-betweening. Our network is able to generate in-between motions in real time until the averaged distance between each character and its corresponding keypose falls below a predefined threshold. The inference time is 125 ms per 10 frames. Qualitative results for Boxing, ReMoCap and InterHuman dataset can be found in Fig. 1, 11, 12 and the supplementary video. Comparisons with baselines are shown in Fig. 5. For Phase Betweener, neither of training approaches yields realistic interactions, as the network fails to implicitly capture the complex relationships between characters. Cross-Interaction Attention also fails to generate interactive in-between motions and suffers from severe error accumulation, as it relies heavily on historical motions for synthesis and struggles with long-term prediction. Moreover, CondMDI struggles with keypose alignment (see Fig. 6). Since it generates entire motion sequences in an offline manner, we loosen the keypose activation threshold in our real-time system to maintain continuity. This adjustment ensures smooth rendering but reduces alignment accuracy.

Controllability. The keyposes are controllable by users. We demonstrate its generalizability on different keyposes when provided with the same initial motion (See Fig. 10 and ’Controllability’ in supplementary video). Through customizing the root positions and rotations of the keyposes or sampling poses from database, our system is robust to produce responsive in-between interactive motion sequences in real time.

Diversity. There are multiple possible transitions that can occur between keyframes. To demonstrate the prediction diversity, we generate three in-between motions for the same keypose condition, as shown in Fig. 7. Our network is capable of generating varied in-between motions while maintaining realistic interactions.

5.4 Quantitative Results

Reconstruction Quality. To evaluate the reconstruction precision, we follow [Harvey et al. 2020] to report the average L2 norm of positions (L2P) and the average L2 norm of quaternions (L2Q) between the ground truth and the generated in-between motions in world coordinate, covering different lengths of in-betweening in short term (refer to Table 1). Our method demonstrates comparable performance across all three datasets and achieves superior reconstruction quality compared to baselines. Further results of long-term reconstruction error are provided in supplementary document.

Interaction Quality. Inspired by [Wang et al. 2022], we evaluate the quality of interactions based on the classification results of the discriminator. Note that the discriminator used for evaluation is trained independently on each of the three datasets using fake samples generated by a different network than the one being tested, so that to ensure an unbiased and effective assessment. Our method outperforms the comparison baselines in terms of interaction quality. It also proves the effectiveness of using the PJD dynamics to represent and assess the interaction quality for periodic motions.

Long-horizon Quality. To assess prediction quality in long horizon, in Table 1, we also report the latent distribution differences with ground truth using Frechet Inception Distance (FID) and Normalized Power Spectrum Similarity (NPSS) [Harvey et al. 2020]. Our method shows robust performance in long transitions, but without the Motion Refiner it can produce highly unrealistic motions (as

Table 1. Quantitative results compared with previous methods and ablated versions. All comparison methods and ablated networks are trained on Boxing dataset only.

	L2P↓		L2Q↓		Foot↓	Interaction↑			Diversity↑		FID↓			100×NPSS↓		
Frames	30	50	30	50	50	40	60	80	30	50	100	150	200	100	150	200
Boxing Dataset																
Ours	0.192	0.208	0.254	0.279	0.342	**0.914**	**0.911**	0.905	1.104	1.802	**0.282**	**0.287**	**0.289**	**1.398**	**1.430**	**1.531**
w/o interaction modeling	0.243	0.253	0.291	0.307	0.336	0.864	0.862	0.862	**1.112**	**1.831**	0.428	0.436	0.439	1.837	2.020	2.348
w/o Inter. GAN	**0.191**	**0.202**	**0.248**	0.266	0.340	0.904	0.902	0.899	1.090	1.787	0.504	0.529	0.556	2.016	2.306	2.551
w/o Motion Refiner	0.235	0.251	0.288	0.301	0.345	0.908	0.901	0.894	1.102	1.788	0.696	0.722	0.730	2.197	2.528	2.814
Phase Betweener (separate)	0.211	0.214	0.257	0.263	0.345	0.866	0.863	0.859	1.089	1.670	0.519	0.534	0.565	2.119	2.304	2.663
Phase Betweener (combined)	0.202	0.210	0.250	0.259	0.350	0.880	0.877	0.872	0.278	0.281	0.547	0.559	0.570	2.849	3.105	3.397
Cross-Interaction	0.217	0.218	0.267	0.272	0.474	0.882	0.875	0.871	0.314	0.317	0.482	0.493	0.502	2.147	2.335	2.590
CondMDI	0.212	0.217	0.263	0.266	0.452	0.905	0.902	0.900	0.293	0.306	0.478	0.484	0.496	1.833	1.940	2.056
Dancing Datasets																
Ours (ReMoCap Dataset)	0.205	0.220	0.264	0.278	0.363	0.908	0.907	**0.907**	0.974	1.077	0.306	0.310	0.312	1.448	1.562	1.793
Ours (InterHuman Dataset)	0.201	0.212	0.249	**0.259**	**0.335**	0.901	0.899	0.898	0.965	1.012	0.289	0.295	0.307	1.407	1.494	1.586

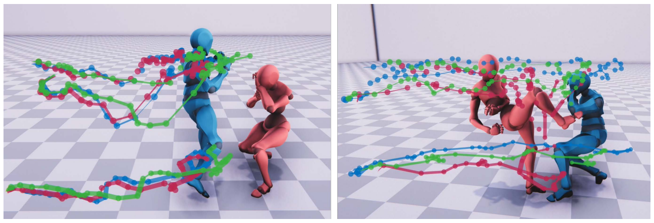


Fig. 7. Qualitative results on predictions diversity. We illustrate historical trajectories in distinct colors for each of the three predictions.

reflected in the FID metric). For NPSS, Phase Betweener (combined) performs the worst in long-horizon synthesis, likely due to the substantially increased complexity of learning a two-character state space with limited capacity of fully-connected layers.

Diversity. We also report the diversity of predictions by measuring the joint positional difference between different samples given the same input and keypose condition. Similar to [Tang et al. 2023], we generate 10 samples for each keypose condition. Our system achieves comparable performance to Phase Betweener. In contrast, Cross-Interaction Attention tends to be deterministic, as its attention mechanism generates similar attention scores for the same input.

Foot Sliding. Following [Zhang et al. 2018], we measure the foot sliding artifact (refer to *Foot* in Table 1) by calculating the averaged foot joint velocity v_f in the first 50 frames when foot height h_f is within threshold $H = 2.5cm$: $v_f \cdot clamp(2 - 2^{h_f/H}, 0, 1)$. It is worth noting that the training datasets inherently contain some foot sliding artifacts. Following [Starke et al. 2020], we apply inverse kinematics on the foot joints to mitigate this issue. Quantitative results compared with ground truth are provided in supplementary document.

User Study. We conduct a user study with 50 participants who have no familiarity with motion in-betweening to assess the visual quality of the generated demos. Each participant is asked to rate the motion quality (on a scale of 1 to 10) for 4 demo cases (2 for boxing dataset and 2 for dancing datasets) generated by different methods, including ground truth. As shown in Fig. 8, our method consistently outperforms the baselines and achieves visual quality comparable to the ground-truth motions. CondMDI achieved higher scores than the other baselines, which may be attributed to its ability to generate fewer unrealistic interactions and produce more continuous transitions across multiple keyposes. Detailed statistics are included in the supplementary document.

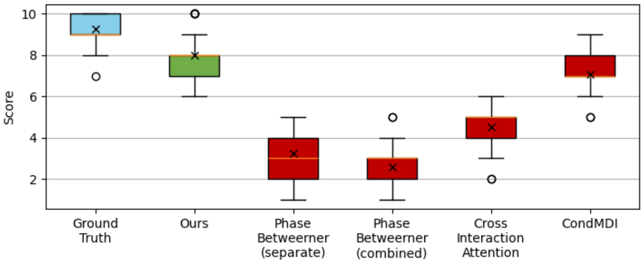


Fig. 8. User study results presented as a box plot of ratings across different methods. Our method achieves scores comparable to the ground truth and surpasses all baseline approaches.

5.5 Ablation Studies

We conduct ablation studies on Boxing dataset to evaluate the effectiveness of different components. To ensure fair comparisons and consistent model complexity, the ablated versions are implemented to have a similar number of parameters as the full model. Qualitative results can be found in Fig. 9,14,13 and the supplementary video.

Cross-Space In-betweening. To evaluate the necessity of FiLM-based interaction conditioning, we conduct an ablation (refer to *w/o interaction modeling*) where each character's motion is predicted independently in its own keypose space without transforming to the other's coordinate system or applying FiLM modulation. Quantitative results indicate a 20% decline in reconstruction performance when dense spatial-temporal relationships between characters are not effectively captured.

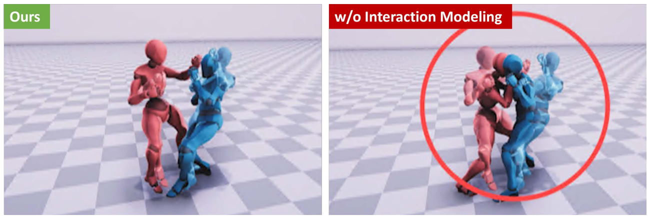


Fig. 9. Qualitative comparison on interaction modeling. On the right, both characters exhibit significant penetration and unrealistic interactions.

Interaction Periodicity. We also test the interaction quality by ablating the discriminator (refer to *w/o Inter. GAN* in the table). We observe a minor improvement in short-term reconstruction precision, likely because removing the discriminator allows the network to focus directly on optimizing the reconstruction loss without the additional constraint of satisfying the adversarial loss. However, we also observe a slight decrease in interaction quality and a significant decline in FID in long transitions. This reflects how modeling the interaction periodicity help learn a robust latent space and maintain long-horizon quality for in-betweening.

Motion Quality. With the Motion Refiner, our network effectively maintain motion quality by reducing the pose errors in long-term synthesis, as shown in Fig. 13 and supplementary video. This is also evident from the lower discrepancy between latent distributions of ground truth and predictions in FID scores.

6 CONCLUSION AND DISCUSSION

In this work, we introduce Cross-Space In-betweening, a novel auto-regressive framework for synthesizing interactive in-between motions. To address the challenges posed by strict keypose and interaction constraints, we maintain interaction and motion quality over long horizons, through modeling of interaction periodicity and refining individual pose errors. Our system enables controllable, long-horizon interaction in-betweening with dense character interactions, and significantly outperforms existing methods across most quantitative metrics.

Limitation. The first limitation of our work is that the motion may not match well with the user-customized keyposes because the imposed spatial constraints can be temporally misaligned (e.g., forcing two characters to punch simultaneously), and the model is unable to infer interactions it has not seen before. Due to the nature of online synthesis, without a global motion planning mechanism, our system does not allow further offline refinement or adjustment of the entire generated in-between sequence for better keypose alignment.

Second, despite showcasing the potential of encoding the periodicity of two-character motions for improving the interaction quality, this strategy is primarily suited for interactions with clear repetitive patterns and does not readily extend to aperiodic interactions (see supplementary document).

Third, as the first work targeting interaction in-betweening, our system currently does not support in-between timing condition as it will further increase modeling complexity.

ACKNOWLEDGMENTS

This research is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1).

REFERENCES

Autodesk. 2025. MotionBuilder. https://www.autodesk.com/products/motionbuilder. Accessed: 2025-01-11.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems* 28 (2015).

Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. 2025. Ready-to-React: Online Reaction Policy for Two-Character Interaction Generation. In *ICLR*.

Zhi Chai and Hong Qin. 2024. Dynamic Motion Transition: A Hybrid Data-driven and Model-driven Method for Human Pose Transitions. *IEEE Transactions on Visualization and Computer Graphics* (2024).

Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. 2023. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia* (2023).

Yuchen Chu and Zeshi Yang. 2024. Real-time Diverse Motion In-betweening with Space-time Control. In *The 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3677388.3696327

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. 2024. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–9.

Inbar Gat, Sigal Raab, Guy Tevet, Yuval Reshef, Amit H. Bermano, and Daniel Cohen-Or. 2025. AnyTop: Character Animation Diffusion with Any Topology. *SIGGRAPH 2025 Conference* (Feb. 2025). https://doi.org/10.48550/arXiv.2502.17327 arXiv:2502.17327 [cs].

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2025. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*. Springer, 418–437.

Purvi Goel, Haotian Zhang, C. Karen Liu, and Kayvon Fatahalian. 2025. Generative Motion Infilling From Imprecisely Timed Keyframes. *Eurographics 2025* (March 2025). https://doi.org/10.48550/arXiv.2503.01016 arXiv:2503.01016 [cs].

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2022. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13053–13064.

Félix G Harvey and Christopher Pal. 2018. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*. 1–4.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 60–1.

Edmond SL Ho and Taku Komura. 2009. Character motion synthesis by topology coordinates. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 299–308.

Seokhyeon Hong, Haemin Kim, Kyungmin Cho, and Junyong Noh. 2024. Long-term Motion In-betweening via Keyframe Prediction. *SCA 2024* (2024).

Isinsu Katircioglu, Costa Georgantas, Mathieu Salzmann, and Pascal Fua. 2021. Dyadic human motion prediction. *arXiv preprint arXiv:2112.00396* (2021).

Jihoon Kim, Taehyun Byun, Seungyoun Shin, Jungdam Won, and Sungjoon Choi. 2022. Conditional motion in-betweening. *Pattern Recognition* 132 (2022), 108894.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

Taesoo Kwon, Young-Sang Cho, Sang I Park, and Sung Yong Shin. 2008. Two-character motion analysis and synthesis. *IEEE transactions on visualization and computer graphics* 14, 3 (2008), 707–720.

Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. 2014. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1314–1321.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision* (2024), 1–21.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. 2025. BeyondMimic: From Motion Tracking to Versatile Humanoid Control via Guided Diffusion. *arXiv e-prints* (2025), arXiv–2508.

Ian Mason, Sebastian Starke, and Taku Komura. 2022. Real-Time Style Modelling of Human Locomotion via Feature-Wise Transformations and Local Motion Phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 1 (May 2022), 1–18. https://doi.org/10.1145/3522618

Mixamo. 2025. Mixamo. https://www.mixamo.com/. Accessed: 2025-01-01.

J Thomas Ngo and Joe Marks. 1993. Spacetime constraints revisited. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*. 343–350.

Boris N Oreshkin, Antonios Valkanas, Félix G Harvey, Louis-Simon Ménard, Florent Bocquelet, and Mark J Coates. 2023. Motion In-Betweening via Deep Δ-Interpolator. *IEEE Transactions on Visualization and Computer Graphics* (2023).

Sang Il Park, Taesoo Kwon, Hyun Joon Shin, and Sung Yong Shin. 2004. Analysis and synthesis of interactive two-character motions. *Technical Note, KAIST, CS/TR-2004* 194 (2004).

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. MMM: Generative Masked Motion Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 1546–1555. https://doi.org/10.1109/CVPR52733.2024.00153

Jia Qin, Youyi Zheng, and Kun Zhou. 2022. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

Charles Rose, Brian Guenter, Bobby Bodenheimer, and Michael F Cohen. 1996. Efficient generation of motion transitions using spacetime constraints. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 147–154.

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

Yijun Shen, Longzhi Yang, Edmond SL Ho, and Hubert PH Shum. 2019. Interaction-based human activity comparison. *IEEE Transactions on Visualization and Computer Graphics* 26, 8 (2019), 2620–2633.

Hubert PH Shum, Taku Komura, Masashi Shiraishi, and Shuntaro Yamazaki. 2008. Interaction patches for multi-character animation. *ACM transactions on graphics (TOG)* 27, 5 (2008), 1–8.

Hubert PH Shum, Taku Komura, and Shuntaro Yamazaki. 2010. Simulating multiple character interactions with collaborative and adversarial goals. *IEEE Transactions on Visualization and Computer Graphics* 18, 5 (2010), 741–752.

Paul Starke, Sebastian Starke, Taku Komura, and Frank Steinicke. 2023. Motion in-betweening with phase manifolds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6, 3 (2023), 1–17.

Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 54–1.

Justin Studer, Dhruv Agrawal, Dominik Borer, Seyedmorteza Sadat, Robert W Sumner, Martin Guay, and Jakob Buhmann. 2024. Factorized Motion Diffusion for Precise and Character-Agnostic Motion Inbetweening. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. 1–10.

Mikihiro Tanaka and Kent Fujiwara. 2023. Role-aware Interaction Generation from Textual Description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15999–16009.

Jeff KT Tang, Howard Leung, Taku Komura, and Hubert PH Shum. 2008. Emulating human perception of motion similarity. *Computer Animation and Virtual Worlds* 19, 3-4 (2008), 211–221.

Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. 2022. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–10.

Xiangjun Tang, Linjun Wu, He Wang, Bo Hu, Xu Gong, Yuchen Liao, Songnan Li, Qilong Kou, and Xiaogang Jin. 2023. RSMT: Real-time Stylized Motion Transition for Characters. *arXiv preprint arXiv:2306.11970* (2023).

Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. 2023. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9601–9611.

Jack M Wang, David J Fleet, and Aaron Hertzmann. 2007. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence* 30, 2 (2007), 283–298.

Xuesong Wang, Ke Jin, Yi Kong, CL Philip Chen, and Yuhu Cheng. 2022. Discriminator-quality evaluation GAN. *IEEE Transactions on Multimedia* 25 (2022), 4081–4093.

Andrew Witkin and Michael Kass. 1988. Spacetime constraints. *ACM Siggraph Computer Graphics* 22, 4 (1988), 159–168.

Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. 2023. ActFormer: A GAN-based transformer towards general action-conditioned 3D human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2228–2238.

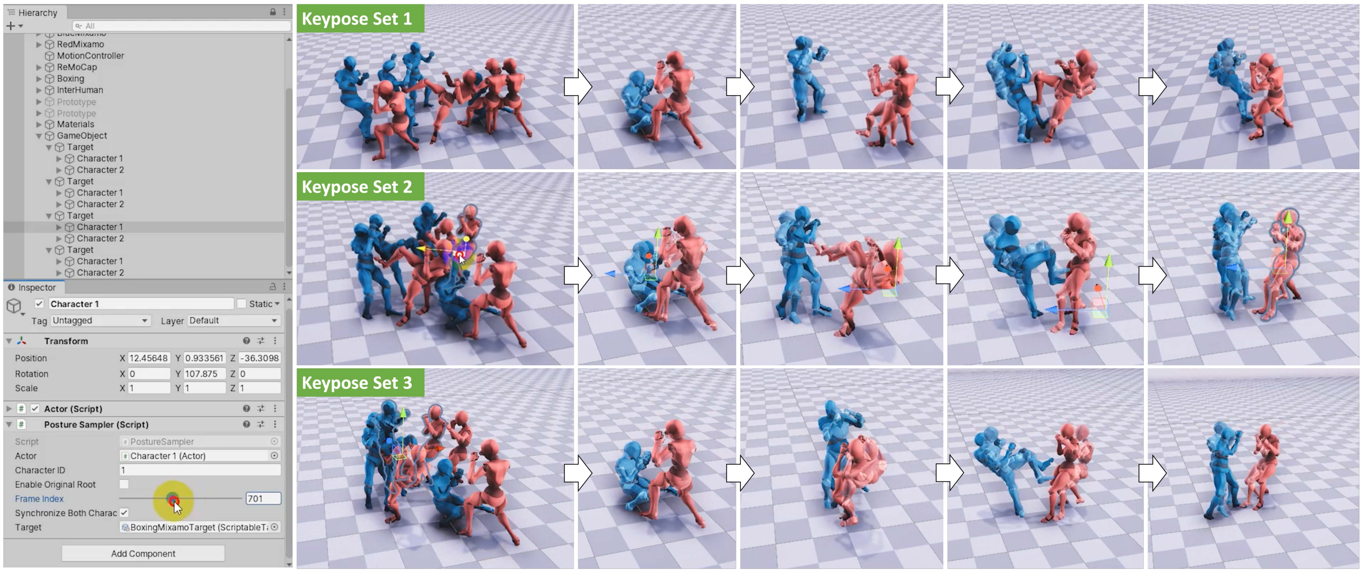
Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. 2024. ReGenNet: Towards Human Action-Reaction Synthesis. *arXiv preprint arXiv:2403.11882* (2024).

Kwan Yun, Seokhyeon Hong, Chaelin Kim, and Junyong Noh. 2025. AnyMoLe: Any Character Motion In-betweening Leveraging Video Diffusion Models. https://doi.org/10.48550/arXiv.2503.08417 arXiv:2503.08417 [cs].

Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 28–35.

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. 2023. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.



☑ Character 1 □ Static

Tag Untagged Layer Default

Transform

Position X 12.45648 Y 0.933561 Z -36.3098

Rotation X 0 Y 107.875 Z 0

Scale X 1 Y 1 Z 1

☑ Actor (Script)

Posture Sampler (Script)

Script PostureSampler

Actor Character 1 (Actor)

Character ID 1

Enable Original Root □

Frame Index 701

Synchronize Both Charac ☑

Target BoxingMixamoTarget (ScriptableT

Fig. 10. By manipulating the root translations and rotations of keyposes via the control panel (left), our system synthesizes interactive in-between motion sequences that dynamically respond to the specified keypose configurations. Detailed animation results are provided in the supplementary video.

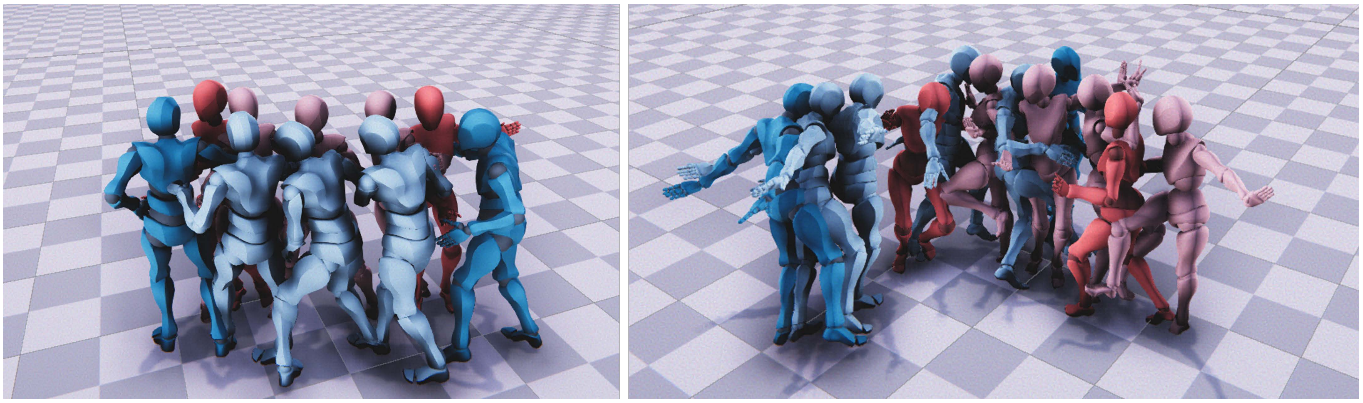


Fig. 11. Qualitative results on ReMoCap dataset. Light blue and pink characters are in-between motion. Blue and red characters are keyposes.

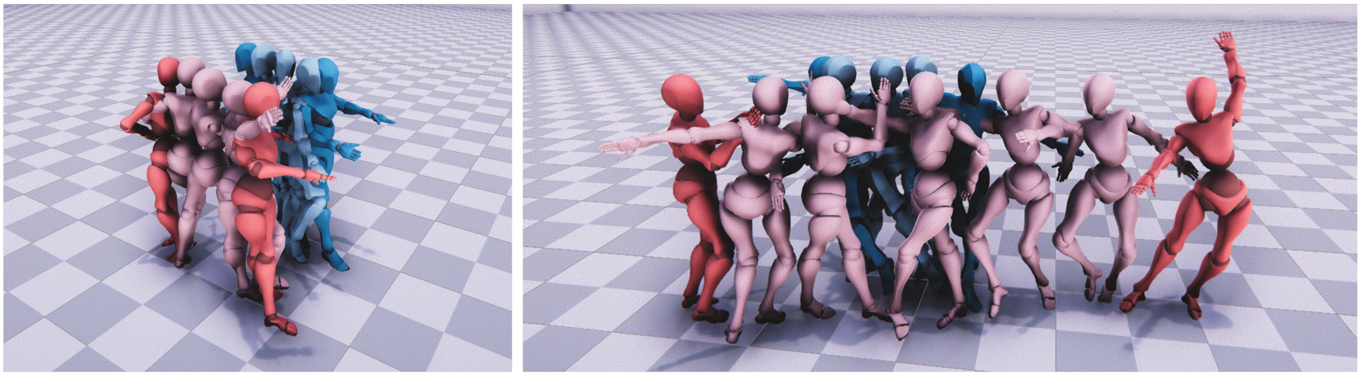


Fig. 12. Qualitative results on InterHuman dataset. Light blue and pink characters are in-between motion. Blue and red characters are keyposes.

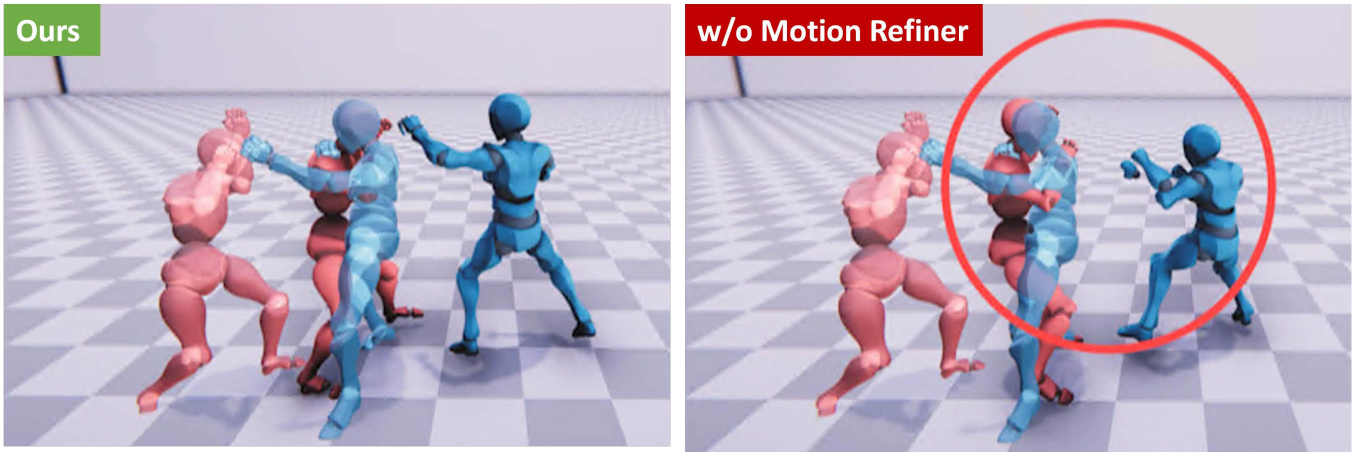


Fig. 13. Qualitative results without the Motion Refiner. The blue character exhibits hand joint deformation after a few seconds of prediction.

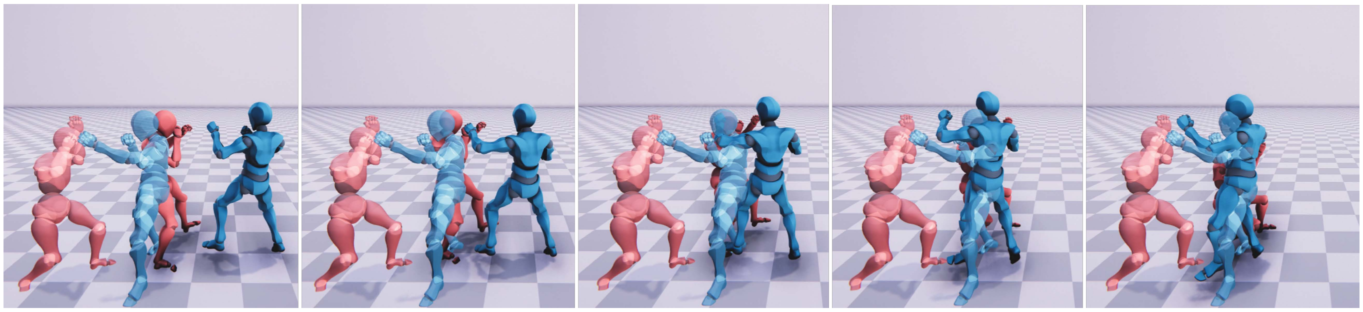


Fig. 14. Qualitative results without modeling interaction periodicity. The blue character is sliding to its keypose without interactive movement.

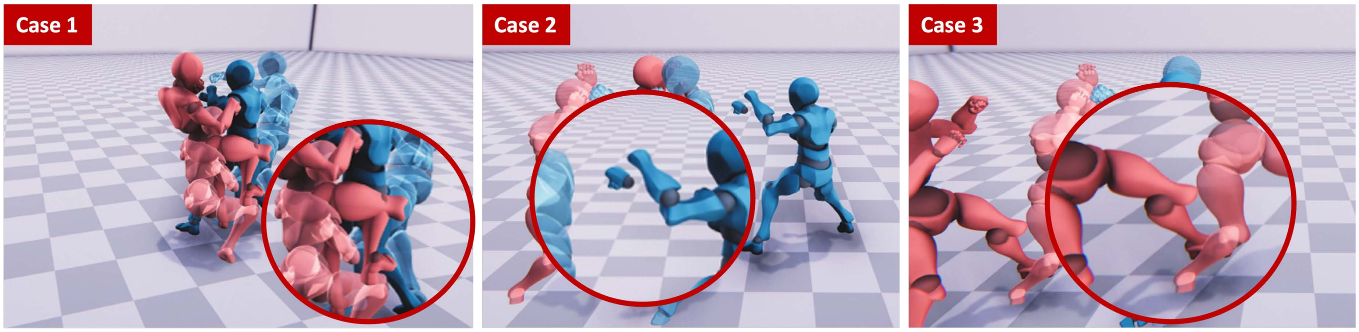


Fig. 15. Examples of deformed bones caused by long-term error accumulation.