HiTMM: Generative Temporal Masked Modeling of Human Interactive Motions

Zicheng Jiao, Yunlian Sun, Hongwen Zhang, Jinhui Tang, Massimo Tistarelli

***Abstract*—We have recently seen some progress in the current field of human-human interaction generation. However, directly generating complex two-person interactive motions remains a significant challenge. Meanwhile, these models typically employ two independent timelines when generating motions for interactive scenarios involving two individuals. This design overlooks the temporal dependencies between motions at each timestep and fails to account for the roles of active and reactive participants during the generation process, often resulting in unrealistic and unnatural motions. In this work, we propose HiTMM, a novel framework for Human interaction generation based on Temporal Masked Modeling. HiTMM first decomposes the human interaction into two separate single-person motions. Individual motions within the interaction belong to the same type, enabling them to be mapped to a shared latent space through a coarse-to-fine approach that produces multi-layer discrete tokens. We then arrange all tokens of the two interacting individuals along a shared timeline. Subsequently, we employ a masked transformer and a residual transformer to model the base-layer and rest-layer motion tokens. Both the base-layer and rest-layer motion tokens are arranged along a single timeline, allowing the model to explicitly capture the temporal order and initiating role embedded in the sequence, where the first individual's motion initiates the interaction. Note that, our model utilizes a shared temporal representation, making it capable of performing temporal editing on specific regions within human interaction sequences. Experimental results show that our model achieves an FID of 5.017 on the InterHuman dataset, surpassing the current state-of-the-art model (vs 5.154 for InterMask), and an FID of 0.373 on the InterX dataset (vs 0.399 for InterMask). Project URL: https://jiaozicheng.github.io/HiTMM/.**

***Index Terms*—Human interaction generation, temporal modeling, masked modeling, discrete motion tokens.**

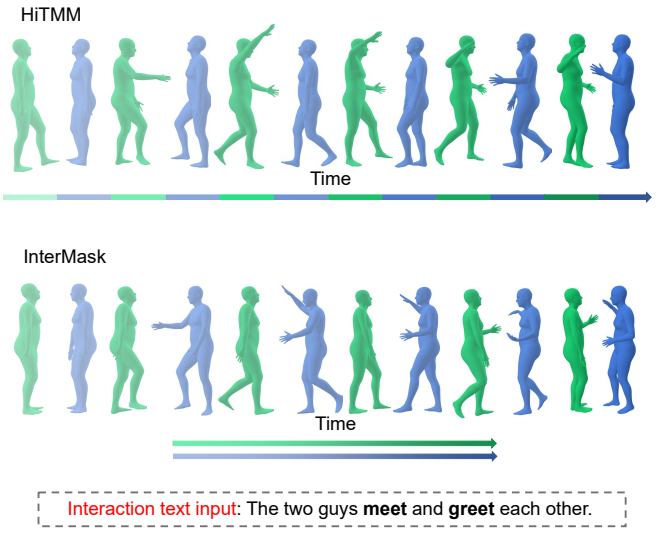

Fig. 1. HiTMM generates human-human interactive motions using a shared timeline under causal temporal control. Given an interaction text input, our method can produce high-quality 3D human motions maintaining temporal continuity.

I. Introduction

HUMAN motion generation plays a crucial role in computer vision and graphics, with extensive applications in animation, robotics, and gaming. Among various generative AI tasks, text-to-motion generation has emerged as one of the most promising directions. Recent advances in motion capture technology [1]–[4] have facilitated the efficient collection of large-scale human motion datasets. Despite significant progress in this field, simultaneously generating reasonable motions for two interactive individuals is highly challenging. First, the motions of both individuals should remain closely correlated at each timestep. Second, it is crucial to preserve the distinct roles of the active and passive participants throughout the generation process. Finally, for downstream applications, the model should generate plausible and natural motions while maintaining consistency throughout the entire sequence.

Currently, most interactive motion generation methods primarily rely on diffusion models. Among these, InterGen [5] stands out as a seminal work in the field. It proposes a specialized diffusion model for human-human interactions, incorporating interaction-centric optimization during training, and yields impressive results. Moreover, InterGen [5] introduces InterHuman, the first comprehensive dataset for human-human interaction generation. Then, in2IN [6] extends the overall interaction description by additionally incorporating individual text descriptions generated through the large language model for further enhancement. Meanwhile, MoMat-MoGen [7] leverages auxiliary motion datasets to retrieve and refine text-conditioned interactions, achieving superior motion quality. Given the relatively limited scale of the InterHuman dataset,

This work was supported in part by the National Natural Science Foundation of China under Grant 62476131 and Grant 62377004, in part by the Major Science and Technology Projects in Jiangsu Province under Grant BG2024042, and in part by the Italian PNRR projects M4C2 Spoke 06 e.INS, HPC Spoke 09 DI-DTPlatform and UrbanDT4TF, FAIR Spoke 10 VisAViT. (Corresponding author: Yunlian Sun.)

Zicheng Jiao, and Yunlian Sun are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jzc6915@njust.edu.cn; yunlian.sun@njust.edu.cn).

Hongwen Zhang is with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China (e-mail: zhanghongwen@bnu.edu.cn).

Jinhui Tang is with the School of Artificial Intelligence, Nanjing Forestry University, Nanjing 210023, China (e-mail: tangjh@njfu.edu.cn.).

Massimo Tistarelli is with the Department of Engineering, University of Sassari, 07100 Sassari, Italy (e-mail: tista@uniss.it).

direct training on it often leads to suboptimal performance. This learning strategy imposes a significant burden on the model, requiring it to generate complete motion information in one pass while simultaneously modeling complex interactions according to textual descriptions. Additionally, the training of diffusion models is notably time-consuming. An alternative and more efficient method involves transforming motions into discrete tokens through vector quantization [8]–[10]. This approach operates by pretraining a Vector Quantized Variational Autoencoder (VQ-VAE) to map motions into a discrete code space, followed by predicting these tokens guided by textual descriptions. InterMask [11] pioneers the application of VQ-VAE for local joint discretization combined with masked modeling, significantly improving the quality of interactive motion generation. However, this method expands the original token sequence length fivefold, resulting in a substantial increase in computational complexity. Furthermore, existing approaches typically model multi-person motion by generating motions for each individual separately following independent timelines. This design risks losing critical interaction information, including fine-grained temporal dependencies between individuals, and also fails to preserve the distinct roles of active and passive participants during interactive motion generation.

Motivated by the outlined issues, we propose **HiTMM**, a novel framework for **H**uman **i**nteraction generation based on **T**emporal **M**asked **M**odeling, as illustrated in Fig. 1. HiTMM is designed with the following key designs: (1) To effectively capture complex human interactive motions, we employ a shared single-person Residual Vector Quantized Variational Autoencoder (RVQ-VAE) to model complete motions of both individuals within the interaction. This strategy treats each individual as the same category, thereby simplifying the representation of human interaction. Compared to traditional VQ-VAE, RVQ-VAE exhibits superior data fitting capabilities, thanks to its multi-layer residual quantization mechanism [9], [12]–[15]. Furthermore, each discrete token represents a motion snippet of the whole body, facilitating applications like temporal editing. (2) In human-human interaction generation, we cross-arrange and serialize motion tokens of both individuals at each timestep. The motion tokens of one individual are temporally linked to those of the other person, enhancing the ability to learn from each other while preserving motion smoothness and continuity. Moreover, by incorporating learnable identity embeddings, this cross-arrangement design enables the model to automatically capture the respective roles of each participant during interaction generation, treating the motion arranged first as the active agent. (3) We utilize the Temporal Masked Transformer (TMT) and the Temporal Residual Transformer (TRT) to learn the entire multi-layer token sequences. TMT and TRT are trained to predict the base-layer and residual motion tokens, respectively. The core design of TMT integrates two specialized attention modules: the Temporal Attention module and the Interactive Attention module. The Temporal Attention module is specifically designed to capture temporal dependencies within the motion token sequences of two individuals, effectively modeling dynamic changes along the time dimension. The Interactive Attention module uses shared self-attention and cross-attention mechanisms to capture individual motion features and model the interactive dependencies between the two individuals.

During the training phase, we initially cross-merge the motion token sequences of two individuals along a single timeline. Subsequently, we employ TMT to mask motion tokens of the base layer for both individuals at a set ratio and perform prediction. Meanwhile, we utilize TRT to progressively predict the motion tokens starting after the base layer until reaching the randomly selected quantized layer. For inference, the process begins with a fully masked set of tokens and proceeds through multiple iterative steps. At each step, the crossover merged tokens with the highest confidence are retained, while the rest are remasked until the iteration completes. Once the base-layer motion tokens are fully predicted, the role occupying the first temporal index position initiates the entire motion sequence as the active participant. Subsequently, TRT progressively predicts the remaining motion token sequences. While each discrete token in InterMask [11] represents a motion snippet of local joints, our approach encodes a motion snippet of the whole body in each discrete token, providing inherent advantages for temporal editing. Thanks to this single-timeline design and discrete tokens of complete motion representations, we can freely edit any continuous region within an interactive motion sequence, similar to the approach used in single-person motion generation. As shown in Table I, HiTMM achieves a superior performance over the current state-of-the-art method InterMask [11] across evaluation metrics including Frechet Inception Distance (FID), R-precision and Average Inference Time per Sentence (AITS) in seconds. Notably, our first-stage motion tokenizer requires only 1 hour and 40 minutes to train, a marked reduction from the 10 hours and 10 minutes needed by spatio-temporal reconstruction stage of InterMask [11]. This efficiency stems directly from our simplified global quantization design.

Overall, the main contributions of this paper are:

- We propose a three-stage pipeline that decouples interactive motion modeling, thereby simplifying the challenging task of directly generating two-person interactions.
- We propose HiTMM, a novel generative temporal masked model that generates interactive motions through a single shared timeline, effectively modeling the temporal dynamics of human-human interactions. By constructing token sequences along this unified timeline, HiTMM captures and incorporates distinct roles of active and passive participants during motion generation.
- Our HiTMM achieves text-conditioned human interaction generation with high fidelity. Experimentally, it sets a new state of the art in this task, achieving an FID of 5.017 (vs 5.154 in InterMask [11]) on InterHuman and 0.373 (vs 0.399 in InterMask [11]) on InterX [5].
- We introduce the first temporal editing method for human-human interaction sequences, effectively reconstructing missing or occluded motion segments.

II. Related Work

A. Human Motion Generation

In recent years, the field of human motion generation has significantly advanced, enabling the creation of complete and

TABLE I
A COMPARATIVE ANALYSIS IS CONDUCTED BETWEEN OUR MODEL AND THE STATE-OF-THE-ART MODEL, INTERMASK [11], ON THE TEXT-CONDITIONED HUMAN INTERACTION DATASET, INTERHUMAN [5].

Methods	FID ↓	R-Precision Top 3 ↑	AITS ↓	Training Time of Motion Tokenizer ↓	What Each Discrete Token Denotes?	Capturing Active / Passive Roles	Exploring Causal Temporal Dependency
InterMask	5.154	0.683	0.81	10h10m	A Motion Snippet of Local Joints	✗	✗
HiTMM	**5.017**	**0.697**	**0.70**	**1h40m**	A Motion Snippet of the Whole Body	✓	✓

detailed motions under various conditions, such as text [8], [9], [16]–[25], audio [26]–[30], music [31]–[33], or motion prefixes [34], [35]. Early works [36], [37] rely mainly on random models, which often produce less realistic and blurry motions. Motion generation based on action labels [38], [39] is inherently class-based and fails to effectively integrate text to produce satisfactory motions. The combination of VAE and transformer frameworks [39], [40] has shown promise, with [41] specifically incorporating text conditions on past frames within a VAE architecture to generate more reasonable motions. The model in [42], trained in a multimodal manner, further strengthens the connection between text and motion. Most recently, diffusion model [43] has emerged as a powerful approach, driving rapid progress in text-driven motion generation. MDM [21] employed a non-categorical diffusion model for text-guided motion generation, while [20] involved adding and denoising noise to the latent space tensors corresponding to motions to generate plausible motions. [44] introduced a sketch-guided motion diffusion framework with a dual-branch time-aware transformer and sketch-aware local attention for intuitive free-hand sketch-to-motion generation. Sport [45] used prompt-conditioned diffusion with body-part contrastive learning to produce dynamically changing motions. Meanwhile, motion quantization models for motion generation are also gaining popularity. [8] quantized motion into discrete tokens and generated realistic motions by regressively predicting these tokens conditioned on textual input. [9], [10] adopted a generative masked modeling framework to enhance text-to-motion generation, producing more refined and plausible motions while offering novel approaches for motion editing. These works provide inspiration for the construction of our model.

B. Human-Human Interaction Generation

Interactive motion generation has witnessed rapid progress in recent years. ComMDM [46] trained a small network to link two pre-trained single-person generation models, MDM [21], enabling the generation of human-human interactive motions. InterGen [5] introduced InterHuman, the first extensive text-based dataset for human interactive motions. Additionally, it designed a diffusion-based model for generating interactive motions, significantly advancing subsequent research in this field. The following studies [6], [7] further improved interactive motion generation by incorporating single-person textual descriptions and motion datasets. Meanwhile, InterX [5] provided the largest existing dataset of human-human interactions, featuring high-quality motion capture, diverse interaction types, and detailed hand gestures. Reactive generation [47]–[49] quickly gained significant interest by introducing the concepts of initiator and reactor into the field of motion generation. However, current diffusion models for human-human interaction generation often face challenges with slow performance in practical applications. Although InterMask [11] employed vector quantization to enhance both the quality and efficiency of interactive motion generation, it overlooked critical temporal dependencies in two interacting individuals at each timestep.

C. Quantized Motion Representation

Transforming data into discrete forms has seen rapid development in recent years. Deep Motion Signatures [50] successfully learned discrete motif words using a contrastive learning approach. TM2T [51] was among the earliest to introduce VQ-VAE [52], converting motion data into discrete tokens and representing motion as a new "language", thus establishing an effective connection between motion and discrete tokens. Building on this, [8], [53], [54] further refined motion quantization using Exponential Moving Average (EMA) and code reset techniques. [9], [10] significantly improved the quality of motion generation by generating discrete tokens through masked prediction. Specifically, MoMask [9] enhanced motion representation by constructing RVQ-VAE to reduce quantization errors, while MMM [10] employed a larger codebook to express single-person motions more effectively. InterMask [11] pioneered the introduction of the VQ-VAE method in the field of interactive motion generation. It quantized human-human motion data into 2D discrete tokens, achieving a quantized representation of joint features in interactive motions. In contrast to InterMask [11], this study innovatively adopts a coarse-to-fine approach to represent complete motion segments within each discrete token, achieving the first comprehensive quantization of motion features in human-human interactions.

III. APPROACH

Our objective is to generate human interactions $\{\mathbf{m}_p\}_{p\in\{a,b\}}$ under the guidance of a textual description c, where $\mathbf{m}_p \in \mathbb{R}^{N\times D}$ represents the motion sequence of an individual, N denotes the time length, D represents the size of motion features, a and b correspond to Person A and Person B in the pair, respectively. As illustrated in Fig. 2, our method comprises three stages. First, we use RVQ-VAE to learn a codebook that maps the motions of each person into multi-layer quantized tokens (Section III-A). Next, we

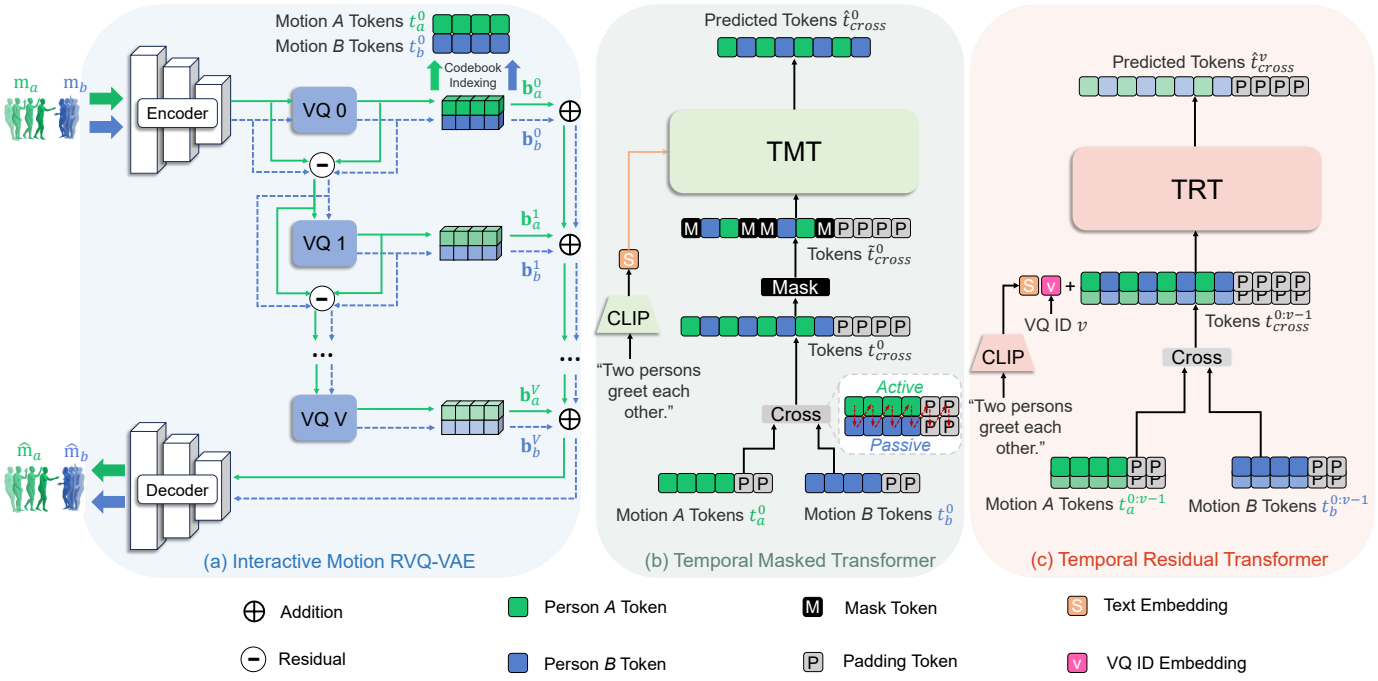

Fig. 2. **Overview of HiTMM.** (a) Here, the motions of each individual are quantized through residual vector quantization (RVQ), $t_a^{0:V}$ and $t_b^{0:V}$. (b) The motion tokens in the base layer of two individuals are first cross-combined to obtain t_{cross}^0, which is then masked and processed through TMT for prediction. (c) The remaining tokens, $t_{cross}^{v>0}$ are progressively predicted layer by layer using TRT based on the tokens $t_{cross}^{0:v-1}$ from original layers.

train a new Temporal Masked Transformer to model the base-layer motion tokens for both individuals (Section III-B). Simultaneously, we develop a Temporal Residual Transformer that progressively generates the complete set of motion tokens from the base-layer motion tokens (Section III-C). The inference process is described in Section III-D.

A. Interactive Motion Tokenizer

The goal of the first stage is to map interactive motions into motion tokens through vector quantization. As shown in Fig. 2(a), we treat each individual's motions as the same data type, allowing utilizing the same single-person RVQ-VAE to model two-person interactive motions. Specifically, we first use an encoder to map the motion sequence $\mathbf{m}_{p(1:N)} \in \mathbb{R}^{N\times D}$ into a latent vector sequence $\tilde{\mathbf{b}}_{p(1:n)} \in \mathbb{R}^{n\times d}$ with a down-sampling ratio of n/N and latent dimension d. In the first layer, similar to VQ-VAE, each of the d-dimensional vector is replaced with the closest codebook entry in the codebook $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{|\mathcal{C}|} \subset \mathbb{R}^d$ through quantization $\mathrm{Q}(\cdot)$, producing the quantized sequence $\mathbf{b}_{p(1:n)} = \mathrm{Q}(\tilde{\mathbf{b}}_{p(1:n)}) \in \mathbb{R}^{n\times d}$. Here, we represent the residual values of layer 0 as: $\mathbf{r}_p^0 = \tilde{\mathbf{b}}_\mathbf{p}$. After the base layer, Residual Quantization (RQ) recursively calculates $\mathbf{b}_p^v$ as the approximation of residual $\mathbf{r}_p^v$, and subsequently computes the next residual $\mathbf{r}_p^{v+1}$ as:

$$\mathbf{b}_p^v = \mathrm{Q}(\mathbf{r}_p^v), \quad \mathbf{r}_p^{v+1} = \mathbf{r}_p^v - \mathbf{b}_p^v, \tag{1}$$

for $v = 0, ..., V$. In general, RVQ-VAE uses a total of $V+1$ layers to represent the latent motion sequence as $V+1$ ordered quantized sequence $\left[\mathbf{b}_{p(1:n)}^v\right]_{v=0}^V = \mathrm{RQ}(\tilde{\mathbf{b}}_{p(1:n)})$, with $\mathbf{b}_{p(1:n)}^v \in \mathbb{R}^{n\times d}$ representing the code sequence of the v-th quantization layer. After residual quantization, the latent sequence of the two-person motions $\mathbf{b_p}$ is the sum of all quantized sequences $\sum_{v=0}^V \mathbf{b}_p^v$, which will then be decoded for motion reconstruction by the decoder. The $V+1$ quantized motion sequences $\left[\mathbf{b}_p^v\right]_{v=0}^V$ can be represented using their corresponding indexes in the codebook $\left[t_{p(1:n)}^v\right]_{v=0}^V$, where $t_{p(1:n)}^v \in \{1, \cdots, |\mathcal{C}|\}^n$.

Following [9], we use motion reconstruction loss and commitment loss to train RVQ-VAE for each quantization layer:

$$\mathcal{L}_{rvq} = \|\mathbf{m}_p - \hat{\mathbf{m}}_p\|_1 + \beta \sum_{v=0}^{V} \|\mathbf{r}_p^v - \mathrm{sg}[\mathbf{b}_p^v]\|_2^2, \tag{2}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation, β serves as a weighting factor used for embedding constraint. $\hat{\mathbf{m}}_a$ and $\hat{\mathbf{m}}_b$ denote the reconstructed motions of the individuals A and B, respectively. For updating the codebook, we apply EMA [52] and codebook reset techniques [8] to ensure optimal performance.

In addition to the basic losses of RVQ-VAE, we incorporate geometric losses from [5] to introduce additional constraints. Specifically, the foot contact loss $\mathcal{L}_{fc}$ encourages the velocity of the feet to be zero when touching the ground, while the velocity loss $\mathcal{L}_{vel}$ encourages the reconstructed motion joints to match the joint velocities of the groundtruth motion sequences. The bone length loss $\mathcal{L}_{bl}$ encourages adjacent joints in the reconstructed motion to match those in the groundtruth motion.

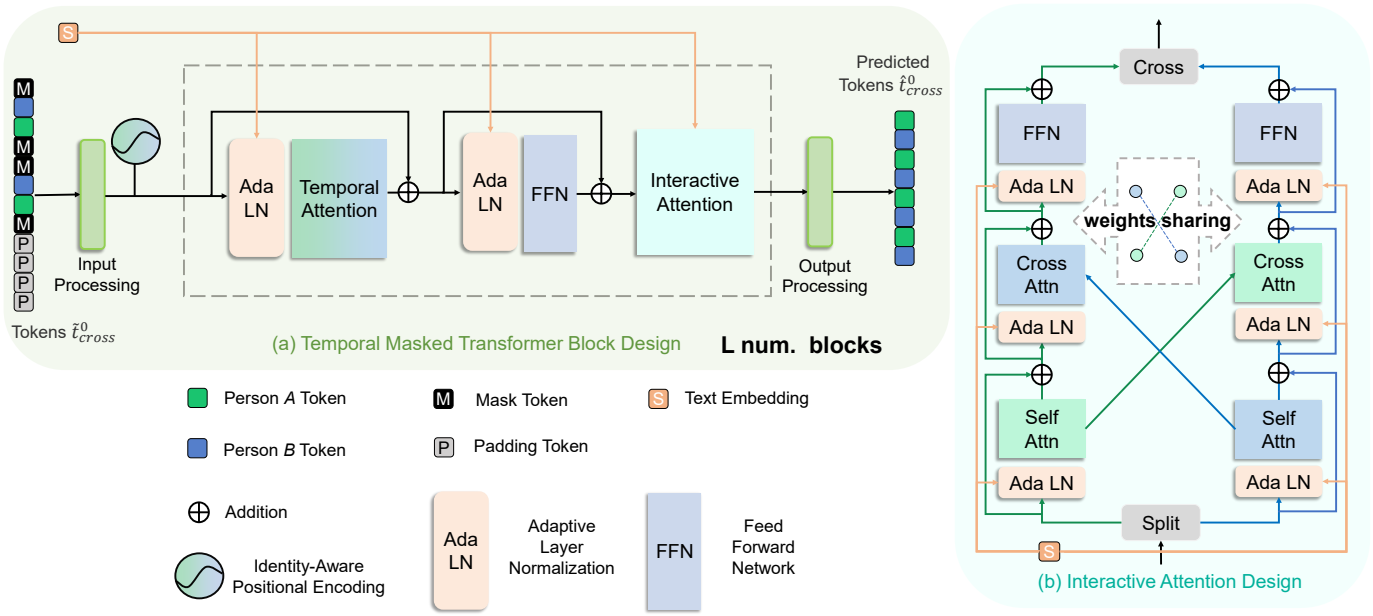

Fig. 3. **TMT.** (a) Each block in TMT employs Temporal and Interactive Attention modules to capture respectively temporal dependencies and interactions in the motion sequences. (b) The Interactive Attention module is designed to learn the interactions between the two persons.

$$\mathcal{L}_{vel} = \frac{1}{N-1}\sum_{i=1}^{N} \|(m_{i+1} - m_i) - (\hat{m}_{i+1} - \hat{m}_i)\|_1,$$
$$\mathcal{L}_{fc} = \frac{1}{N-1}\sum_{i=1}^{N} \|(\hat{m}_{i+1} - \hat{m}_i) \cdot f_i\|_1, \quad (3)$$
$$\mathcal{L}_{bl} = \frac{1}{N-1}\sum_{i=1}^{N} \|B(m_i) - B(\hat{m}_i)\|_1.$$

Here, m_i and $\hat{m}_i$ denote the ground-truth and reconstructed pose at time step i, respectively, within a sequence of total length N. The binary variable $f_i \in \{0, 1\}$ indicates foot contact states for the heel and toe joints, while the function $B(\cdot)$ computes the lengths of bones between adjacent joints.

The overall loss $\mathcal{L}_{rvqvae}$ is the sum of $\mathcal{L}_{rvq}$ and the weighted combination of geometric losses and interactive losses:

$$\mathcal{L}_{rvqvae} = \mathcal{L}_{rvq} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{fc}\mathcal{L}_{fc} + \lambda_{bl}\mathcal{L}_{bl}, \quad (4)$$

where the hyper-parameters λ_{vel}, λ_{fc}, λ_{bl} are carefully calibrated to balance the relative magnitudes of their respective loss terms.

B. Conditional Temporal Masked Transformer

Our temporal masked transformer is designed to jointly construct the motion tokens $\{t^0_a, t^0_b\}$ of the base layer under the guidance of text c and capture the complex temporal relationships between them. Note that for human-human interactions, each motion movement is closely tied to all previous movements of both individuals. If the motions of two persons are arranged along a single timeline similar to that used for single-person motions, the resulting sequence will exhibit a causal relationship. Then, we can apply a special [MASK] token to randomly mask and replace tokens in this sorted token sequence. We use $\tilde{t}^0_{cross}$ to represent the masked sequence. Our goal is to predict the masked tokens and generate $\hat{t}^0_{cross}$ using the textual input c, together with $\tilde{t}^0_{cross}$.

1) Causal Crossover Merge: As shown in Fig. 2(b), we first separate the discrete motion tokens of two individuals and then perform a crossover merge to obtain the crossover motion tokens $t^0_{cross} = \{t^0_{p(\lceil i/2 \rceil)}\}_{i=1}^{2n}$. Here, n denotes the length of the discrete token sequence for each single-person motion, $\lceil \cdot \rceil$ is the ceiling function (rounding up), / represents division operator, and p can be obtained as follows:

$$p = \begin{cases} a, & i \,\%\, 2 = 1 \\ b, & i \,\%\, 2 = 0, \end{cases} \quad (5)$$

where $\%$ denotes the modulo operation.

2) Masking and Remasking Strategies: [MASK] and [PAD] are learnable special-purpose tokens. The [MASK] token is used to denote corrupted input, and the model is trained to predict the original tokens in place of [MASK]. The [PAD] token is employed to pad shorter motion sequences, facilitating batch processing for sequences of varying lengths. The token sequence t^0_{cross} is masked by replacing it with [MASK] following two stages. Firstly, our approach randomly masks the entire motion token sequence. The masking ratio is scheduled with a ratio $\gamma(\tau_i)$, which is controlled by a pre-scheduled function:

$$\gamma(\tau_i) = \cos\left(\frac{\pi\tau}{2}\right) \in [0, 1]$$
$$\tau_i \sim \mathcal{U}(0, 1). \quad (6)$$

Then, we employ a remasking strategy, as used in [55]. If a token is selected for masking, there is a 80% chance that this token will be replaced with [MASK], a 10% chance that it will

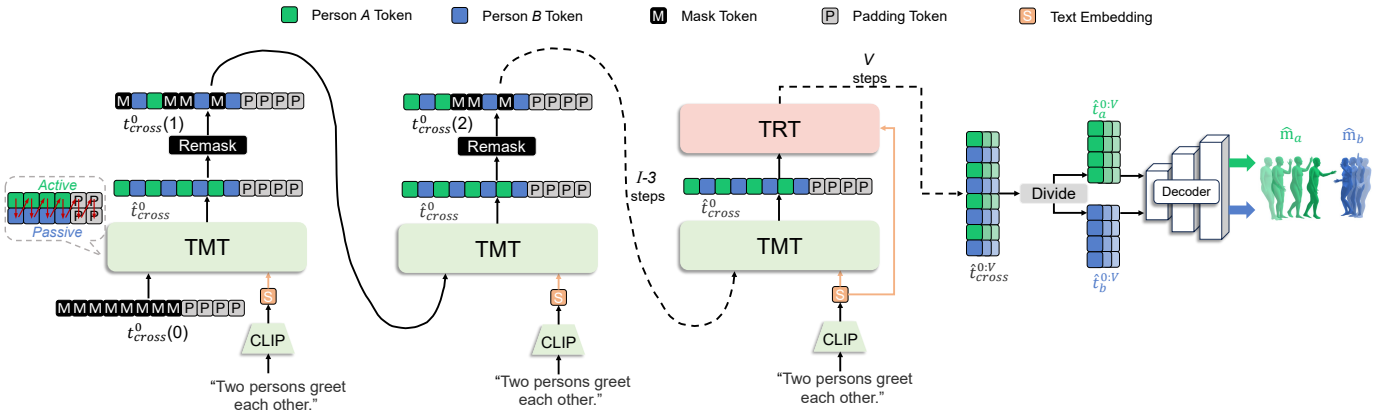

Fig. 4. **Inference Process.** Starting from an empty sequence $t^0_{cross}(0)$, TMT generates the base-layer token sequence $\hat{t}^0_{cross}$ over I iterations. Following this, TRT progressively predicts the remaining-layer token sequences $\hat{t}^{1:V}_{cross}$ within V steps.

be randomly replaced, and a 10% chance that it will be kept unchanged. Finally, we adopt *step unroll masking* from [56], where we remask a portion $(\gamma(\tau_{i+1}))$ of the predicted tokens with the lowest confidence scores and predict them again.

3) Identity-Aware Positional Encoding: In conventional approaches [5], [6], [57], identical positional encoding schemes are applied independently to Person A and Person B. This methodology presents two significant limitations: first, it fails to capture the distinct identity information of each individual; second, when motion sequences are interleaved in cross-modal inputs, it induces positional confusion, thereby impairing the model's ability to discern the respective positional contexts of A and B.

To mitigate these issues, we introduce a novel identity-aware positional encoding scheme that incorporates learnable person-specific embeddings. Specifically, we first apply temporal positional encoding to the entire cross-sequence to establish chronological order. We then assign role-based identities, where the person at odd indices (1st, 3rd, 5th, etc.) of the cross-sequence is labeled as "Active", while the person at even indices (2nd, 4th, 6th, etc.) is designated as "Passive". This dual encoding scheme enables the model to simultaneously capture temporal dynamics and distinguish between characteristic interactive behaviors, such as initiative and responsiveness during motion generation.

During training, the assignment of Active/Passive roles to specific persons is randomly alternated across different sequences to prevent the model from developing fixed character associations. When Person A is assigned the active role and Person B is assigned the passive role, the encoding is formulated as:

$$\begin{aligned} \mathbf{X}'_{\text{cross}} &= \mathcal{P}(\mathbf{X}_{\text{cross}}) + \mathbf{R}_{\text{roles}}, \\ \mathbf{R}_{\text{roles}} &= [\mathbf{p}^A_{\text{active}}, \mathbf{p}^B_{\text{passive}}, \mathbf{p}^A_{\text{active}}, \mathbf{p}^B_{\text{passive}}, \dots]. \end{aligned} \tag{7}$$

Here, $\mathcal{P}(\cdot)$ represents the standard cosine positional encoding applied to the entire cross-sequence $\mathbf{X}_{\text{cross}} = [\mathbf{x}^1_a, \mathbf{x}^1_b, \mathbf{x}^2_a, \mathbf{x}^2_b, \dots]$, providing continuous temporal indices $[0, 1, 2, 3, \dots, 2n-1]$. The role embedding matrix $\mathbf{R}_{\text{roles}}$ contains alternating role-specific embeddings, where $\mathbf{p}^A_{\text{active}}$ and $\mathbf{p}^B_{\text{passive}}$ denote the active and passive embeddings assigned to Person A and Person B respectively.

When roles are reversed during training (Person B as active, Person A as passive), the embedding assignment becomes $[\mathbf{p}^B_{\text{active}}, \mathbf{p}^A_{\text{passive}}, \mathbf{p}^B_{\text{active}}, \mathbf{p}^A_{\text{passive}}, \dots]$. This stochastic role-swapping strategy enhances the model's adaptability and generalization capability by encouraging it to capture contextually appropriate role assignments rather than relying on predetermined character associations.

4) Temporal Attention: As illustrated in Fig. 3(a), the input masked tokens are processed through a processing step to obtain the token embeddings $\mathbf{e}^{(l)}_{cross}$ for block l, where $l \in \{1, 2, \dots \mathbf{L}\}$. Furthermore, we utilize adaptive normalization layers [58] for regularization. The block architecture begins with a temporal attention module, followed by a feedforward network (FFN). This design captures the temporal dependencies throughout the entire motion token sequence, embedding $\mathbf{e}^{(l)}_{cross}$ into a crossover context vector $\mathbf{g}^{(l)}_{cross}$. The computation of $\mathbf{g}^{(l)}_{cross}$ is formulated as follows:

$$\begin{aligned} \mathbf{g}^{(l)}_{cross} &= FFN(Attn(\mathbf{Q}^{ta}, \mathbf{K}^{ta}, \mathbf{V}^{ta})) \\ &= FFN(softmax(\frac{\mathbf{Q}^{ta}(\mathbf{K}^{ta})^T}{\sqrt{W}})\mathbf{V}^{ta}), \end{aligned} \tag{8}$$

where the temporal query $\mathbf{Q}^{ta}$, temporal key $\mathbf{K}^{ta}$, and temporal value $\mathbf{V}^{ta}$ matrices are computed as linear projections of the crossover embeddings $\mathbf{e}^{(l)}_{cross}$. W is the number of channels in the attention layer.

5) Interactive Attention: As illustrated in Fig. 3(b), this module primarily consists of two attention mechanisms with shared parameters: self-attention and cross-attention. The self-attention mechanism ensures that each motion token sequence focuses on its own information, while the cross-attention mechanism facilitates interaction between the motion information of the two individuals. Using the formulation in Eq.5, we decompose the representation $\mathbf{g}^{(l)}_{cross}$ into its constituent components $\mathbf{g}^{(l)}_a$ and $\mathbf{g}^{(l)}_b$. Then, the self-attention mechanism embeds the current hidden states $\mathbf{g}^{(l)}$ into a context vector $\mathbf{c}^{(l)}$:

$$\begin{aligned} \mathbf{c}^{(l)}_a &= Attn(\mathbf{Q}^{sa}_a, \mathbf{K}^{sa}_a, \mathbf{V}^{sa}_a), \\ \mathbf{c}^{(l)}_b &= Attn(\mathbf{Q}^{sa}_b, \mathbf{K}^{sa}_b, \mathbf{V}^{sa}_b), \end{aligned} \tag{9}$$

where the attention matrices $\{\mathbf{Q}_a^{sa}, \mathbf{K}_a^{sa}, \mathbf{V}_a^{sa}\}$ and $\{\mathbf{Q}_b^{sa}, \mathbf{K}_b^{sa}, \mathbf{V}_b^{sa}\}$ are obtained by applying linear projections to their respective $\mathbf{g}_a^{(l)}$ and $\mathbf{g}_b^{(l)}$. Subsequently, the cross-attention mechanism embeds $\mathbf{c}^{(l)}$ into a context vector $\mathbf{e}^{(l+1)}$:

$$\begin{aligned}\mathbf{e}_a^{(l+1)} &= FFN(Attn(\mathbf{Q}_a^{ca}, \mathbf{K}_b^{ca}, \mathbf{V}_b^{ca})),\\ \mathbf{e}_b^{(l+1)} &= FFN(Attn(\mathbf{Q}_b^{ca}, \mathbf{K}_a^{ca}, \mathbf{V}_a^{ca})),\end{aligned} \tag{10}$$

where the $\{\mathbf{Q}_a^{ca}, \mathbf{K}_a^{ca}, \mathbf{V}_a^{ca}\}$ and $\{\mathbf{Q}_b^{ca}, \mathbf{K}_b^{ca}, \mathbf{V}_b^{ca}\}$ matrices are linear projections of $\mathbf{c}_a^{(l)}$ and $\mathbf{c}_b^{(l)}$ respectively. After obtaining $\mathbf{e}_a^{(l+1)}$ and $\mathbf{e}_b^{(l+1)}$, we can apply the formulation in Eq.5 to perform crossover merge operation, yielding $\mathbf{e}_{cross}^{(l+1)}$.

6) Training Objective: As defined in Eq.11, our TMT f_θ is optimized by minimizing the negative log-likelihood of target predictions, calculated by determining the cross-entropy loss between the one-hot encoded actual labels and the predictions generated by the model.

$$\mathcal{L}_{\text{mask}} = -\sum_{i=1}^{2n} \mathbb{E}_{\tilde{t}^0_{\text{cross}(i)}=[\text{MASK}]} \log f_\theta\big(t^0_{\text{cross}(i)} \big| \tilde{t}^0_{\text{cross}(i)}, c\big) \tag{11}$$

C. Conditional Temporal Residual Transformer

As shown in Fig. 2(c), we learn a TRT to model the remaining crossover tokens of V layers. Following the approach used for TMT, we employ Causal Crossover Merge to generate motion token sequences $[t^v_{cross}]_{v=1}^{V} = \left[\{t^v_{p(\lceil i/2 \rceil)}\}_{i=1}^{2n}\right]_{v=1}^{V}$, where $\lceil \cdot \rceil$ is the ceiling function, / denotes division operator, p is defined in Eq. 5.

Next, we randomly select a layer $q \in [1, V]$ for learning. TRT f_ϕ is trained to predict the v-th layer tokens based on the text c, the RQ layer indicator v, and the token embeddings. Here, the token embeddings summarize the embeddings of all previous layers. The overall training objective of our TRT is:

$$\mathcal{L}_{res} = -\sum_{v=1}^{q}\sum_{i=1}^{2n} \log f_\phi(t^v_{cross(i)} | t^{1:v-1}_{cross(i)}, c, v). \tag{12}$$

D. Inference

The inference process, as shown in Fig. 4, involves three stages. Initially, starting from a fully masked sequence, TMT generates token sequences for both individuals over I iterations. In each iteration i, TMT predicts the probabilities for the masked tokens, samples new tokens, and then remasks those tokens with the lowest $\lceil \gamma(\frac{i}{I}) \cdot 2n \rceil$ confidence scores until i reaches I. We use a cosine schedule $\gamma\left(\frac{i}{I}\right)$ to control the number of tokens retained, which increases as the iteration count i progresses. Once TMT completes its prediction, we can obtain an identity-sensitive temporal token sequence. TRT then proceeds to gradually predict the token sequences for the remaining quantization layers. Subsequently, the fully predicted crossover token sequences are split according to the method described in Eq. 5, yielding individual token sequences for person A and person B. Finally, the token sequences for both individuals are decoded and projected back to motion sequences through the RVQ-VAE decoder. Classifier Free Guidance (CFG) [59] is employed for both TMT and TRT to improve optimization, following the approach in [60]. The final logits ψ_g are computed by interpolating between the unconditional logits ψ_u (i.e., $c = \emptyset$) and the conditional logits ψ_c, scaled by a guidance factor s:

$$\psi_g = \psi_u + s \cdot (\psi_c - \psi_u). \tag{13}$$

IV. Experiments

A. Datasets

Our experiments are conducted on the two largest interactive motion datasets: InterHuman [5] and InterX [57].

InterHuman consists of 7,779 interactive motion sequences, each paired with 3 distinct textual annotations. It adopts the Amass [61] skeleton representation with 22 joints. Each single-person motion is represented as $\mathbf{m}_p = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f]$, where $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$ denotes global position, $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ represents global velocity, $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ indicates local rotation, and $\mathbf{c}^f \in \mathbb{R}^4$ signifies binary foot contact features, with $\mathbf{m}_p \in \mathbb{R}^{262}$. Here, N_j denotes the number of joints.

InterX, currently the largest dataset for human-human interaction, contains 11,388 interactive motion sequences, each also paired with 3 textual annotations. It follows the SMPL-X [62] skeleton representation, which includes 56 joints covering the body, hands, and face, along with root orientation and translation, represented as $\mathbf{m}_p \in \mathbb{R}^{336}$.

B. Evaluation Metrics

We adopt the evaluation metrics used by InterGen [5]. These include the *Frechet Inception Distance* (FID), which measures the dissimilarity between the distributions of generated interactions and the distributions of real interactions. We use *R-Precision* and *Multimodal Distance* (MM Dist) to assess the match between generated interactions and input text. We also compute the *Diversity* to evaluate the range of variations within the generated interaction distributions, and *Multi-Modality* (MModality) to calculate the average variance of generated interactions from a single text prompt. To compute these metrics, we borrow the text and motion encoders from [5], [57] to extract latent space features.

C. Implementation Details

Our HiTMM model is implemented using PyTorch. We use the frozen CLIP-ViTL/14 model [42] as our text encoder. The interactive motion RVQ-VAE consists of 6 quantization layers, each with a codebook of 512 512-dimensional codes and a downscale factor of 4. TMT contains 6 layers and 6 heads, with a latent dimension of 384. TRT contains 8 layers and 8 heads, with a latent dimension of 384. The batch size is set to 512 for training RVQ-VAE, and 52, 64 for training TMT and TRT. The learning rate for all models in HiTMM is set to 2e-4. During inference, we use CFG scale of 2 and 4 for TMT and TRT on both datasets. Finally, the number of iterations I during inference is set to 12 for both datasets.

TABLE II

QUANTITATIVE EVALUATION ON THE INTERHUMAN AND INTERX TEST SETS. THE IN2IN MODEL UTILIZES BOTH INTERACTION TEXT AND INDIVIDUAL ANNOTATIONS, WHILE IN2IN* RELIES SOLELY ON INTERACTION TEXT. ± INDICATES A 95% CONFIDENCE INTERVAL AND → MEANS THE CLOSER TO GROUND TRUTH THE BETTER. RED FACE INDICATES THE BEST RESULT, WHILE BLUE REFERS TO THE SECOND BEST.

Dataset	Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
		Top 1	Top 2	Top 3				
Inter Human	Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
	TEMOS [17]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$6.342^{\pm.015}$	$6.939^{\pm.71}$	$0.535^{\pm.014}$
	T2M [19]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
	MDM [21]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$7.125^{\pm.018}$	$7.602^{\pm.045}$	$2.350^{\pm.080}$
	ComMDM [46]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$6.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
	InterGen [5]	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$2.141^{\pm.063}$
	MoMat-MoGen [7]	$0.449^{\pm.004}$	$0.591^{\pm.003}$	$0.666^{\pm.004}$	$5.674^{\pm.085}$	$3.790^{\pm.001}$	$8.021^{\pm.35}$	$1.295^{\pm.023}$
	in2IN* [6]	$0.425^{\pm.008}$	$0.576^{\pm.008}$	$0.662^{\pm.009}$	$5.535^{\pm.120}$	$3.803^{\pm.002}$	$7.953^{\pm.047}$	$1.215^{\pm.023}$
	in2IN [6]	$0.455^{\pm.004}$	$0.611^{\pm.005}$	$0.687^{\pm.005}$	$5.177^{\pm.103}$	$3.790^{\pm.002}$	$7.940^{\pm.030}$	$1.061^{\pm.038}$
	InterMask [11]	$0.449^{\pm.004}$	$0.599^{\pm.005}$	$0.683^{\pm.004}$	$5.154^{\pm.061}$	$3.790^{\pm.002}$	$7.944^{\pm.033}$	$1.737^{\pm.020}$
	HiTMM	$0.452^{\pm.004}$	$0.613^{\pm.003}$	$0.697^{\pm.003}$	$5.017^{\pm.054}$	$3.789^{\pm.001}$	$7.971^{\pm.030}$	$0.916^{\pm.034}$
InterX	Ground Truth	$0.429^{\pm.004}$	$0.626^{\pm.003}$	$0.736^{\pm.003}$	$0.002^{\pm.0002}$	$3.536^{\pm.013}$	$9.734^{\pm.078}$	-
	TEMOS [17]	$0.092^{\pm.003}$	$0.170^{\pm.003}$	$0.238^{\pm.002}$	$29.258^{\pm.069}$	$6.867^{\pm.013}$	$4.738^{\pm.078}$	$0.672^{\pm.041}$
	T2M [19]	$0.184^{\pm.010}$	$0.298^{\pm.006}$	$0.396^{\pm.005}$	$5.481^{\pm.382}$	$9.576^{\pm.006}$	$2.771^{\pm.151}$	$2.761^{\pm.042}$
	MDM [21]	$0.203^{\pm.009}$	$0.329^{\pm.007}$	$0.426^{\pm.005}$	$23.701^{\pm.057}$	$9.548^{\pm.014}$	$5.856^{\pm.077}$	$3.490^{\pm.061}$
	ComMDM [46]	$0.090^{\pm.002}$	$0.165^{\pm.004}$	$0.236^{\pm.004}$	$29.266^{\pm.067}$	$6.870^{\pm.017}$	$4.734^{\pm.067}$	$0.771^{\pm.053}$
	InterGen [5]	$0.207^{\pm.004}$	$0.335^{\pm.005}$	$0.429^{\pm.005}$	$5.207^{\pm.216}$	$9.580^{\pm.011}$	$7.788^{\pm.208}$	$3.686^{\pm.052}$
	InterMask [11]	$0.403^{\pm.005}$	$0.595^{\pm.004}$	$0.705^{\pm.005}$	$0.399^{\pm.013}$	$3.705^{\pm.017}$	$9.046^{\pm.073}$	$2.261^{\pm.081}$
	HiTMM	$0.425^{\pm.003}$	$0.614^{\pm.004}$	$0.727^{\pm.003}$	$0.373^{\pm.012}$	$3.677^{\pm.013}$	$9.371^{\pm.064}$	$2.316^{\pm.065}$

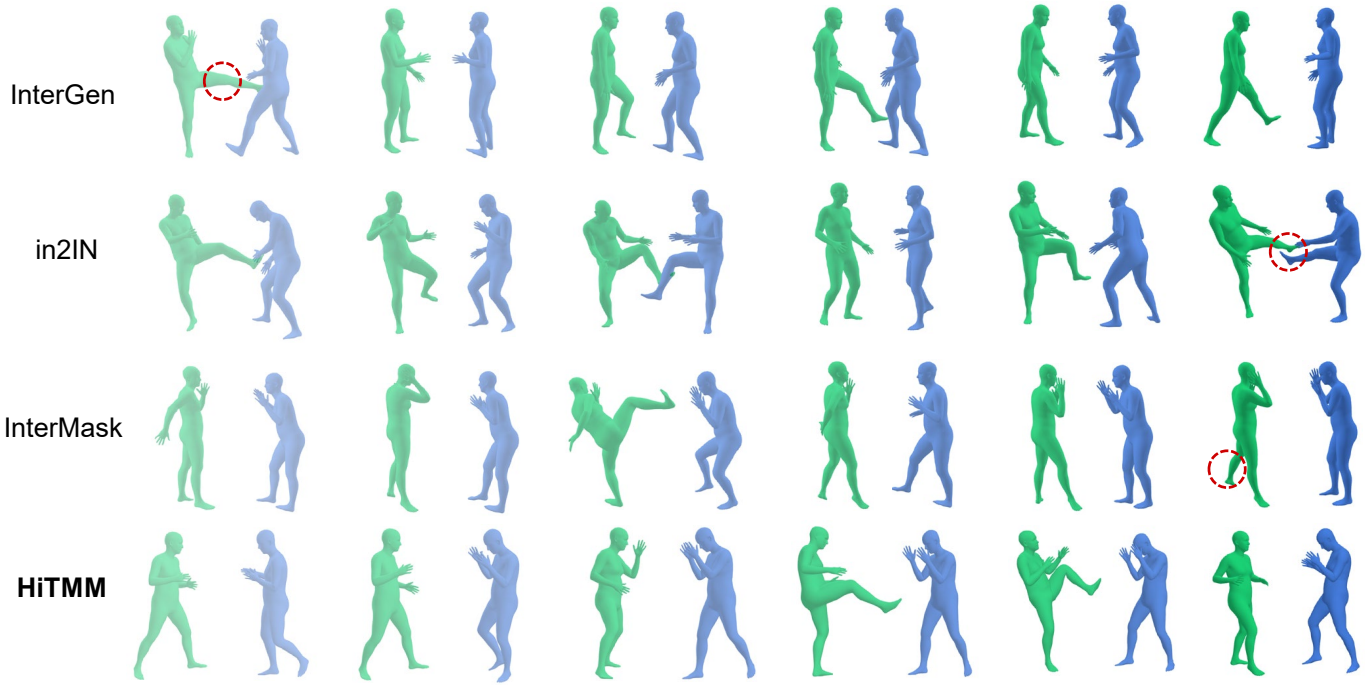

Fig. 5. **Interaction Description**: One person kicks the right leg first and then the left leg towards another person. The X-axis represents time.

D. Comparison with Previous Work

We evaluate our method with state-of-the-art approaches [5]–[7], [11], [17], [19], [21], [46] on the InterHuman and InterX datasets.

1) Quantitative Comparisons: Table II presents the evaluation of our HiTMM compared to previous methods. Following previous works [5], [11], We report the average values of 20 repeated generations with a 95% confidence interval. Overall, HiTMM establishes new state-of-the-art results on both datasets. Specifically, it achieves the lowest FID score (5.017 on InterHuman and 0.373 on InterX) while simultaneously outperforming all baselines for only interaction text in R-Precision and MM Dist. Notably, our model, which relies solely on interaction text, outperforms in2IN [6], a method

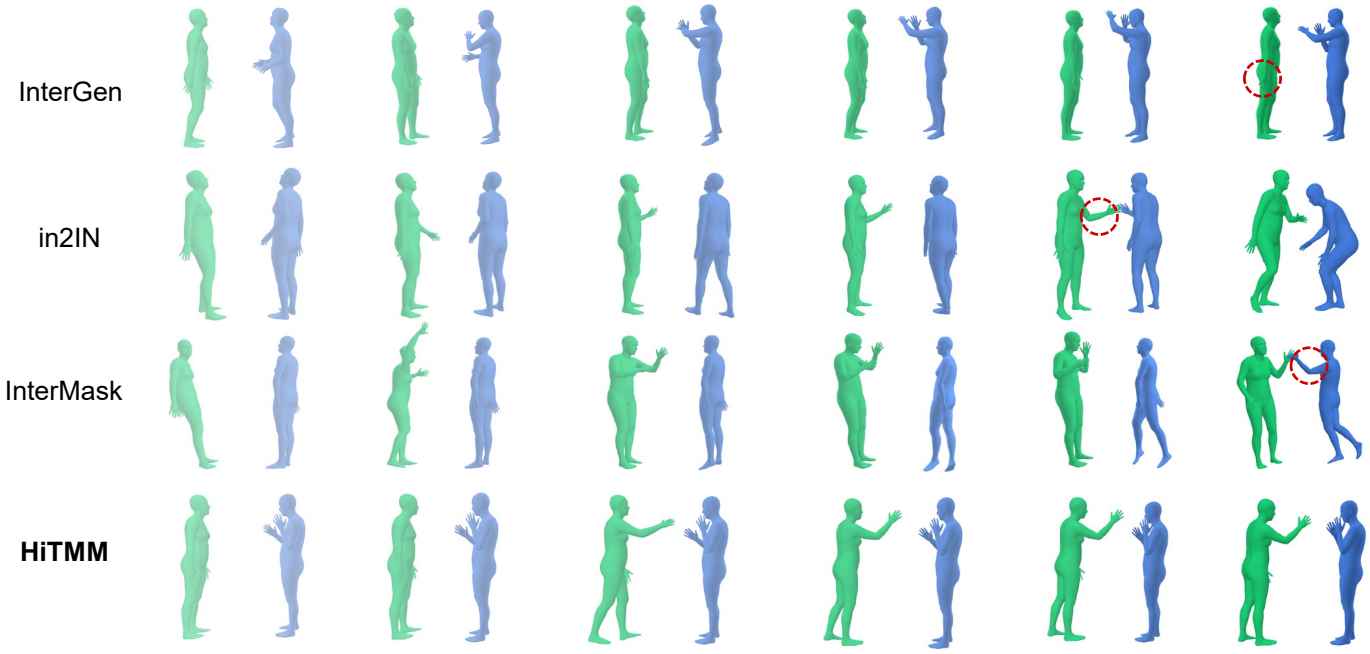

Fig. 6. **Interaction Description**: One person greets with his right hand, while the other person places his hands on his chest. The X-axis represents time.

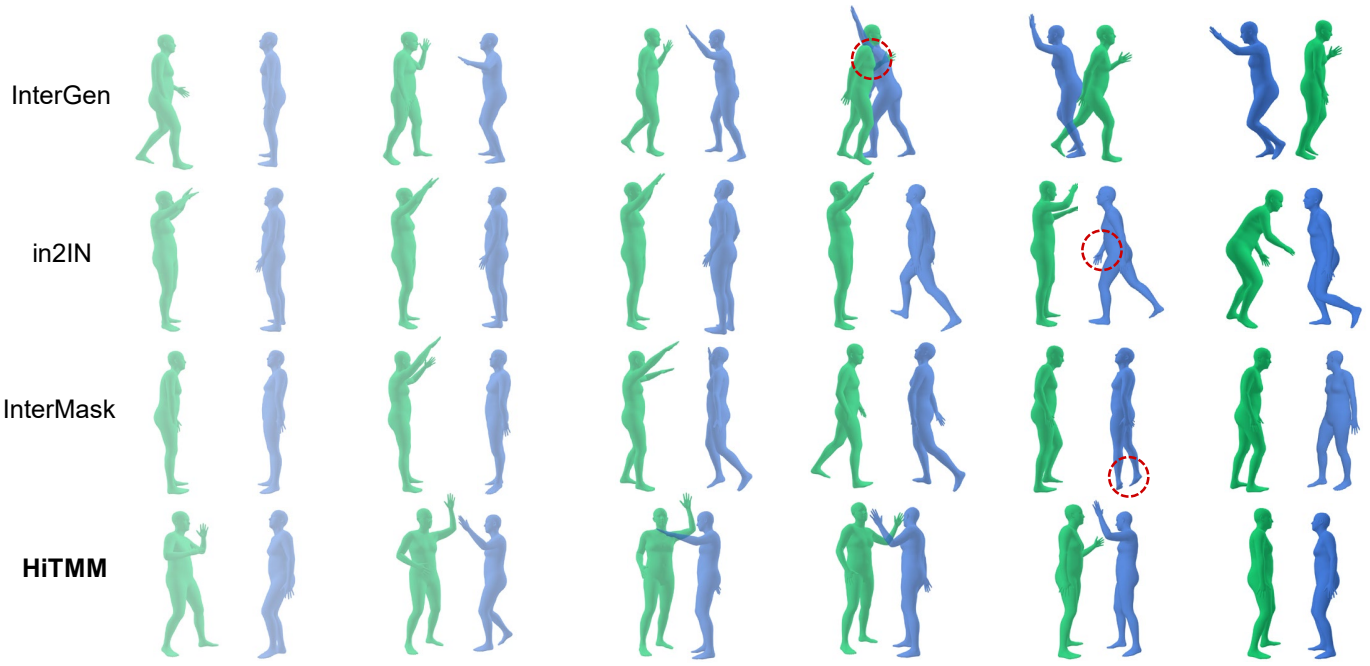

Fig. 7. **Interaction Description**: Two people exchange greetings with a wave of the hand. The X-axis represents time.

that utilizes both individual text information and interaction text, on key metrics such as R-Precision and FID. This further demonstrates the superiority of our approach. Although multimodality contributes to motion diversity, we prioritize the generation quality over MModality. In [9], it is revealed that an overemphasis on MModality without considering the quality of the generated results may lead to optimization issues, resulting in random outputs for any given input.

2) User Study: We further conduct a user study on the InterHuman dataset to evaluate our model against InterGen [5],

TABLE III
COMPARISON OF RECONSTRUCTION QUALITY AND TRAINING TIME BETWEEN OUR PROPOSED INTERACTIVE MOTION RVQ-VAE AND INTERMASK [11]'S SPATIO-TEMPORAL VQ-VAE. **BOLD** INDICATES THE BEST RESULT.

Method	FID ↓	R-Precision Top-3↑	MM Dist ↓	Training Time ↓
Spatio-temporal VQ-VAE	0.976	0.667	3.797	10h10m
Ours (RVQ-VAE)	**0.956**	**0.673**	**3.795**	**1h40m**

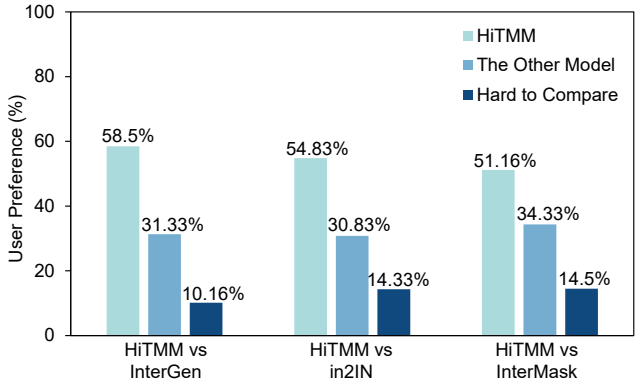

Fig. 8. User study comparing our HiTMM with InterGen [5], in2IN [6] and InterMask [11].

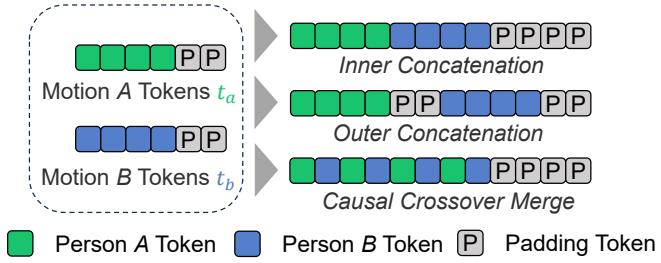

Fig. 9. Different token merging methods.

TABLE IV
ABLATION STUDY RESULTS ON THE INTERHUMAN TEST SET TO VALIDATE DIFFERENT MERGING METHODS. **BOLD** INDICATES THE BEST RESULT.

Token Merging Methods	R-Precision Top-1 ↑	FID ↓	MM Dist ↓	Diversity →
Inner Concatenation	0.427	5.120	3.795	7.982
Outer Concatenation	0.442	5.113	3.791	7.996
Causal Crossover Merge	**0.452**	**5.017**	**3.789**	**7.971**

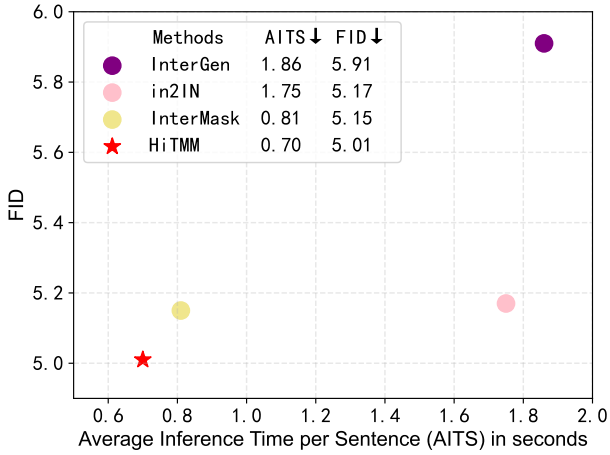

Fig. 10. The comparison of motion generation quality (FID score) and speed (AITS) between HiTMM and SOTA methods on the InterHuman dataset. All tests are conducted on the same NVIDIA 3090. A model closer to the origin indicates superior performance.

in2IN [6] and InterMask [11]. For HiTMM, InterGen [5] and InterMask [11], we generate motion sequences based on 30 different interaction texts. For in2IN [6], in addition to the interaction texts, we incorporate individual text information to compare with the best version of in2IN [6]. To ensure diversity in expertise, we recruit 20 participants for this user study, including 10 students from our field and 10 from other disciplines. They are asked to compare HiTMM side by side with each SOTA method, including InterGen [5], in2IN [6], and InterMask [11]. For each comparison, participants are randomly required to select one from three options: "HiTMM", "The Other Model", or "Hard to Compare" based on comprehensive evaluation of both generation quality and text alignment. As shown in Fig. 8, 58.5% of the participants preferred HiTMM over InterGen (31.33%), 54.83% preferred HiTMM over in2IN (30.83%) and 51.16% preferred HiTMM against InterMask (34.33%).

3) Qualitative Comparisons: We compare our generations with the current state-of-the-art models: InterGen [5], in2IN [6] and InterMask [11]. As shown in Fig.5, Fig.6 and Fig.7, HiTMM generates highly realistic interactions, at the same time well aligned with textual descriptions. Through qualitative evaluation, we observe that HiTMM consistently outperforms InterGen [5], in2IN [6] and InterMask [11] across various scenarios. For instance, in the scenario described as "One person kicks the right leg first and then the left leg towards another person", InterGen incorrectly shows kicking the left leg first, while in2IN depicts one person kicking the left leg and the other kicking the right leg. InterMask shows one person kicking the right leg but not subsequently kicking the left leg. In the case of "One person greets with his right hand, while the other person places his hands on his chest", InterGen does not show a person raising their right hand. Meanwhile, both in2IN and InterMask incorrectly show the person raising their left hand. Finally, for the prompt "Two people exchange greetings with a wave of the hand", InterGen suffers from severe interpenetration issues between the two individuals, while in2IN only depicts one person waving, thereby failing to capture the concept of "exchange". While InterMask generates motions consistent with the text, it exhibits noticeable artifacts such as abnormal foot movements. In contrast, our method generates high-quality motions that closely match the interactive text descriptions, demonstrating superior alignment and realism. For fair comparison, the images shown in Fig. 5, Fig. 6, and Fig. 7 are sampled using the same fixed time interval as all compared methods. Please refer to the supplementary demo videos for dynamic visualizations.

4) Comparison of Quantization Methods: Table III presents a comparison between our method and the spatio-temporal VQ-VAE of InterMask [11] in terms of reconstruction quality and training time. Our coarse-to-fine quantization approach not only yields more realistic reconstruction results but also requires significantly less training time. Specifically, for training, our method completes in just 1 hour and 40 minutes on a single NVIDIA 3090 GPU, which is significantly faster than the 10 hours and 10 minutes required by InterMask's spatio-

TABLE V
ABLATION STUDIES CONDUCTED ON THE INTERHUMAN TEST SET TO VERIFY THE ESSENTIAL COMPONENTS OF THE PROPOSED TMT. W/O DENOTES "WITHOUT". **BOLD** FACE INDICATES THE BEST RESULT.

Method	R-Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
Individual Mask	$0.441^{\pm.006}$	$0.603^{\pm.005}$	$0.683^{\pm.003}$	$5.214^{\pm.078}$	$3.790^{\pm.001}$	**7.943**$^{\pm.032}$	$0.915^{\pm.032}$
w/o Identity-Aware Positional Encoding	$0.443^{\pm.004}$	$0.597^{\pm.004}$	$0.685^{\pm.005}$	$5.113^{\pm.071}$	$3.790^{\pm.001}$	$7.960^{\pm.035}$	**0.904**$^{\pm.033}$
w/o Temporal Attention	$0.441^{\pm.005}$	$0.599^{\pm.005}$	$0.684^{\pm.004}$	$5.230^{\pm.072}$	$3.790^{\pm.001}$	$7.963^{\pm.033}$	$0.908^{\pm.033}$
w/o Interactive Attention	$0.445^{\pm.005}$	$0.602^{\pm.003}$	$0.681^{\pm.004}$	$5.519^{\pm.079}$	$3.790^{\pm.001}$	$7.982^{\pm.030}$	$0.886^{\pm.035}$
HiTMM	**0.452**$^{\pm.004}$	**0.613**$^{\pm.003}$	**0.697**$^{\pm.003}$	**5.017**$^{\pm.054}$	**3.789**$^{\pm.001}$	$7.971^{\pm.030}$	$0.917^{\pm.035}$

TABLE VI
ABLATION STUDIES ON THE RESIDUAL QUANTIZATION LAYERS DURING TRAINING. **BOLD** REPRESENTS THE BEST RESULT.

Residual Quantization Layers	R-Precision Top-1 ↑	FID ↓	MM Dist ↓	Diversity →	MModality ↑
(V,0)	0.382	7.512	3.813	7.822	0.830
(V,1)	0.400	6.935	3.806	7.834	0.897
(V,2)	0.423	6.608	3.801	7.913	0.912
(V,3)	0.440	5.386	3.795	7.892	0.911
(V,4)	0.446	5.203	3.794	**7.940**	0.941
(V,5)	**0.452**	**5.017**	**3.789**	7.971	0.916
(V,6)	0.447	5.226	3.794	7.983	**0.955**
(V,7)	0.432	5.782	3.795	8.001	0.946

TABLE VII
ABLATION STUDIES ON QUALITY INFLUENCED BY THE NUMBER OF CODES AND THE CODE DIMENSION. **BOLD** INDICATES THE BEST RESULT.

[number of code , code of dimension]	R-Precision Top-1 ↑	FID ↓	MM Dist ↓	Diversity →
[256 , 1024]	0.434	5.853	3.801	**7.942**
[512 , 512]	**0.452**	**5.017**	**3.789**	7.971
[1024 , 256]	0.447	5.563	3.793	7.847

temporal VQ-VAE.

5) Speed and Efficiency Advantages: In addition to its exceptional fidelity and text alignment performance, HiTMM also demonstrates significant advantages in training efficiency and inference speed. In terms of training efficiency, HiTMM completes all training phases in just 3 days on a single NVIDIA 3090 GPU, which is 70% more efficient than both InterGen [5] and in2IN [6] requiring 5 days with two NVIDIA 3090 GPUs. Regarding inference speed, as shown in Fig. 10, HiTMM achieves excellent inference efficiency while maintaining the best FID score (5.017), outperforming all other baseline methods.

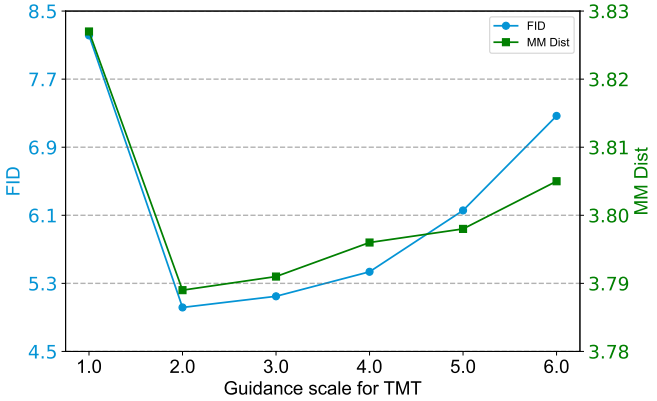

Fig. 11. Evaluation of the guidance scale s for TMT during inference.

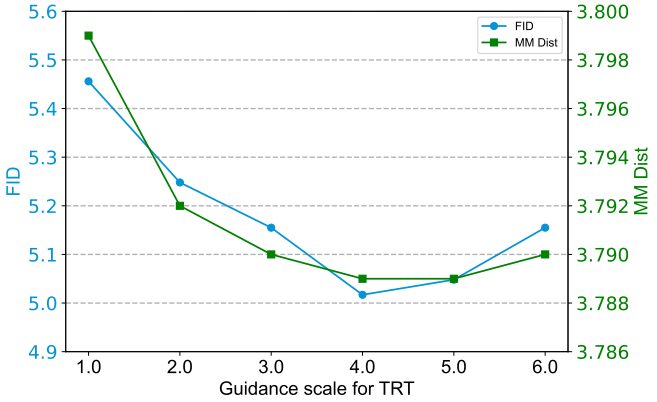

Fig. 12. Evaluation of the guidance scale s for TRT during inference.

E. Ablation Study

1) Merging Methods: Fig. 9 illustrates different merging methods for two-person motion tokens. *Inner Concatenation* refers to concatenating the valid tokens of both individuals while padding the remaining tokens. *Outer Concatenation* denotes concatenating directly the entire set of motion tokens of the two persons. *Causal Crossover Merge* represents causally merging the two-person motion tokens. We apply these merging methods to both TMT and TRT, with the ablation results shown in Table IV. Among the three merging methods, *Causal Crossover Merge* achieves the best performance, effectively combining the human-human motion tokens along a single timeline and significantly improving the quality of motion generation.

2) TMT: Table V presents ablation studies to evaluate the key components of our TMT. We compare different modeling approaches, attention mechanisms, and masking strategies. *Individual Mask* indicates randomly masking the discrete token sequences of person A and person B separately. Notably, the removal of any key component leads to a decrease in performance. Our complete HiTMM model outperforms all ablated versions, achieving the best results on primary metrics such as FID, R-Precision, and MM Dist. These results highlight

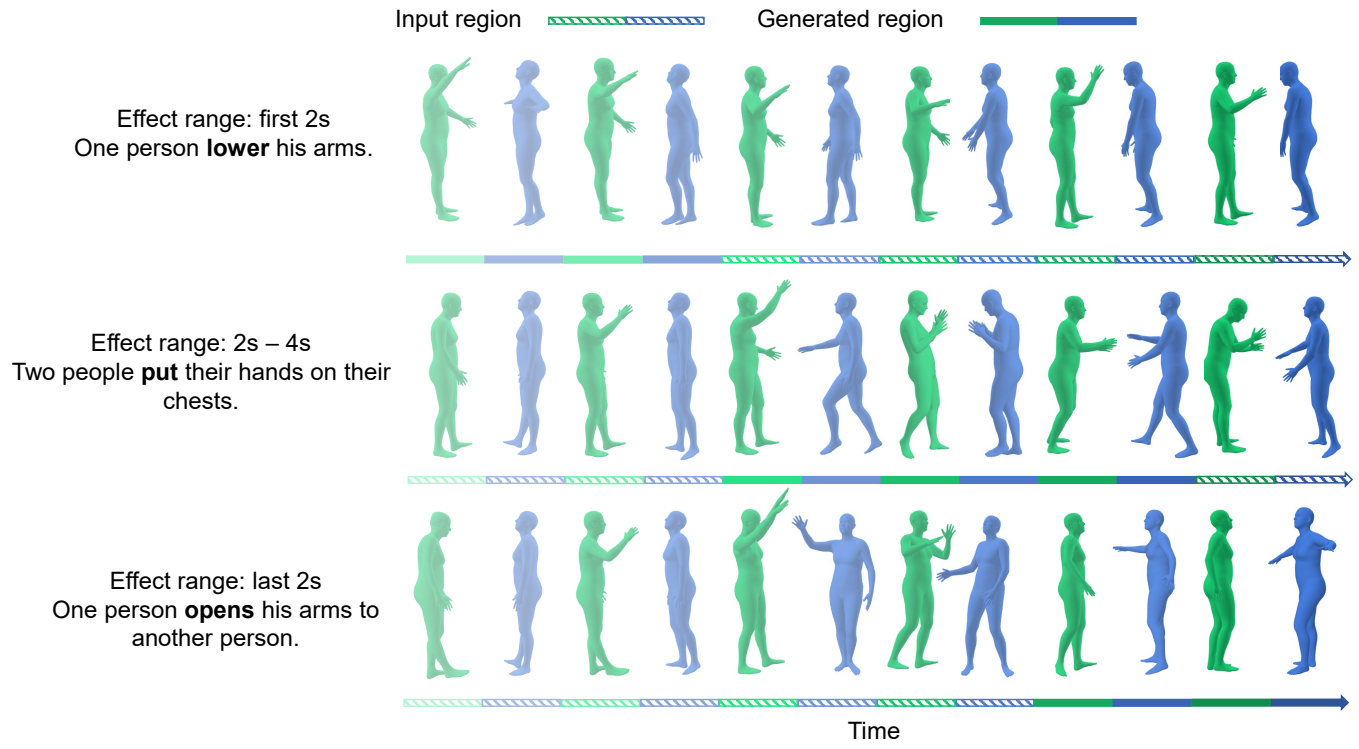

Fig. 13. **Examples of temporal editing.** The slashed region indicates the motion range provided by the reference sequence. The solid region represents the motion range generated by HiTMM, which depends on the text prompt to the left. The X-axis represents time.

the importance of all components in TMT architecture for modeling human interactions.

3) Residual Quantization Layers (V): Table VI shows the impact of different quantization layers. More quantization layers enables more accurate reconstruction in the Interactive Motion Tokenizer, but also increases the burden on TRT for generating residual tokens. We observe that when $V = 5$, the optimal balance is achieved between reconstruction quality and computational cost.

4) Codebook Size: Table VII demonstrates the results of increasing the number of codebook entries from 256 to 1024 while reducing the dimensionality of the codebook latent space from 1024 to 256. While a larger codebook size enables richer motion representation, it may introduce redundancy in the codebook. Our experiments reveal that the optimal balance is achieved with 512 codebook entries coupled with a 512-D latent space, which delivers peak performance without compromising efficiency.

F. Inference Hyper-parameter

The CFG scale is a critical hyperparameter in the masked modeling inference process. As shown in Fig. 11 and Fig. 12, we demonstrate the performance curves of FID and MM Dist by scanning different s values for TMT and TRT on the InterHuman dataset. The optimal guidance scale for TMT inference is set around $s = 2$. Over-guidance during inference may even degrade performance. For TRT, the best performance peaks at $s = 4$ with an FID of 5.017, which is better than $s = 5$ with an FID of 5.048, while the MM Dist remains the same at 3.789 in both cases.

TABLE VIII
QUANTITATIVE RESULTS OF TEMPORAL EDITING FOR THE PREFIX, IN-BETWEENING, AND SUFFIX GENERATION.

Editing Location	R-Precision Top-1 ↑	FID ↓	MM Dist ↓	Diversity →
Prefix	0.458	4.012	3.785	7.948
In-betweening	0.455	4.465	3.787	7.974
Suffix	0.459	4.101	3.785	7.965

G. Application: Temporal Editing

By leveraging a unified timeline structure where each discrete token encodes full motion dynamics, along with its inherent mask-driven design, HiTMM is capable of freely editing any continuous temporal region within an interaction sequence, similar to how it operates in single-person scenarios. In Fig. 13, we demonstrate the ability of HiTMM to temporally edit specific regions within a human interaction sequence. The interaction sequence is generated based on the textual description "These two people greet each other". Then, these regions can be freely selected from the first half, the middle, or the latter half of the sequence. In particular, we can mask tokens in specific regions and follow the same inference steps described in Section III-D. We conduct a quantitative evaluation by performing inpainting on three temporal segments: the initial 25% (Prefix), the middle 25%-75% (In-betweening), and the

final 25% (Suffix). The corresponding quantitative results for temporal editing are presented in Table VIII.

V. LIMITATION AND FUTURE WORK

Although HiTMM demonstrates strong performance in fidelity and text-motion alignment for interactive motion generation, there exist some limitations. Firstly, occasional body jittering may occur when visualizing SMPL [63] meshes, due to the absence of physical constraints and interactions with the physical environment during the motion generation process. Therefore, integrating interactions with a physical environment represents a promising direction for future work to enhance the physical plausibility of the generated motions. Secondly, HiTMM requires the target motion length as input for generation, a limitation stemming from the inherent design of current masked generation models. This could potentially be addressed by pretraining the t2m [19] model on paired interaction datasets and applying the text2length sampling. Finally, for fast-changing motions, sudden position swapping between the two characters may occur. This may be caused by the overly compact discrete tokens representing paired interactions, which limits the model's ability to fully capture the specific roles of each individual. To address this, enhancing the identity information of both characters while ensuring temporal coherence presents a promising avenue for future research. We plan to investigate and address the aforementioned limitations in future work.

VI. CONCLUSION

In this work, we present HiTMM, a novel temporal masked motion model that generates identity-aware, high-quality 3D human interactive motions from textual descriptions while maintaining causal temporal control. HiTMM comprises three key components: (1) an interactive motion tokenizer that transforms 3D human interactions into multi-layer latent space tokens; (2) a conditional temporal masked motion transformer that predicts the base-layer motion tokens conditioned on interaction text; and (3) a conditional temporal residual motion transformer that predicts the remaining motion tokens using the same text condition. HiTMM reformulates the generation of two-person motions as the generation of human motions along a single timeline. It also incorporates identity-aware guidance into the generation process, thereby further enhancing the temporal controllability of human-human interaction synthesis. Extensive experiments demonstrate that HiTMM outperforms state-of-the-art methods both qualitatively and quantitatively. Benefiting from its superior temporal control and masking properties, HiTMM achieves temporal motion editing across multiple contexts in the field of human interactive motion generation.

REFERENCES

[1] Y. Tian, H. Zhang, Y. Liu, and L. Wang, "Recovering 3d human mesh from monocular images: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 12, pp. 15 406–15 425, 2023.

[2] W. Yao, H. Zhang, Y. Sun, and J. Tang, "Staf: 3d human mesh recovery from video with spatio-temporal alignment fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[3] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 446–11 456.

[4] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu, "Pymaf-x: Towards well-aligned full-body model regression from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 287–12 303, 2023.

[5] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "Intergen: Diffusion-based multi-human motion generation under complex interactions," *International Journal of Computer Vision*, pp. 1–21, 2024.

[6] P. Ruiz-Ponce, G. Barquero, C. Palmero, S. Escalera, and J. García-Rodríguez, "in2in: Leveraging individual information to generate human interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 1941–1951.

[7] Z. Cai, J. Jiang, Z. Qing, X. Guo, M. Zhang, Z. Lin, H. Mei, C. Wei, R. Wang, W. Yin *et al.*, "Digital life project: Autonomous 3d characters with social intelligence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 582–592.

[8] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan, "Generating human motion from textual descriptions with discrete representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 730–14 740.

[9] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1900–1910.

[10] E. Pinyoanuntapong, P. Wang, M. Lee, and C. Chen, "Mmm: Generative masked motion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1546–1555.

[11] M. G. Javed, C. Guo, L. Cheng, and X. Li, "Intermask: 3d human interaction generation via collaborative masked modeling," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=ZAyuwJYN8N

[12] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.

[13] J. Martinez, H. H. Hoos, and J. J. Little, "Stacked quantizers for compositional vector compression," *arXiv preprint arXiv:1411.2173*, 2014.

[14] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[15] H. Yao, Z. Song, Y. Zhou, T. Ao, B. Chen, and L. Liu, "Moconvq: Unified physics-based motion control via scalable discrete representations," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–21, 2024.

[16] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, "Maskedmimic: Unified physics-based character control through masked motion inpainting," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–21, 2024.

[17] M. Petrovich, M. J. Black, and G. Varol, "Temos: Generating diverse human motions from textual descriptions," in *European Conference on Computer Vision*. Springer, 2022, pp. 480–497.

[18] E. Pinyoanuntapong, M. U. Saleem, P. Wang, M. Lee, S. Das, and C. Chen, "Bamm: bidirectional autoregressive motion model," in *European Conference on Computer Vision*. Springer, 2024, pp. 172–190.

[19] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.

[20] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[21] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=SJ1kSyO2jwu

[22] W. Song, X. Jin, S. Li, C. Chen, A. Hao, and X. Hou, "Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4361–4371, 2023.

[23] X. Gao, Y. Yang, Z. Xie, S. Du, Z. Sun, and Y. Wu, "Guess: Gradually enriching synthesis for text-driven human motion generation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 12, pp. 7518–7530, 2024.

[24] X. Shi, W. Yao, C. Luo, J. Peng, H. Zhang, and Y. Sun, "Fg-mdm: Towards zero-shot human motion generation via chatgpt-refined descriptions," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 446–461.

[25] W. Yao, Y. Sun, H. Zhang, Y. Liu, and J. Tang, "Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 14, 2026, pp. 11 901–11 909.

[26] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, and X. Wang, "Tm2d: Bimodality driven 3d dance generation via music-text integration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9942–9952.

[27] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, "Bailando: 3d dance generation by actor-critic gpt with choreographic memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 050–11 059.

[28] Z. Zhou and B. Wang, "Ude: A unified driving engine for human motion generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5632–5641.

[29] J. Zhang, M. Zhu, Y. Zhang, Z. Zheng, Y. Liu, and K. Li, "Speechact: Towards generating whole-body motion from speech," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[30] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, L. Bao, and Z. He, "Audio2gestures: Generating diverse gestures from audio," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[31] 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 transactions on visualization and computer graphics*, vol. 29, no. 8, pp. 3519–3534, 2022.

[32] Z. Yang, Y.-H. Wen, S.-Y. Chen, X. Liu, Y. Gao, Y.-J. Liu, L. Gao, and H. Fu, "Keyframe control of music-driven 3d dance generation," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[33] J. Zhang, Y. Sun, H. Zhang, and J. Tang, "Edmg: Towards efficient long dance motion generation with fundamental movements from dance genres," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 10 447–10 456.

[34] Z. Liu, S. Wu, S. Jin, S. Ji, Q. Liu, S. Lu, and L. Cheng, "Investigating pose representations and motion contexts modeling for 3d motion prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 681–697, 2022.

[35] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9489–9497.

[36] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, "Dance revolution: Long-term dance generation with music via curriculum learning," in *International conference on learning representations*, 2020.

[37] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek, "Synthesis of compositional animations from textual descriptions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1396–1406.

[38] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2021–2029.

[39] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 985–10 995.

[40] C. Guo, X. Zuo, S. Wang, X. Liu, S. Zou, M. Gong, and L. Cheng, "Action2video: Generating videos of human 3d actions," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 285–315, 2022.

[41] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action composition for 3d humans," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 414–423.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[43] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[44] Q. Wang, X. Zheng *et al.*, "Doodle your motion: Sketch-guided human motion generation [j]," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[45] B. Ji, Y. Pan, Z. Liu, S. Tan, and X. Yang, "Sport: From zero-shot prompts to real-time motion generation," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[46] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," in *The Twelfth International Conference on Learning Representations*, 2024.

[47] L. Xu, Y. Zhou, Y. Yan, X. Jin, W. Zhu, F. Rao, X. Yang, and W. Zeng, "Regennet: Towards human action-reaction synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1759–1769.

[48] B. Chopin, H. Tang, N. Otberdout, M. Daoudi, and N. Sebe, "Interaction transformer for human reaction generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 8842–8854, 2023.

[49] M. Tanaka and K. Fujiwara, "Role-aware interaction generation from textual description," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15 999–16 009.

[50] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir, "Deep motifs and motion signatures," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–13, 2018.

[51] C. Guo, X. Zuo, S. Wang, and L. Cheng, "Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts," in *European Conference on Computer Vision*. Springer, 2022, pp. 580–597.

[52] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[53] T. Lucas, F. Baradel, P. Weinzaepfel, and G. Rogez, "Posegpt: Quantization-based 3d human motion generation and forecasting," in *European Conference on Computer Vision*. Springer, 2022, pp. 417–435.

[54] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang, "Motiongpt: Finetuned llms are general-purpose motion generators," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7368–7376.

[55] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.

[56] S. Kim, D. Jo, D. Lee, and J. Kim, "Magvlt: Masked generative vision-and-language transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 338–23 348.

[57] L. Xu, X. Lv, Y. Yan, X. Jin, S. Wu, C. Xu, Y. Liu, Y. Zhou, F. Rao, X. Sheng *et al.*, "Inter-x: Towards versatile human-human interaction analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 260–22 271.

[58] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[59] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI

[60] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 315–11 325.

[61] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[62] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[63] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

Zicheng Jiao received the B.E. degree from the Xizang University, Lasa, China, in 2023. He is now a Ph.D. student in the School of Computer Science and Engineering at Nanjing University of Science and Technology, Nanjing, China. His research interests include computer vision, motion generation.

Yunlian Sun received the M.E. degree in computer science and technology from Harbin Institute of Technology, China, in 2010, and the Ph.D. degree in ingegneria elettronica, informatica e delle telecomunicazioni from the University of Bologna, Italy, in 2014. After the Ph.D. degree, she worked as a Post-Doctoral Researcher with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. She is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include computer vision, pattern recognition, generative AI and embodied AI.

Hongwen Zhang received the B.E. degree from the South China University of Technology, Guangzhou, China, in 2015, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021, respectively. He has been working as a Post-Doctoral Researcher at Tsinghua University and is currently an Associate Professor at the School of Artificial Intelligence, Beijing Normal University. His research interests include computer vision, computer graphics, and their applications in 3D human modeling.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing Forestry University, Nanjing, China. He has authored more than 200 articles in toptier journals and conferences. His research interests include multimedia analysis and computer vision. Dr.Tang was a recipient of the Best Paper Awards in ACM MM 2007 and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015. He has served as an Associate Editor for the IEEE TNNLS, IEEE TKDE, IEEE TMM, and IEEE TCSVT. He is a Fellow of IAPR.

Massimo Tistarelli (Senior Member, IEEE) received the Ph.D. degree in computer science and robotics from the University of Genoa, Genoa, Italy, in 1991. He is currently a tenured Full Professor of Computer Science and the Director of the Computer Vision Laboratory with the University of Sassari, Sassari, Italy. His main research interests cover biological and artificial vision (in particular, recognition, 3-D reconstruction, and dynamic scene analysis), pattern recognition, biometrics, visual sensors, robotic navigation, and visuomotor coordination. He has coauthored over 100 scientific papers in peer-reviewed books, conferences, and international journals. He is one of the world-recognized leading researchers in biometrics. He is an Associate Editor of IEEE TPAMI, Pattern Recognition Letters, and Image and Vision Computing. He is a fellow of the IAPR and an IEEE Distinguished Lecturer Program member.