InterMamba: Efficient Human-Human Interaction Generation With Adaptive Spatio-Temporal Mamba

Zizhao Wu, Yingying Sun, Yiming Chen, Xiaoling Gu, Ruyu Liu, and Jiazhou Chen

***Abstract*—Human-human interaction generation has garnered significant attention in motion synthesis due to its vital role in understanding humans as social beings. However, existing methods typically rely on transformer-based architectures, which often face challenges related to scalability and efficiency. To address these challenges, we propose InterMamba, a novel and efficient human–human interaction generation method built on the Mamba framework, designed to capture long-sequence dependencies effectively while enabling real-time feedback. Specifically, we introduce an adaptive spatio-temporal Mamba framework that utilizes two parallel SSM branches with an adaptive mechanism to integrate the spatial and temporal features of motion sequences. To further enhance the model's ability to capture dependencies within individual motion sequences and the interactions between different individual sequences, we develop two key modules: the self adaptive spatio-temporal Mamba module and the cross adaptive spatio-temporal Mamba module, enabling efficient feature learning. Extensive experiments demonstrate that our method achieves the state-of-the-art results on both two interaction datasets with remarkable quality and efficiency. Compared to the baseline method InterGen, our approach not only improves accuracy but also reduces the parameter size to just 66 M (36% of InterGen's), while achieving an average inference speed of 0.57 seconds, which is 46% of InterGen's execution time.**

***Index Terms*—Human-Human Interaction, Human Motion Synthesis, Text-driven Motion Generation, Mamba.**

I. Introduction

HUMAN-HUMAN interaction generation (HHI) is a crucial component in many graphics and vision applications, such as gaming, virtual simulation, and embodied AI. In recent years, with the rapid development of deep generative models [47], [50], [53], particularly diffusion models [16], significant progress has been achieved for single-person human motion generation in terms of controllability [20], [48], diversity [12], and quality [56]. Building on these advancements, researchers have begun exploring more complex, interaction-centered human motion generation [24], [41], [42], including Human-Human interaction generation.

Compared to single-human motion generation, human-human interaction generation is more challenging, as it requires simultaneously ensuring high-quality individual motion generation while accurately modeling interactions between individuals. To address these issues, significant efforts have been made, leading to notable progress. For example, InterGen [24] introduces a diffusion-based transformer framework, where each individual's latent features are conditioned on those of others through cross-attention mechanisms that share weights, yielding promising results. ComMDM [42] employs lightweight neural layers to bridge pre-trained single-person motion diffusion models, effectively enabling multi-person interaction modeling. InterMask [18] proposes a masked VQVAE framework for the task in the discrete space by utilizing the VQVAE [47] and Inter-M transformer. However, we note that most existing methods are built upon the Transformer architecture. While Transformer-based approaches effectively model intricate contextual relationships, benefiting from their powerful attention mechanism, they also inherit the drawback of quadratic complexity, which limits their scalability and efficiency for long sequences. Although various techniques such as windowing [29], sliding [33], sparsification [39], hashing [31], Ring Attention [26], and Flash Attention [5]—or their combinations—have been introduced to mitigate the quadratic complexity of the Transformer, it remains a bottleneck in terms of scalability and efficiency.

Recently, Structured State Space Models (SSMs) have gained attention for their efficiency and effectiveness in handling long sequences. Mamba [9], [10], building on this foundation, offers a promising solution for capturing long-range dependencies by introducing input selectivity and leveraging the scan method. Unlike transformers, which scale quadratically with sequence length, Mamba achieves linear or near-linear complexity while maintaining strong performance on extended sequences. This advantage has positioned it as a leading approach in continuous long-sequence analysis, delivering the state-of-the-art results in areas like natural language processing and computer vision. Therefore, it is natural to introduce Mamba into the human

Received 22 April 2025; revised 9 November 2025; accepted 14 November 2025. Date of publication 21 November 2025; date of current version 6 February 2026. This work was supported in part by the Zhejiang Province Leading Geese Science and Technology Program under Grant 2025C02155, in part by NSFC under Grant 62471168, Grant 62202137, and Grant 62172367, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LMS25F020009. Recommended for acceptance by J. Noh. *(Zizhao Wu and Yingying Sun are co-first authors.) (Corresponding author: Zizhao Wu.)*

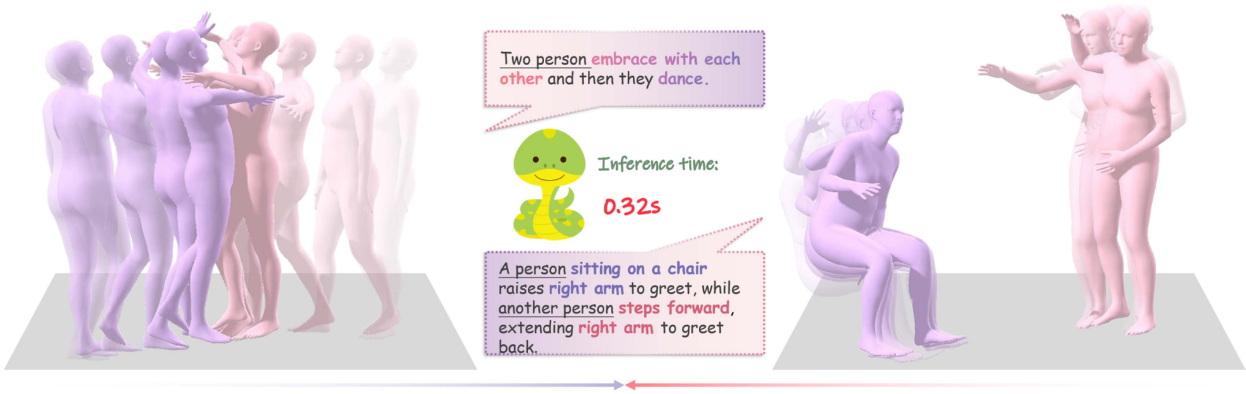
Zizhao Wu, Yingying Sun, and Yiming Chen are with the Department of Digital Media Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: wuzizhao@hdu.edu.cn; yingyingsun@hdu.edu.cn; chenyiming@hdu.edu.cn).

Xiaoling Gu is with the Department of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: guxl@hdu.edu.cn).

Ruyu Liu is with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: lry@hznu.edu.cn).

Jiazhou Chen is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: cjz@zjut.edu.cn).

Digital Object Identifier 10.1109/TVCG.2025.3635116



In this paper, we introduce an efficient human-to-human interaction generation method based on the Mamba framework, designed to achieve real-time, high-fidelity, text-driven dual-person motion synthesis.

motion generation paradigm to solve the scalability problem of existing transformer-based methods.

However, directly applying the existing Vanilla Mamba to our human-human interaction framework presents the following two challenges: Firstly, the vanilla Mamba module is specialized in 1D sequence modeling; how to extend it to process the spatio-temporal 3D motion sequences remains a challenge. The key lies in achieving an accurate fusion of spatio-temporal features while maximizing efficiency. Secondly, due to the complex interactions between human motions, another critical challenge is how to extend Mamba to model intricate multi-person interactions and communication effectively.

To tackle the above challenges, we propose InterMamba, an Adaptive Spatio-Temporal Mamba (ASTM) framework for human–human interaction generation. InterMamba leverages SSM operations along both temporal and spatial dimensions, with adaptive parameters that dynamically adjust their respective contributions, thereby achieving high scalability and efficiency in modeling the context of long-range sequences. To achieve precise and efficient modeling of spatiotemporal single-person motion sequences and complex human-to-human motion interactions, we propose a Self-ASTM module and a Cross-ASTM module, respectively. Specifically, the Self-ASTM captures long-range dependencies within individual characters by leveraging attention-based spatio-temporal Mamba, while the Cross-ASTM explicitly establishes connections between paired individuals and facilitates interaction-aware information exchange through the Local Interaction Information Aggregation (LIIA).

We conduct our experiments on two famous benchmark datasets: InterHuman [24] and InterX [52], and the results have demonstrated that our approach achieves superior performance and lower computational footprint (as shown in Fig. 1), setting new precedents for efficient, scalable human-human interaction generation.

In summary, we make the following three contributions in this paper:

- We propose InterMamba, a novel Mamba-based framework for human-to-human interaction generation. By leveraging adaptive spatio-temporal SSMs, our approach effectively captures long-range spatial and temporal dependencies, providing a highly efficient framework for modeling complex interactions.
- We explore the Self-ASTM and the Cross-ASTM modules, which are specifically designed to capture long-range dependencies within individual motion sequences and model the intricate interactions between individuals, respectively.
- Our method achieves the state-of-the-art results, demonstrating significant improvements in both generation quality and real-time efficiency.

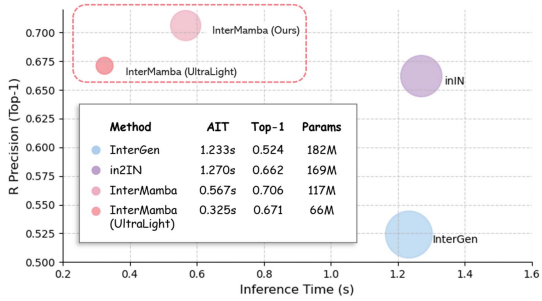


Fig. 1. This figure presents a comparative analysis of different methods on the Interhuman dataset in terms of R-Precision (Top-1) and inference time, with the bubble size indicating the number of model parameters. From the figure, our InterMamba and InterMamba (UltraLight), which are highlighted in a red dashed box, achieve superior performance compared to other methods.

II. Related Work

In this section, we briefly review relevant literature on single-human motion generation, human-human interaction generation, as well as SSMs.

A. Single-Human Motion Generation

Single-human Motion Generation involves synthesizing single human movements based on various control conditions, including past motions [15], [30], [35], [50], music [1], [22], [46], action labels [13], [37], scenes [3], [48], and text [11], [20], [23], [40]. Among these conditions, text-driven motion generation

has gained significant attention due to its ability to provide a more intuitive and accessible way to represent human intent. Methods like MotionCLIP [44] integrate CLIP-based semantic knowledge into motion representations, while approaches such as TEMOS [38] and T2M [12] employ Transformer-based VAEs to generate motion from textual descriptions. Further advancements, including AttT2M [62] and TM2D [8], incorporate spatial-temporal body-part encoders into VQ-VAE to enhance the expressiveness of discrete motion representations.

Recently, diffusion models have emerged as a dominant paradigm for human motion generation due to their strong generative capabilities and diversity. Methods like Motion Diffusion Model (MDM) [45], MotionDiffuse [54], and FLAME [20] adopt diffusion models with Transformer-based backbones to enable various conditioning strategies for coherent motion synthesis. Since these approaches operate directly on raw motion data, making them computationally expensive and susceptible to noise. To address this, MLD [4] introduces latent-space diffusion to improve motion quality, OmniControl [51] and FineMoGen [56] enhance text-driven motion generation with more precise spatial-temporal control. Additionally, ReMoDiffuse [55] and Make-An-Animation [2] extend the generalization ability of diffusion models by incorporating large-scale in-the-wild data, such as human pose images and videos, into the training process.

Recent research highlights that human motions in real-world applications are inherently influenced by interactions with surrounding objects, other individuals, and the environment. These findings have led to growing interest in interaction-centered motion generation, encompassing tasks such as human-object interaction [21], [36], human-human interaction [24], [41], [42], and human-scene interaction [19], [49]. In this work, we focus specifically on the task of human-human interaction generation.

B. Human-Human Interaction Generation

Compared to single-human motion generation, human-human interaction generation is more challenging, as it requires simultaneously ensuring high-quality individual motion generation while accurately modeling interactions between individuals. Many researchers have made valuable contributions to address this challenge. For example, building on the strong performance of MDM [45], ComMDM [42] applies fine-tuning and minor modifications to adapt it for various motion-related tasks. Notably, ComMDM pioneers the task of human-human interaction motion generation by incorporating a communication transformer block between two instances of a pre-trained MDM [45]. InterGen [24] makes a significant contribution to this area by introducing a dedicated human interaction dataset and proposing an interaction-aware diffusion model. This model leverages a novel weight-sharing mechanism with a global motion representation, achieving high-quality human interaction synthesis. The in2IN [41] further advances the field by introducing a diffusion model that conditions motion generation not only on overall interaction descriptions but also on individual actions. This approach enhances both intra-person motion diversity and inter-person coordination, improving the realism and coherence of generated interactions.

Although the above-mentioned methods have achieved promising results, we observe that they are all built on the Transformer architecture, which suffers from quadratic time complexity, posing limitations in handling long motion sequences and computational efficiency. Therefore, this paper focuses on addressing the scalability and efficiency challenges of existing approaches by proposing a novel Mamba-based human-human interaction motion generation method.

C. Selective State Space Models

Recently, State Space Models (SSMs) [7], [9], [10], [14] have demonstrated remarkable performance in handling long-context tasks, offering both high-quality generation and efficient inference. However, the earlier SSMs struggled with effectively modeling discrete and information-dense data. To address this limitation, Mamba [9] introduced a data-dependent SSM layer and a selection mechanism utilizing parallel scan (S6), significantly enhancing its applicability across various domains.

Since then, Mamba has sparked widespread interest in the deep learning community, leading to a surge of research in computer vision. SSMs are now proving their vast potential across diverse applications. For instance, Vim [63] presents a novel vision backbone integrating bidirectional Mamba blocks for visual representation learning, achieving modeling capabilities comparable to ViT [6] while reducing computational complexity to subquadratic time with linear memory requirements. VMamba [28] introduces 2D Selective Scan (SS2D), a four-way scanning mechanism designed for spatial domain processing, ensuring that each image patch captures rich contextual information. VideoMamba [34] extends these concepts to video understanding, combining the strengths of convolution and attention within a pure SSM framework, demonstrating efficiency and effectiveness in both short-term and long-term video tasks. Similarly, PointMamba [27] adapts SSMs for 3D vision applications by enabling global modeling with linear complexity, making it a compelling choice for point cloud analysis. In the motion domain, MotionMamba [59] introduces a hierarchical temporal Mamba block alongside a bidirectional spatial Mamba block to facilitate real-time and high-quality single-human motion generation. InfiniMotion [58] focuses on generating long motion sequences with improved memory capacity. Motion Avatar [61] extends motion generation to human and animal forms with generalizable avatar representations. KMM [57] incorporates keyframe masking to guide motion synthesis, and Motion Anything [60] proposes a general-purpose motion generation pipeline across various input modalities.

In this work, we introduce Mamba into the human-human interaction motion generation paradigm for the first time. We show that, compared to other methods, our approach achieves significant improvements in both generation quality and efficiency.

III. Preliminaries

A. Diffusion Model

Following DDPM [16] and InterGen [24], our model adopts a diffusion framework with forward and reverse processes for

motion generation. The forward process adds time-dependent noise to the motion data $p_0(x)$, transforming it into $p_t(x)$ over t steps until it approximates $\mathcal{N}(0, \mathbf{I})$. The reverse process removes the noise to reconstruct $p_0(x)$ via a generative model parameterized by θ. According to [16], we simplify the multi-step diffusion process into a single step, expressed as:

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{1}$$

where $\bar{\alpha}_t \in (0,1)$ follows a monotonically decreasing schedule. As t approaches infinity, $\bar{\alpha}_t \to 0$, causing x_t to approach a standard normal distribution $\mathcal{N}(0, \mathbf{I})$. The reverse process aims to learn f_θ to iteratively reconstruct the motion, generating $\hat{x}_0$ conditioned on the input text c. Instead of predicting the noise, we directly estimate $\hat{x}_0$, following the approach in [54]. The training objective is defined as:

$$\mathscr{L}_{\text{t}} = \mathbb{E}_{x_0,t}\left[\|x_0 - f_\theta(x_t, t, c)\|_2^2\right]. \tag{2}$$

B. Mamba Model

Selective state space models (SSMs) have emerged as a promising approach for sequence modeling, offering efficient computation and compact data storage while achieving strong performance across diverse domains. Inspired by particular continuous system, traditional SSMs maps a 1-dimensional function or sequence $x(t)$ to $y(t)$ through an implicit latent state $h(t)$:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ denotes the evolution parameter, and $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$ denotes the projection parameters. However, these parameters are time-invariant matrices and vectors, which constrains capacities on modeling discrete and information-dense data. To resolve this problem, Mamba proposes a novel selection mechanism that parameterizes the SSM parameters based on the input data to affect interactions along the sequence. The rewritten architecture formulation is defined as below:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}_t h_{t-1} + \bar{\mathbf{B}}_t x_t \\ y_t &= \mathbf{C}_t h_t, \end{aligned} \tag{4}$$

where $\bar{\mathbf{A}}_t$, $\bar{\mathbf{B}}_t$, and $\mathbf{C}_t$ are dynamically updating matrixs and vectors by the time step, meaning that the model has changed from time-invariant to time-varying. As the discrete versions of the continuous system, Mamba [9] transform the continuous parameters $\mathbf{A}, \mathbf{B}$ to discrete parameters $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t$ with a timescale parameter Δ. The discretization rule, zero-order hold (ZOH), is defined as follows:

$$\begin{aligned} \bar{\mathbf{A}} &= exp(\Delta\mathbf{A}) \\ \bar{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}, \end{aligned} \tag{5}$$

where $(\Delta\mathbf{A})^{-1}$ denotes the inverse of matrix $\Delta\mathbf{A}$, $\mathbf{I}$ denotes the identity matrix. Additionally, Mamba imply a Selective Scan Mechanism (SSMs) as its core SSM operator. Thus, mamba makes these parameters change depending on the input which can be formulated as:

$$\begin{aligned} \mathbf{B} &= Linear_N(x) \\ \mathbf{C} &= Linear_N(x) \\ \Delta &= \tau_\Delta(P + LayerNorm(Linear_1(x))), \end{aligned} \tag{6}$$

where P is a learnable parameter, τ_Δ is SoftPlus function, and $Linear_d(\cdot)$ is a parameterized projection to dimension d. The parameters $\mathbf{B} \in \mathbb{R}^{(B,L,N)}$, $\mathbf{C} \in \mathbb{R}^{(B,L,N)}$, and $\Delta \in \mathbb{R}^{(B,L,D)}$ are directly derived from the input data $x \in \mathbb{R}^{(B,L,D)}$.

IV. Method

Our objective is to generate a high-quality human-human interaction motion sequence that aligns with the action described in the text prompt while maintaining high efficiency. In our implementation, we use $\{x_p\} = \{x_p^i\}_{i=1}^n, p \in \{a, b\}$ to denote the outcome motion sequences, where n is the frame number of the sequence. For each pose $x_p^i, p \in \{a, b\}$, we define it as: $x_p^i = [j^g, v^g, r^l, c^f], p \in \{a, b\}$, where $j^g \in \mathbb{R}^{3J}$ denotes the global joint positions, $v^g \in \mathbb{R}^{3J}$ represents the global velocities, $r^l \in \mathbb{R}^{6J}$ refers to 6D representation of local rotations and $c^f \in \mathbb{R}^4$ corresponds to the binary foot-ground contact features. Fig. 3 illustrates the pipeline of our method, which mainly consists of Adaptive Spatio-Temporal Mamba modules, Self-Role Mamba Blocks, Cross-Role Mamba Blocks and LIIA interaction modules. Specifically, given a text prompt, our method first employs the CLIP encoder to achieve the textual features embedding. Then the textural features are combined with the initial hidden states $h_a^{(t)}$ and $h_b^{(t)}$ separately and are fed into our framework. Within the framework, our method employs the Self-ASTM Blocks for modeling the long-range motion contexts of individuals and Cross-ASTM Blocks for modeling the intricate interaction contexts between individuals. To further enhance the representation, we exploit the LIIA module that aggregates human-human interaction representations. After N iterations, our method ultimately generates an efficient and high-quality human-human interaction sequence that aligns with the semantics of the text input.

In the following, we begin with the introduction of the backbone of our framework. In Section IV-B, we detail the Self-ASTM module. Next, we present LIIA module in Section IV-C. In Section IV-D, we introduce the Cross-ASTM module. Finally, in Section IV-E we provide the objective function of our method.

A. Adaptive Spatio-Temporal Mamba

As described in Section III, the Mamba model is inherently well-suited for sequential modeling due to its selection architectural design. However, its feature extraction is primarily focused on the temporal dimension, lacking explicit modeling of spatial dynamics. Consequently, Mamba exhibits performance bottlenecks in such spatio-temporal entangled scenarios. This limitation becomes apparent when addressing complex tasks such as 3D human interaction generation. To overcome this challenge, we propose a novel parallel spatio-temporal modeling framework named Adaptive Spatio-Temporal Mamba (ASTM). ASTM decomposes the modeling process into two complementary branches: a Spatial State Space Model (Spatial SSM)

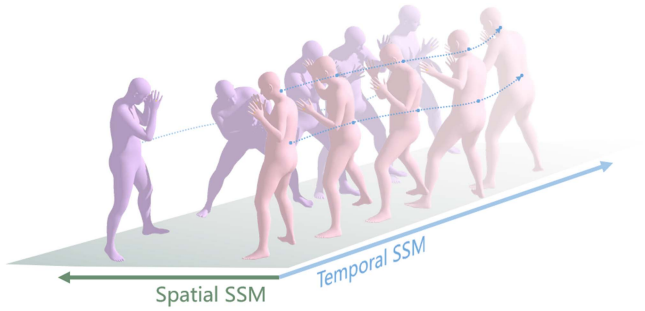


Fig. 2. The spatial and temporal scanning process, where the spatial SSM captures intra-frame joint relationships and the temporal SSM models inter-frame relationships.

and a Temporal State Space Model (Temporal SSM). Specifically, as illustrated in Fig. 2, Spatial SSM focuses on capturing intra-frame spatial dependencies, effectively extracting structural information such as the interactions between skeletal joints within each frame. In contrast, the Temporal SSM is dedicated to modeling inter-frame dynamics, capturing long-range motion trajectories and temporal causality. To be specific, given the input feature tensor $h \in \mathbb{R}^{(B,L,D)}$, where B is the batch size, L is the sequence length, and D is the feature dimension, the temporal SSM processes the input as follows:

$$h_t = LayerNorm(SSM_{temp}(Conv_{temp}(Linear(h)))), \tag{7}$$

where Linear$(\cdot)$ projects input to a latent space, $Conv_{spat}(\cdot)$ extracts local spatial patterns, $SSM_{temp}(\cdot)$ is SSM based module which captures long-range temporal dependencies, and LayerNorm$(\cdot)$ is layer normalization which improves training stability and convergence. Similar to the temporal branch, the spatial SSM adopts the same module structure but operates along the spatial dimension. The spatial output can be computed by:

$$h_s = LayerNorm(SSM_{spat}(Conv_{spat}(Linear(h')))), \tag{8}$$

where all modules are identical in structure, the input $h' \in \mathbb{R}^{(B,D,L)}$ is derived by transposing the input tensor to swap the sequence and feature dimensions. To adaptively integrate the learned spatial and temporal representations, we introduce a learnable fusion mechanism. Instead of naïvely concatenating or averaging features, we allow the model to dynamically reweight the contributions from h_s and h_t. The final fused representation is computed as:

$$z = w_\alpha h_t + w_\beta h_s, \tag{9}$$

where w_α and w_β are jointly optimized during training and kept fixed during inference. These parameters are not shared across different ASTM blocks, each block maintains its own set of modulation weights to allow for layer-specific adaptations.

Benefiting from this decoupled modeling and adaptive fusion approach, ASTM significantly enhances the model's capacity to represent the spatio-temporal characteristics of complex motion, leading to improved structural coherence and temporal consistency in human interaction generation tasks.

B. Self-ASTM Block

To enable the model to effectively learn the motion features of each individual in human-human interaction, we designed the self Adaptive Spatio-Temporal Mamba Block (Self-ASTM) based on the ASTM, as illustrated in Fig. 3. This module processes the motion of each individual separately as input and leverages ASTM to extract the most salient spatio-temporal features for each character. Specifically, given the motion feature of a single individual $h_p^{(t)}, p \in \{a, b\}$ as input, the output feature can be calculated as:

$$\begin{aligned}
\bar{h}_{\mathrm{p}} &= LayerNorm(h_p), \\
\hat{h}_p &= ASTM(\bar{h}_{\mathrm{p}}), \\
q &= \sigma(Linear(\bar{h}_{\mathrm{p}})), \\
\widetilde{h}_p &= \hat{h}_p \star q, \\
h_p^S &= h_p + Linear(\widetilde{h}_p),
\end{aligned} \tag{10}$$

where $LayerNorm(\cdot)$ is the layer normalization, $Linear(\cdot)$ is the linear layer for feature mapping, σ is the activation function and the SiLU function is used in our experiment. $ASTM(\cdot)$ can be calculated from Section III-B.

C. Local Interaction Information Aggregation

Information in dyadic interactions is typically highly abstract and densely entangled. However, existing methods often rely on raw individual features and utilize mutual attention mechanisms to model feature-level relationships. Those approaches are prone to introducing noise and struggle to align deeper semantic representations and interactive information. To address this issue, we investigate Local Interaction Information Aggregation(LIIA) module, as illustrated in Fig. 5, which consists of two convolution layers and a residual connection. Given the motion feature of a single individual $h_p, p \in \{a, b\}$ and text feature embeddings $cond_t$, the output joint representation h_{inter} can be formulated as:

$$\begin{aligned}
h_{ab} &= concat(h_a^S, h_b^S), \\
\hat{h}_{ab} &= AdaLN(h_{ab}, cond_t), \\
h_{inter} &= conv_{3\times3}(conv_{1\times1}(\hat{h}_{ab})).
\end{aligned} \tag{11}$$

where $conv_{k\times k}$ denotes the convolution with kernel size k, $AdaLN(\cdot)$ is the adaptive layer normalization to reduce the impact of noise features on the input and optimize the training process. We note that this strategy not only facilitates the modeling of local feature interactions but also ensures a more comprehensive and robust integration of contextual information.

D. Cross-ASTM Block

While the Self-ASTM efficiently models individual motion features, it overlooks the interaction details between individuals, which are crucial for human-human interaction generation task. To address this problem, inspired by [25], we introduce

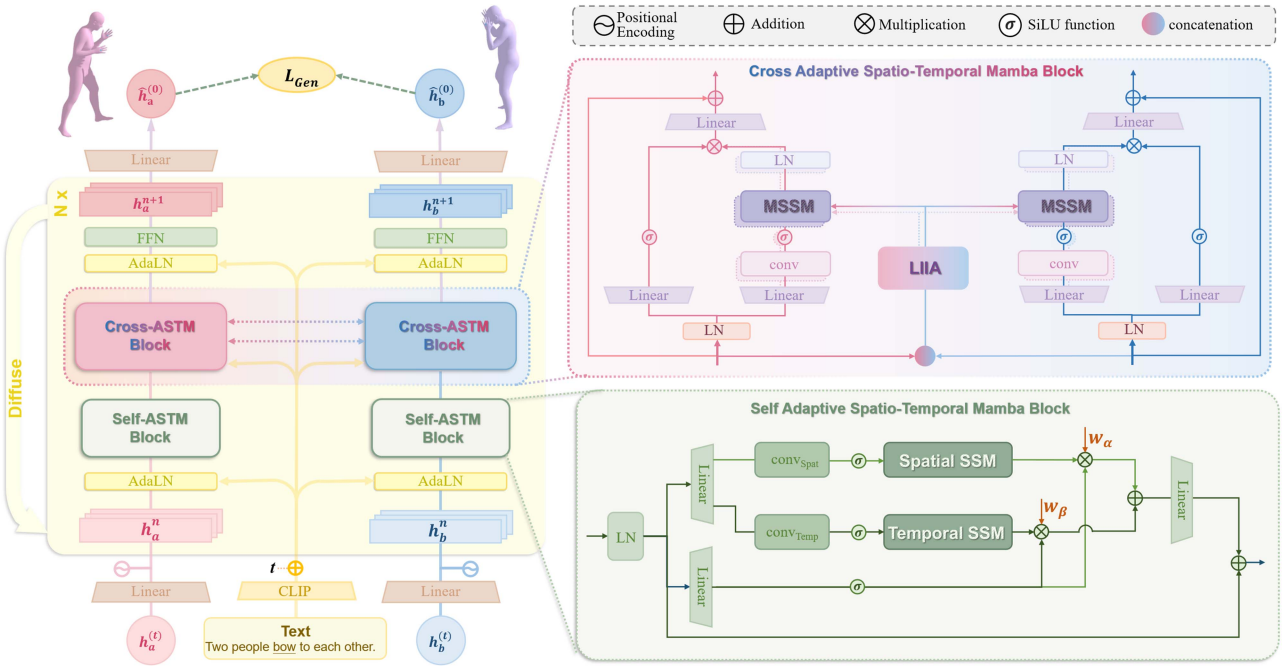
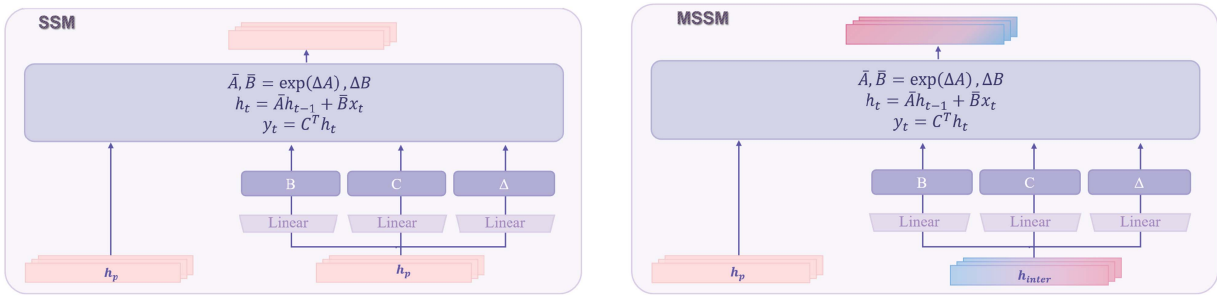


Fig. 3. **The framework of InterMamba.** The key components include Self Adaptive Spatio-Temporal Mamba module (Self-ASTM), Cross Adaptive Spatio-Temporal Mamba module (Cross-ASTM), and Local Interaction Information Aggregation module (LIIA).



(a) Schematic Diagram of the State Space Model (SSM)

(b) Schematic Diagram of the Mixed State Space Model (Mix-SSM)

Fig. 4. (a) This figure illustrates the core module of the ASTM in Section IV-B —- **State Space Model (SSM)**. Given an input sequence h_p, the input first undergoes linear transformations through (5) to calculate the parameters B, C, and Δ. Then, the final output can be calculated through (4) and (5). (b) This figure illustrates the core module of the $\text{ASTM}_{\text{cross}}$ in Section IV-D —- **Mix State Space Model (MSSM)**. This module, based on the State Space Model (SSM), enables each individual to perceive the mutual interaction features dynamically, leading to improved modeling of contextual dependencies. Given h_p and h_{inter}, which is calculated from Section IV-C, the parameters B, C, and Δ can be calculated through x_{inter}, and x_t is the input sequence of the MSSM. Then, the final output can be calculated through (4) and (5). The details of Mix SSM are provided in Section IV-D. (a) Schematic Diagram of the State Space Model (SSM). (b) Schematic Diagram of the Mixed State Space Model (Mix-SSM).

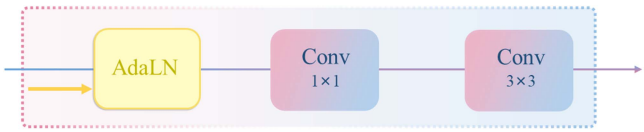


Fig. 5. The structure of Local Interaction Information Aggregation (LIIA) consists of AdaLN for adaptive normalization, followed by 1× 1 and 3× 3 convolutions for feature refinement and local interaction modeling.

a novel cross Adaptive Spatio-Temporal Mamba block (Cross-ASTM) which extends the Self-ASTM to explicitly model cross-individual interaction relationships. As illustrated in Fig. 3, the overall architecture of the cross-ASTM mirrors that of the Self-ASTM, with one key difference: the original $ASTM$ module is replaced but the newly designed $ASTM_{cross}$ module. Specifically, given the inputs $h_p^S \in \mathbb{R}^{(B,L,D)}$ and $h_{inter} \in \mathbb{R}^{(B,L,D)}$, obtained through the Self-STM and LIIA, the output of the Cross-ASTM block can be calculated as:

$$
\begin{aligned}
\bar{h}_p &= \text{LayerNorm}(h_p^S), \\
\hat{h}_p &= ASTM_{cross}(\bar{h}_p, h_{inter}), \\
q &= \hat{\sigma}(Linear(\bar{h}_p)), \\
\widetilde{h}_p &= \hat{h}_p \star q, \\
h_{\text{out}} &= h_p^S + \text{Linear}(\widetilde{h}_p),
\end{aligned} \tag{12}
$$

where $LayerNorm(\cdot)$ is the layer normalization, $Linear(\cdot)$ is the linear layer for feature mapping, $\hat{\sigma}$ is the activation function. Within this module, the $ASTM_{cross}$ function is designed to extract spatio-temporal features by jointly considering both individual representations and their interactions. It is based on the Mix State Space Model (MSSM) architecture and extends it by introducing two parallel branches dedicated to spatial and temporal modeling, respectively. The internal computation of this mechanism is formally defined as follows:

$$\begin{aligned} c_s &= LayerNorm\left(MSSM_{temp}\left(Conv_{temp}\right.\right. \\ &\quad \left.\left. \times \ (Linear(h_p^S), h_{inter})\right)\right), \\ c_t &= LayerNorm\left(MSSM_{soat}\left(Conv_{spat}\right.\right. \\ &\quad \left.\left. \times \ (Linear(h_p^{S'}), h'_{inter})\right)\right), \\ \hat{h}_p &= \alpha_c * c_s + \beta_c * c_t. \end{aligned} \tag{13}$$

where the $h_p^{S'} \in \mathbb{R}^{(B,D,L)}$ and $h'_{inter} \in \mathbb{R}^{(B,D,L)}$ are obtained by swapping the last two dimensions of h_p^S and h_{inter}, respectively, so that the inputs are aligned along the spatial dimension. The parameters α_c and β_c are learnable scalars, consistent with the definition in (9).

The modules $MSSM_{\text{spat}}$ and $MSSM_{\text{temp}}$ are instantiated from the base MSSM, which extends the standard State Space Model (SSM) by integrating interaction information. Unlike conventional SSMs that operate solely on individual feature sequences, MSSM simultaneously processes both individual features and interaction features for enhanced feature fusion. Specifically, the interaction features h_{inter} are used to generate the parameters (B, C, Δ), while the individual features h_p serve as the query input, as defined in (6). The output is computed using (4) and (5). This design enables the model to better capture and emphasize interaction-relevant information. The architecture of MSSM is illustrated in Fig. 4.

E. The Objective Function

We adopt the same loss functions as those used in InterGen [24], which encompass several loss functions, and can be formulated as:

$$\begin{aligned} \mathscr{L}_{total} &= \mathscr{L}_{diff} + \lambda_{vel}\mathscr{L}_{vel} + \lambda_{foot}\mathscr{L}_{foot} + \lambda_{BL}\mathscr{L}_{BL} \\ &\quad + \lambda_{DM}\mathscr{L}_{DM} + \lambda_{RO}\mathscr{L}_{RO}, \end{aligned} \tag{14}$$

where $\mathscr{L}_{diff}$ refers to the diffusion loss, $\mathscr{L}_{vel}$ denotes the joint velocity loss, $\mathscr{L}_{foot}$ refers to the foot contact loss, $\mathscr{L}_{BL}$ represents the bone length loss, $\mathscr{L}_{DM}$ corresponds to the masked joint Distance Map (DM) loss and $\mathscr{L}_{RO}$ signifies the relative orientation loss. All loss terms are weighted by hyper-parameters $\lambda_{vel}, \lambda_{foot}, \lambda_{BL}, \lambda_{DM}$ and λ_{RO}, that balance their influence on the final result.

V. Experiments

In this section, we first outline the details of our experimental setup in Section V-A, followed by both quantitative and qualitative evaluations to assess the effectiveness of our method in Section V-B. In Section V-D, we present a comprehensive ablation study.

A. Experiment Setup

Datasets: We conduct our experiments on two publicly available benchmark datasets: InterHuman [24] and InterX [52]. InterHuman [24] is the first human-human interaction dataset with natural language annotations. It contains 6,022 motion sequences spanning various categories of human actions, annotated with 16,756 unique descriptions composed of 5,656 distinct words, with a total duration of 6.56 hours. InterX [52] adopts the SMPL-X representation, which includes 55 joints covering the face, body, and hands. The dataset consists of 11,000 interaction sequences, accompanied by more than 34,000 fine-grained textual annotations at the body-part level, totaling over 8.1 million frames.

Implementation Details: Our framework consists of two stacked ASTM modules (with an ultra-light variant containing only one ASTM module). Each ASTM module employs a state size of 16. For text encoding, we utilize the frozen CLIP-ViT-L/14 model. The diffusion process is trained with 1,000 timesteps, while inference employs the DDIM sampling strategy with 50 timesteps and $\eta = 0$. We adopt the cosine noise schedule [32] and classifier-free guidance [17], setting random CLIP embeddings to zero with a probability of 10% during training and using a guidance coefficient of 3.5 during sampling. The model is optimized using the AdamW optimizer with a learning rate of 10^{-4} and a weight decay of 2×10^{-5}. All models are trained for 2,000 epochs with a batch size of 50 on two NVIDIA RTX 4090 GPUs. And the inference is conducted on a single NVIDIA RTX 4090 GPU.

Evaluation Metrics: We adopt the same evaluation metrics as InterGen [24] to assess our model's performance, which include: (1) **Frechet Inception Distance (FID):** measures the discrepancy in latent distributions between the generated and real datasets. (2) **R-Precision:** evaluates text-motion alignment by determining the probability of the correct text appearing in the top-k results ($k = 1, 2, 3$) after ranking. (3) **Diversity:** quantifies the variation in generated motions within the dataset. (4) **Multimodality (MModality):** assesses the diversity of motions generated for the same text input. (5) **Multi-modal Distance (MM Dist):** calculates the distance between motion features and their corresponding text features.

B. Comparison With Baselines

1) Quantitative Comparison: To ensure a comprehensive comparison, we evaluate our approach against several baseline methods, including TEMOS [38], T2M [12], MDM [45], ComMDM [42], InterGen [24], RIG [43] and in2IN [41]. Among these methods, only InterGen [24] and in2IN [41] are originally designed for human–human interaction generation, InterGen [24] represents a milestone in the field as it introduces the first human-human interaction dataset, while in2IN [41] is the most recent advancement. Therefore, we mainly discuss the performance of our method against these two approaches.

TABLE I
THE QUANTITATIVE COMPARISONS ON THE INTERHUMAN [24] TEST SET. WE RUN ALL THE EVALUATIONS 20 TIMES. ± INDICATES A 95% CONFIDENCE INTERVAL. ↑ (↓) INDICATES THAT A HIGHER (LOWER) RESULT CORRESPONDS TO BETTER PERFORMANCE, AND → MEANS THE CLOSER TO GROUND-TRUTH THE BETTER. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE REFERS TO THE SECOND BEST RESULT.

Methods	R Precision↑			*FID* ↓	*MMDist* ↓	*Diversity* →	*MModality* ↑
	Top1	Top2	Top3				
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.003}$	$3.755^{\pm.007}$	$7.948^{\pm.064}$	-
TEMOS [38]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$5.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M [12]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$4.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM [45]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$6.125^{\pm.018}$	$7.602^{\pm.045}$	$2.355^{\pm.080}$
ComMDM* [42]	$0.067^{\pm.013}$	$0.125^{\pm.018}$	$0.184^{\pm.015}$	$38.643^{\pm.098}$	$13.211^{\pm.013}$	$3.520^{\pm.058}$	$0.217^{\pm.018}$
ComMDM [42]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.011}$	$7.069^{\pm.054}$	$5.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
RIG [43]	$0.285^{\pm.010}$	$0.409^{\pm.014}$	$0.521^{\pm.013}$	$6.775^{\pm.069}$	$4.876^{\pm.018}$	$7.311^{\pm.043}$	$2.096^{\pm.065}$
InterGen [24]	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.012}$	$7.387^{\pm.029}$	$\mathbf{2.141}^{\pm.063}$
in2IN [41]	$0.425^{\pm.008}$	$0.576^{\pm.006}$	$0.662^{\pm.009}$	$\mathbf{5.535}^{\pm.120}$	$3.803^{\pm.002}$	$7.983^{\pm.047}$	$1.215^{\pm.023}$
InterMamba (Ours)	$\mathbf{0.475}^{\pm.005}$	$\mathbf{0.625}^{\pm.006}$	$\mathbf{0.706}^{\pm.004}$	$5.945^{\pm.007}$	$\mathbf{3.785}^{\pm.003}$	$\mathbf{7.963}^{\pm.033}$	$0.993^{\pm.026}$
InterMamba (light)	$0.443^{\pm.005}$	$0.592^{\pm.005}$	$0.671^{\pm.004}$	$8.082^{\pm.095}$	$3.801^{\pm.004}$	$7.991^{\pm.033}$	$0.965^{\pm.020}$

TABLE II
THE QUANTITATIVE COMPARISONS ON THE INTER-X [52] TEST SET. WE RUN ALL THE EVALUATIONS 20 TIMES. ± INDICATES A 95% CONFIDENCE INTERVAL. **BOLD** INDICATES THE BEST RESULT, AND UNDERLINE REFERS TO THE SECOND BEST RESULT.

Methods	R Precision↑			*FID* ↓	*MMDist* ↓	*Diversity* →	*MModality* ↑
	Top1	Top2	Top3				
Real	$0.429^{\pm.004}$	$0.626^{\pm.003}$	$0.736^{\pm.003}$	$0.002^{\pm.0002}$	$3.536^{\pm.013}$	$9.734^{\pm.078}$	-
TEMOS [38]	$0.092^{\pm.003}$	$0.171^{\pm.003}$	$0.238^{\pm.002}$	$29.258^{\pm.0694}$	$6.867^{\pm.013}$	$4.738^{\pm.078}$	$0.672^{\pm.041}$
T2M [12]	$0.184^{\pm.010}$	$0.298^{\pm.006}$	$0.396^{\pm.005}$	$5.481^{\pm.3820}$	$9.576^{\pm.006}$	$5.771^{\pm.151}$	$2.761^{\pm.042}$
T2M* [12]	$0.325^{\pm.004}$	$0.487^{\pm.005}$	$0.593^{\pm.005}$	$4.506^{\pm.020}$	$8.5347^{\pm.055}$	$5.771^{\pm.151}$	$2.761^{\pm.042}$
MDM [45]	$0.203^{\pm.009}$	$0.329^{\pm.007}$	$0.426^{\pm.005}$	$23.701^{\pm.0569}$	$9.548^{\pm.014}$	$5.856^{\pm.077}$	$3.490^{\pm.061}$
MDM(GRU) [45]	$0.179^{\pm.006}$	$0.299^{\pm.005}$	$0.387^{\pm.007}$	$32.617^{\pm.1221}$	$9.557^{\pm.019}$	$7.003^{\pm.134}$	$3.430^{\pm.035}$
ComMDM [42]	$0.090^{\pm.002}$	$0.165^{\pm.004}$	$0.236^{\pm.003}$	$29.266^{\pm.0668}$	$6.870^{\pm.017}$	$4.734^{\pm.067}$	$0.771^{\pm.053}$
InterGen [24]	$0.207^{\pm.004}$	$0.335^{\pm.005}$	$0.429^{\pm.005}$	$5.207^{\pm.2160}$	$9.580^{\pm.018}$	$7.788^{\pm.208}$	$\mathbf{3.686}^{\pm.052}$
InterMamba (Ours)	$\mathbf{0.416}^{\pm.004}$	$\mathbf{0.605}^{\pm.004}$	$\mathbf{0.711}^{\pm.004}$	$\mathbf{1.873}^{\pm.029}$	$\mathbf{3.862}^{\pm.016}$	$\mathbf{8.459}^{\pm.046}$	$2.427^{\pm.044}$
InterMamba (light)	$0.406^{\pm.005}$	$0.594^{\pm.005}$	$0.705^{\pm.005}$	$2.431^{\pm.051}$	$3.982^{\pm.015}$	$8.407^{\pm.073}$	$2.237^{\pm.064}$

For other methods that were originally proposed for single-person motion generation, we follow the strategies described in InterGen and InterX to extend them to human-human interaction scenarios by modifying their network input and output dimensions to accommodate the non-canonical two-person representation introduced in InterHuman [24]. Particularly, for ComMDM [42], we report two sets of results to ensure fairness: (1) the original few-shot setting provided in the paper, which involves pretraining and fine-tuning on 10 samples (denoted with $*$); (2) a re-implemented version trained from scratch on the InterHuman training set using the same data representation (denoted without $*$). In addition, we apply the same setup to T2M [12], using the Inter-X dataset.

Table I and Table II present the quantitative comparison of our InterMamba with various human-human interaction generation methods on the InterHuman [24] and InterX [52] datasets, respectively. Each evaluation is conducted 20 times, except for MModality, which is run 5 times, and we report the averaged results along with a 95% confidence interval. The results demonstrate that InterMamba consistently outperforms previous approaches, achieving the highest R-Precision across all settings (0.475/0.625/0.706 on InterHuman and 0.4573/0.642/0.742 on Inter-X), indicating superior text-motion alignment. Additionally, it attains the lowest MM Dist (3.785 on InterHuman and 3.594 on Inter-X), ensuring improved structural coherence and smoother motion transitions. Furthermore, InterMamba exhibits the highest Diversity scores (7.963 on InterHuman and 9.194 on Inter-X), reflecting its strong capacity to generate a broad range of human interactions. While its FID scores (5.945 and 0.517, respectively) are slightly higher than a few baselines, this trade-off enhances motion quality and semantic consistency. Additionally, InterMamba achieves a competitive MModality score (0.993 on InterHuman and 2.593 on Inter-X), demonstrating its ability to generate diverse and interaction-aware motion patterns. These results further validate that InterMamba effectively balances accuracy, coherence, and diversity, making it as the state-of-the-art solution for text-driven human-human interaction motion generation.

2) Qualitative Comparison: In Fig. 6 and Fig. 7, we qualitatively compare our method with InterGen [24] and in2IN [41], both trained on the InterHuman dataset using identical text prompts. From Fig. 6, for the prompt: *The two persons approach each other and bow down while bending their waist*, InterGen exhibits unnatural limb positioning, rigid motion, and inconsistent spacing, lacking proper arm-torso contact. InterMamba, however, ensures fluid movement, synchronized waist bending, and realistic posture alignment. We attribute this to the design of ASTM, which selects different scanning directions to preserve spatial coherence. For the prompt: *Two people embrace each other*, InterGen struggles with spatial coordination, leading to

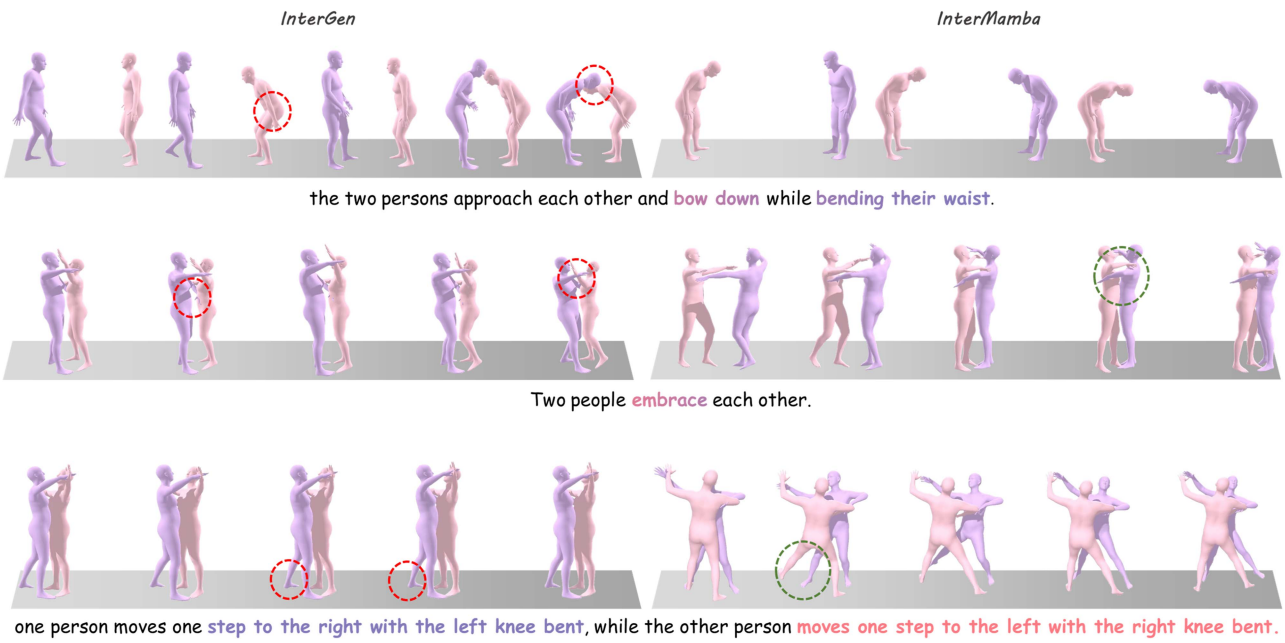


Fig. 6. Qualitative comparison of interaction generation results from **InterGen** and **InterMamba**. **Red circles** indicate implausible or inaccurate motions, while **green circles** highlight more plausible interactions.

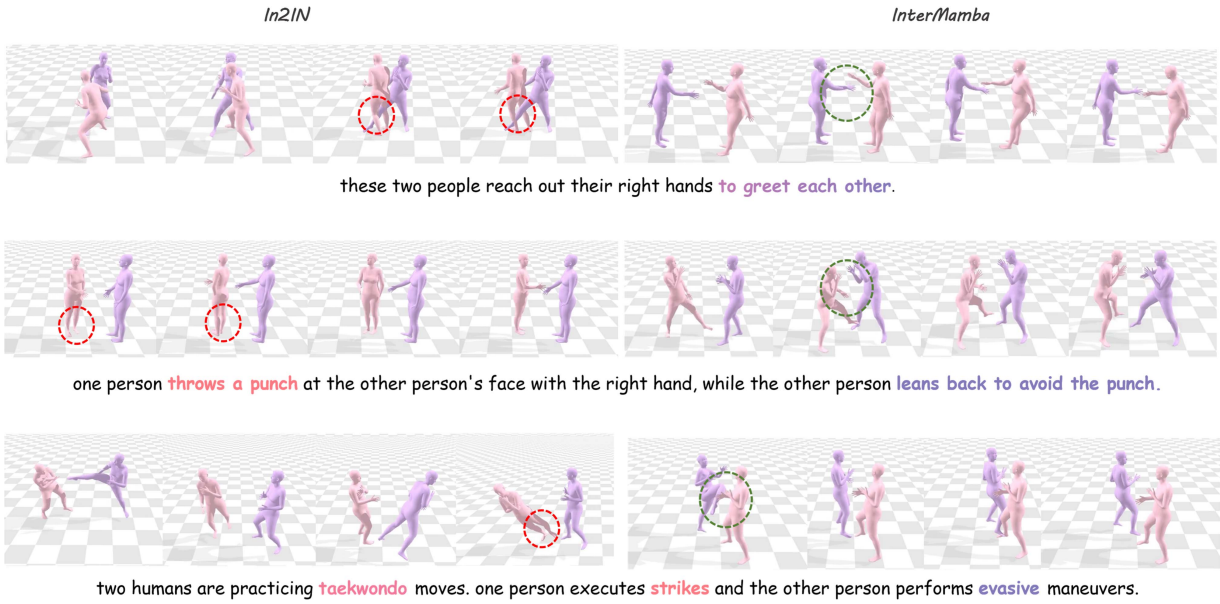


Fig. 7. Qualitative comparison of interaction generation results from **in2IN** and **InterMamba**. **Red circles** indicate implausible or inaccurate motions, while **green circles** highlight more plausible interactions.

misaligned hand positions and an unnatural embrace. In comparison, InterMamba produces a well-executed, natural interaction with realistic body contact. We attribute this to the LIIA module, which effectively captures fine-grained interaction details. For the prompt: *One person moves one step to the right with the left knee bent, while the other moves one step to the left with the right knee bent*, InterGen struggles to maintain correct foot placement and knee bending, due to its weaker ability to capture fine-grained textual details. In contrast, InterMamba, leveraging AdaLN and Mamba's strong contextual information extraction capabilities, accurately reflects the prompt, ensuring synchronized stepping and spatial coherence. Besides, InterMamba exhibits superior interaction naturalness and spatial awareness, producing smooth transitions and physically plausible interperson contact.

Furthermore, from Fig. 7, it is evident that In2IN suffers from severe foot sliding artifacts and unnatural motion patterns, especially in dynamic scenes. These artifacts severely degrade the

TABLE III
THIS TABLE COMPARES THE COMPUTATIONAL COMPLEXITY OF OUR METHOD WITH THAT OF OTHER APPROACHES, BASED ON THREE INDICATORS: AVERAGE INFERENCE TIME, PARAMETERS, AND FLOPS. AS SHOWN IN THE TABLE, OUR INTERMAMBA ACHIEVES THE HIGHEST OVERALL EFFICIENCY ACROSS ALL METRICS.

Methods	Inference time(s)	Params (M)	FLOPs (G)
InterGen [24]	1.233	182	68.33
in2IN [52]	1.260	160	33.72
Ours	0.567	117	33.53
Ours (UltraLight)	**0.325**	**66**	**14.05**

realism and physical plausibility of the generated interactions. In contrast, InterMamba generates more temporally consistent and physically grounded motions, further demonstrating its advantage in fine-grained interaction modeling.

Overall, these qualitative results highlight that our InterMamba significantly surpasses InterGen and In2IN in generating semantically accurate, structurally coherent, and interaction-consistent human motions, establishing it as a more effective solution for text-driven human-human interaction generation.

3) Complexity Analysis: To assess the efficiency of our approach, we conduct a complexity analysis comparing our method against several existing approaches. The results, summarized in Table III, include inference time, model parameters, and FLOPs (floating point operations per second). From the table, InterGen exhibits the highest computational cost, requiring 1.233 s for inference with 182 M parameters and 68.33 G FLOPs, making it computationally expensive. in2IN has a slightly higher inference time (1.260 s) but with 160 M parameters, indicating similar computational demands as InterGen, though FLOPs data is unavailable. In contrast, our InterMamba significantly reduces computational complexity, achieving an inference time of 0.567 s—nearly 2.2× faster than InterGen—while maintaining a more efficient model size (117M parameters, 33.53G FLOPs). Furthermore, InterMamba (UltraLight) optimizes efficiency even further, achieving the lowest inference time (0.325 s) with just 66M parameters and 14.05G FLOPs, making it the most lightweight and computationally efficient solution. Compared to InterGen and in2IN, InterMamba runs more than twice as fast with significantly lower FLOPs, demonstrating its scalability and efficiency for real-time applications.

C. User Study

To evaluate the subjective quality of different generation methods from a human perspective, we conducted a user study. The result is shown in Fig. 8. In our user study, a total of 30 volunteers (15 males and 15 females, aged 19–25) participated in answering the questions, all of whom had prior experience with MAYA or 3ds Max.

Each participant was asked to watch 30 groups of videos, where each group contained outputs generated by the three methods (In2IN, InterGen, and InterMamba). The three results in each group were randomly ordered and anonymized to avoid bias. Participants were instructed to select the result they found most natural or preferable in each case. All selections were recorded and aggregated, and the final statistics were used to compare the overall user preference for the three methods.

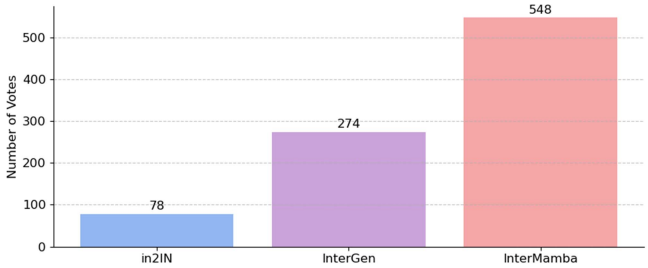


Fig. 8. User Study comparing our InterMamba, InterGen and in2IN.

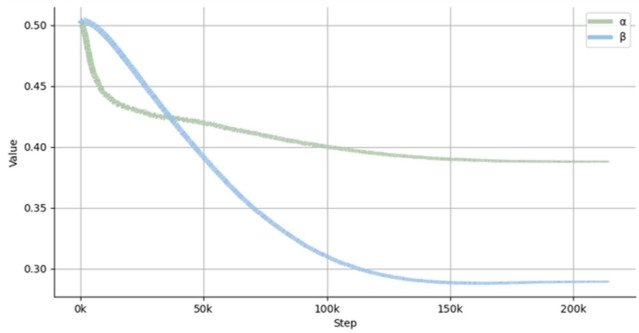


Fig. 9. As the training progresses, the variation process of the adaptive parameters α and β.

As shown in the Fig. 8, InterMamba received the highest number of votes, accounting for 60.9% of all selections. In comparison, InterGen and in2IN received 30.4% and 8.7%, respectively. These findings suggest that InterMamba produces outputs more aligned with human preference and confirm the effectiveness of our method, particularly in terms of naturalness and diversity.

D. Ablation Study

To assess the effectiveness of each module, we conduct comprehensive ablation study to quantify the significance of each module and explore their interactions within the overall framework. First, we present the importance of the adaptive parameters during training in Fig. 9 and Table IV. Next, we examine the effectiveness of the three key components in our InterMamba framework by incrementally adding the proposed modules to the baseline InterGen [24], with the results shown in Table V.

The Effect of Adaptive parameters: To evaluate the effectiveness of the adaptive Spatio-Temporal SSM and different fusion strategies, we conduct a detailed ablation study on the InterHuman dataset. In each ASTM block, we introduce two trainable scalar parameters, w_α and w_β, which are independently learned during training and fixed during inference. These parameters are not shared across modules and are not conditioned on the input, allowing localized and stable spatial-temporal adaptation. As shown in Fig. 9, w_β gradually decreases over training steps, indicating increased temporal emphasis, while w_α remains stable, reflecting its persistent contribution to spatial

TABLE IV
QUANTITATIVE ABLATION STUDY ON THE INTERHUMAN DATASET EVALUATING THE IMPACT OF ADAPTIVE SPATIO-TEMPORAL FUSION AND ALTERNATIVE FUSION STRATEGIES. **BOLD** INDICATES THE BEST RESULT IN EACH CATEGORY, AND UNDERLINE REFERS TO THE SECOND BEST RESULT.

Methods	R Precision↑			*FID* ↓	*MMDist* ↓	*Diversity* →	*MModality* ↑
	Top1	Top2	Top3				
w/o Temporal Mamba	$0.464^{\pm.005}$	$0.603^{\pm.004}$	$0.680^{\pm.006}$	$\mathbf{5.539}^{\pm.132}$	$3.789^{\pm.001}$	$7.840^{\pm.031}$	$0.844^{\pm.022}$
w/o adaptive param.	$0.453^{\pm.005}$	$0.625^{\pm.005}$	$0.703^{\pm.005}$	$6.121^{\pm.141}$	$3.786^{\pm.001}$	$7.791^{\pm.030}$	$0.829^{\pm.024}$
w/ adaptive param.	$\mathbf{0.475}^{\pm.005}$	$\mathbf{0.625}^{\pm.006}$	$\mathbf{0.706}^{\pm.004}$	$5.945^{\pm.007}$	$\mathbf{3.785}^{\pm.003}$	$\mathbf{7.963}^{\pm.033}$	$0.993^{\pm.026}$
linear Fusion	$0.448^{\pm.006}$	$0.600^{\pm.004}$	$0.678^{\pm.004}$	$7.680^{\pm.137}$	$3.797^{\pm.001}$	$8.061^{\pm.033}$	$1.059^{\pm.027}$
Gate Fusion	$0.442^{\pm.006}$	$0.589^{\pm.005}$	$0.664^{\pm.005}$	$10.961^{\pm.145}$	$3.806^{\pm.001}$	$8.208^{\pm.027}$	$\mathbf{1.070}^{\pm.018}$
Plus Fusion	$0.443^{\pm.004}$	$0.589^{\pm.005}$	$0.666^{\pm.005}$	$10.775^{\pm.123}$	$3.808^{\pm.001}$	$8.185^{\pm.035}$	$1.036^{\pm.020}$

TABLE V
THE EFFECTIVENESS OF EACH MODULE IN INTERMAMBA ON THE INTERHUMAN DATASET. THE SELF-ASTM REFERS TO THE SELF-ASTM BLOCK IN SECTION IV-B, CROSS-ASTM DENOTES THE CROSS-ASTM BLOCK IN SECTION IV-D, AND LIIA STANDS FOR THE LOCAL INTERACTION INFORMATION AGGREGATION MODULE IN SECTION IV-C. **BOLD** FONT HIGHLIGHTS THE BEST PERFORMANCE IN EACH CATEGORY, AND UNDERLINE REFERS TO THE SECOND BEST RESULT.

Self-ASTM	Cross-ASTM	LIIA	Params (M)	R Precision (Top 3) ↑	*FID* ↓	*MMDist* ↓	*Diversity* →	*MModality* ↑
			182	$0.624^{\pm.010}$	$\mathbf{5.918}^{\pm.079}$	$5.108^{\pm.012}$	$7.387^{\pm.029}$	$\mathbf{2.141}^{\pm.063}$
✓			180	$0.614^{\pm.012}$	$8.524^{\pm.060}$	$3.814^{\pm0.001}$	$8.032^{\pm0.031}$	$0.949^{\pm0.023}$
✓	✓		117	$0.605^{\pm.002}$	$13.184^{\pm.012}$	$3.841^{\pm0.001}$	$7.960^{\pm0.030}$	$1.045^{\pm0.026}$
✓	✓	✓	**117**	$\mathbf{0.705}^{\pm.004}$	$5.945^{\pm.007}$	$\mathbf{3.785}^{\pm.003}$	$\mathbf{7.963}^{\pm.033}$	$0.993^{\pm.026}$

encoding. Table IV presents quantitative results under various settings. Without temporal Mamba, the model achieves the best FID (5.539) but lower R-Precision (Top1: 0.464), suggesting structurally plausible but temporally less coherent motion. We observe that removing the adaptive parameters—replacing the learned weighted fusion with a simple summation of the temporal and spatial features—leads to training instability and eventual failure of convergence and higher FID (6.121), further confirming that the adaptive parameterization is essential for stable and effective training. When adaptive parameters are included, our model achieves the best overall performance, with the highest R-Precision (0.475/0.625/0.706), the lowest MM Distance (3.785), and the highest Diversity (7.963), demonstrating enhanced text-motion alignment and generation diversity.

We further compare three commonly used fusion strategies: Linear Fusion achieves relatively high Diversity (8.061) but suffers from worse FID (7.680) and R-Precision, implying semantic inconsistency; Gate Fusion obtains the highest MModality (1.070), indicating better interaction variety, but with poor FID (10.961); Plus Fusion shows similar trends.

These results confirm that our adaptive fusion strategy not only stabilizes training but also effectively balances spatial-temporal information, achieving superior precision, coherence, and diversity in human-human interaction motion generation.

Effect of Self-ASTM Block: To demonstrate the superiority of the self-ASTM module, we compare its standalone performance against other configurations. The first row in Table V shows that using only the Self-ASTM results in an R-Precision of **0.371** and an FID of **5.918**, indicating that SRM establishes a solid baseline for individual motion modeling. However, without inter-human interaction understanding, the model's text-motion alignment is suboptimal, highlighting the need for additional components to capture complex interactions.

Effect of Cross-ASTM Block: Adding the cross-ASTM block further improves the model's R-Precision to **0.409**, but significantly increases the FID to **8.524**. This trade-off highlights that while Cross-ASTM enhances inter-character motion coordination and interaction, it introduces challenges in maintaining consistent and coherent motions across the characters. As a result, the quality of motion generation degrades, leading to visual inconsistencies and a reduction in overall motion fluidity.

Effect of Local Interaction Information Aggregation: Further integrating the Local Interaction Information Aggregation (LIIA) alongside Self-ASTM and Cross-ASTM significantly boosts the R-Precision to **0.705**, the highest among all configurations, while maintaining a competitive FID of **5.945**. These results confirm that LIIA plays a crucial role in refining inter-human interactions and enhancing semantic consistency, effectively balancing high motion quality with strong text alignment.

VI. CONCLUTION

In this paper, we introduce InterMamba, an efficient and high-quality framework for text-driven human-human interaction generation. Our approach effectively tackles two major challenges in the field: efficiently capturing individual motion features and interaction details from text while ensuring real-time inference speed. To achieve this, we propose an Adaptive Spatio-Temporal Mamba framework (ASTM) to dynamically extract and integrate spatial and temporal features. Building on this foundation, we develop two core modules: Self-ASTM, which captures long-range dependencies for each character using AST, and Cross-ASTM, which explicitly models human-human interactions with the assistance of Local Interaction Information Aggregation (LIIA). These modules enable InterMamba to generate realistic human-human interaction efficiently. Our extensive experiments demonstrate that InterMamba sets a new

benchmark for real-time, high-quality human-human interaction generation, laying the groundwork for future advancements in text-driven Human-Human interaction generation.

VII. Limitations

Despite InterMamba's advancements, several limitations remain. First, although inference speed is improved by integrating Mamba, the method still relies on diffusion-based generation, which involves multiple iterations. Reducing these steps speeds up inference but may affect motion quality. Future work could explore alternative models—like flow-based or hybrid approaches—to improve both speed and quality. Second, while Mamba helps with long-range dependencies, the generated motions may still lack detail in subtle interactions, emotions, or physical realism. Enhancing contact modeling and biomechanical constraints is vital for more expressive results. Finally, practical deployment needs improvement. Real-world applications like avatars and games require better user control, real-time adaptation, and multi-task optimization. Future work should focus on adaptive strategies like user feedback and multi-agent collaboration to bridge the gap between algorithms and real-world use.

References

[1] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 8, pp. 3519–3534, Aug. 2023.

[2] S. Azadi, A. Shah, T. Hayes, D. Parikh, and S. Gupta, "Make-an-animation: Large-scale text-conditional 3D human motion generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 14993–15002.

[3] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 387–404.

[4] X. Chen et al., "Executing your commands via motion diffusion in latent space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18000–18010.

[5] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–14.

[6] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[7] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry Hippos: Towards language modeling with state space models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–27.

[8] K. Gong et al., "TM2D: Bimodality driven 3D dance generation via music-text integration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9908–9918.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. 1st Conf. Lang. Model.*, 2024, pp. 1–32.

[10] A. Gu et al., "Combining recurrent, convolutional, and continuous-time models with linear state-space layers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 572–585.

[11] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "MOMask: Generative masked modeling of 3D human motions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1900–1910.

[12] C. Guo et al., "Generating diverse and natural 3D human motions from text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5152–5161.

[13] C. Guo et al., "Action2Motion: Conditioned generation of 3D human motions," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2021–2029.

[14] R. Hasani, M. Lechner, T.-H. Wang, M. Chahine, A. Amini, and D. Rus, "Liquid structural state-space models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–22.

[15] A. Hernandez, J. Gall, and F. Moreno-Noguer, "Human motion prediction via spatio-temporal inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7134–7143.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

[17] J. Ho and T. Salimans, "Classifier-free diffusion guidance," pp. 1–8, 2022, *arXiv:2207.12598*.

[18] M. G. Javed, C. Guo, L. Cheng, and X. Li, "InterMask: 3D human interaction generation via collaborative masked modeling," in *Proc. 13th Int. Conf. Learn. Representations*, 2025, pp. 1–22.

[19] N. Jiang et al., "Scaling up dynamic human-scene interaction modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1737–1747.

[20] J. Kim, J. Kim, and S. Choi, "Flame: Free-form language-based motion synthesis & editing," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 7, pp. 8255–8263.

[21] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu, "Controllable human-object interaction synthesis," in *Proc. Eur. Conf. Comput. Vis.*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., 2024, pp. 54–72.

[22] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI Choreographer: Music conditioned 3D dance generation with AIST," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13401–13412.

[23] H. Liang et al., "OMG: Towards open-vocabulary motion generation via mixture of controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 482–493.

[24] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "InterGEN: Diffusion-based multi-human motion generation under complex interactions," *Int. J. Comput. Vis.*, vol. 132, pp. 3463–3483, 2024.

[25] B. Lin et al., "MTMamba: Enhancing multi-task dense scene understanding by Mamba-based decoders," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 314–330.

[26] H. Liu, M. Zaharia, and P. Abbeel, "Ring attention with blockwise transformers for near-infinite context," in *Proc. 12th Int. Conf. Learn. Representations*, 2024, pp. 1–23.

[27] J. Liu et al., "Point Mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy," pp. 1–17, 2024, *arXiv:2403.06467*.

[28] Y. Liu et al., "VMamba: Visual state space model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 1–21.

[29] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[30] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2891–2900.

[31] S. Miao, Z. Lu, M. Liu, J. Duarte, and P. Li, "Locality-sensitive hashing-based efficient point transformer with applications in high-energy physics," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 1–18.

[32] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[33] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-Transformer: Hierarchical vision transformer with local self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2082–2091.

[34] J. Park, H. Kim, K. Ko, M. Kim, and C. Kim, "VideoMamba: Spatio-temporal selective state space model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 1–18.

[35] X. Peng, S. Mao, and Z. Wu, "Trajectory-aware body interaction transformer for multi-person pose forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 17121–17130, doi: 10.1109/CVPR52729.2023.01642.

[36] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang, "HOI-DIFF: Text-driven synthesis of 3D human-object interactions using diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2025, pp. 2878–2888.

[37] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3D human motion synthesis with transformer VAE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10985–10995.

[38] M. Petrovich, M. J. Black, and G. Varol, "TEMOS: Generating diverse human motions from textual descriptions," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 480–497.

[39] M. Pietruszka, Ł. Borchmann, and Ł. Garncarek, "Sparsifying transformer models with trainable representation pooling," in *Proc. 60th*

Annu. Meeting Assoc. Comput. Linguistics, Dublin, Ireland, May 2022, pp. 8616–8633.
[40] E. Pinyoanuntapong, P. Wang, M. Lee, and C. Chen, "MMM: Generative masked motion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1546–1555.
[41] P. Ruiz-Ponce, G. Barquero, C. Palmero, S. Escalera, and J. Garcıá-Rodrıguez, "in2in: Leveraging individual information to generate human interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2024, pp. 1941–1951.
[42] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," pp. 1–13, 2023, *arXiv:2303.01418*.
[43] M. Tanaka and K. Fujiwara, "Role-aware interaction generation from textual description," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15953–15963.
[44] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "MotionClip: Exposing human motion generation to clip space," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 358–374.
[45] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," in *Proc. 11th Int. Conf. Learn. Representations*, 2023, pp. 1–16.
[46] J. Tseng, R. Castellon, and K. Liu, "EDGE: Editable dance generation from music," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 448–458, 2023.
[47] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6306–6315.
[48] J. Wang, S. Yan, B. Dai, and D. Lin, "Scene-aware generative network for human motion synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12206–12215.
[49] Z. Wang et al., "Move as you say, interact as you can: Language-guided human motion generation with scene affordance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 433–444.
[50] Z. Wu, S. Mao, C. Zhang, Y. Wang, and M. Zeng, "Contrastive disentanglement for self-supervised motion style transfer," *Multimedia Tools Appl.*, vol. 83, no. 27, pp. 70523–70544, 2024.
[51] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "OmniControl: Control any joint at any time for human motion generation," in *Proc. 12th Int. Conf. Learn. Representations*, 2024, pp. 1–19.
[52] L. Xu et al., "Inter-X: Towards versatile human-human interaction analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22260–22271.
[53] L. Xu et al., "ActFormer: A GAN-based transformer towards general action-conditioned 3D human motion generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2228–2238.
[54] M. Zhang et al., "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, 2024.
[55] M. Zhang et al., "RemoDiffuse: Retrieval-augmented motion diffusion model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 364–373.
[56] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, and Z. Liu, "FineMoGen: Fine-grained spatio-temporal motion generation and editing," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 13981–13992.
[57] Z. Zhang et al., "KMM: Key frame mask Mamba for extended motion generation," pp. 1–15, 2024, *arXiv:2411.06481*.
[58] Z. Zhang et al., "InfiniMotion: Mamba boosts memory in transformer for arbitrary long motion generation," pp. 1–15, 2024, *arXiv:2407.10061*.
[59] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, and H. Tang, "Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective SSM," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 265–282.
[60] Z. Zhang et al., "Motion anything: Any to motion generation," pp. 1–18, 2025, *arXiv:2503.06955*.
[61] Z. Zhang et al., "Motion Avatar: Generate human and animal avatars with arbitrary motion," in *Proc. British Mach. Vis. Conf.*, 2024, pp. 1–17.
[62] C. Zhong, L. Hu, Z. Zhang, and S. Xia, "ATTT2M: Text-driven human motion generation with multi-perspective attention mechanism," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 509–519.
[63] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 1–18.

Zizhao Wu received the Ph.D. degree from the Department of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2013. He is currently an associate professor with the Faculty of Digital Media Technology, Hangzhou Dianzi University, Hangzhou. His research interests mainly include computer vision and computer graphics.

Yingying Sun is currently working toward the MS degree with Hangzhou Dianzi University, Hangzhou, China. Her research interests include computer graphics and computer vision.

Yiming Chen is currently working toward the Undergraduate degree Hangzhou Dianzi University, Hangzhou, China. Her research interests include computer graphics and computer vision.

Xiaoling Gu received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2017. She is currently an associate professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou. Her research interests include the areas of multimedia analysis, visual-language learning, and image generative models.

Ruyu Liu received the PhD degree in control science and engineering from the Zhejiang University of Technology, Hangzhou, China, in 2021. She is currently a Marie Sktodowska-Curie Postdoctoral Fellow. She has authored or coauthored more than 60 publications, including top-tier IEEE transactions (*IEEE Transactions on Industrial Electronics*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Transactions on Industrial Informatics*, and *IEEE Transactions on Artificial Intelligence*). Her research interests include intelligent robotics and embodied AI perception. She has secured competitive grants, including the EU Horizon Program, NSFC Youth Program, and China Postdoctoral Science Foundation. She is also a Young Editorial Board member for the Metaverse.

Jiazhou Chen received the joint PhD degree in computer science from the Institut National de Recherche en Informatique et en Automatique (INRIA) Bordeaux Sud-Ouest, Talence, France, and also from the State Key Laboratory of Computer Aided Design (CAD) and Computer Graphics (CG), Zhejiang University, Hangzhou, China, in 2012.

He is currently an associate professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. He is also the deputy director with the Institute of Computer Software. His research interests include computer graphics, digital humanities, and visualization.