2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.00672

TIMotion: Temporal and Interactive Framework for Efficient Human-Human Motion Generation

Yabiao Wang[1,2*], Shuo Wang[2*], Jiangning Zhang[2], Ke Fan[3], Jiafu Wu[2], Zhucun Xue[1], Yong Liu[1†]
[1]Zhejiang University [2]Youtu Lab, Tencent [3]Shanghai Jiao Tong University
https://aigc-explorer.github.io/TIMotion-page/

Abstract

Human-human motion generation is essential for understanding humans as social beings. Current methods fall into two main categories: single-person-based methods and separate modeling-based methods. To delve into this field, we abstract the overall generation process into a general framework MetaMotion, which consists of two phases: temporal modeling and interaction mixing. For temporal modeling, the single-person-based methods concatenate two people into a single one directly, while the separate modeling-based methods skip the modeling of interaction sequences. The inadequate modeling described above resulted in sub-optimal performance and redundant model parameters. In this paper, we introduce TIMotion (Temporal and Interactive Modeling), an efficient and effective framework for human-human motion generation. Specifically, we first propose Causal Interactive Injection to model two separate sequences as a causal sequence leveraging the temporal and causal properties. Then we present Role-Evolving Scanning to adjust to the change in the active and passive roles throughout the interaction. Finally, to generate smoother and more rational motion, we design Localized Pattern Amplification to capture short-term motion patterns. Extensive experiments on InterHuman and InterX demonstrate that our method achieves superior performance.

1. Introduction

In the field of generative computer vision, human motion generation has significant implications for computer animation [18, 20], game development [3, 33], and robotic control [26, 36, 37]. In recent years, there have been remarkable advancements in human motion generation, driven by various user-specified conditions such as action categories [9, 22], speeches [2, 12], and natural language prompts [11, 24]. Among these, many approaches leveraging large language models [1] and diffusion models [43] have yielded impressive results in generating realistic and diverse motions, benefiting from their powerful modeling capabilities. Despite this progress, most existing methods are designed primarily for single-person scenarios, thereby neglecting a crucial element of human motion: the complex and dynamic interactions between individuals. We address the challenge of generating two-person motion by fully utilizing the temporal and interactive dynamics between the individuals.

To better explore human-human motion generation, we abstract a general framework *MetaMotion*, as shown on the left of Fig. 1, which consists of two phases: temporal modeling and interaction mixing. Previous approaches have prioritized interaction mixing over temporal modeling and can be divided into two primary categories: single-person-based methods and separate modeling-based methods. As shown in Fig. 1(a), the single-person-based methods (*e.g.* MDM [32]) concatenate two individuals into one, which is then fed into the existing single-person motion generation module, *e.g.* DiT. And the separate modeling-based methods (*e.g.* InterGen [16]), as shown in Fig. 1(b), model two individuals individually and then extract motion information from both themselves and each other, using self-attention and cross-attention mechanisms, respectively. Following the general logic of *MetaMotion*, we introduce the Temporal and Interactive Framework, as depicted in Fig. 1(c), which models the human-human causal interactions. In fact, this effective temporal modeling method can simplify the design of the interaction mixing module and reduce the number of learnable parameters.

In this paper, we design an effective temporal modeling method and introduce TIMotion, a temporal and interactive framework for efficient human-human motion generation. Specifically, we first propose Causal Interactive Injection to model the two single-person motion sequences as a causal interaction sequence, according to the temporal causal prop-

*Equal contributions.
†Corresponding author.

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.00672

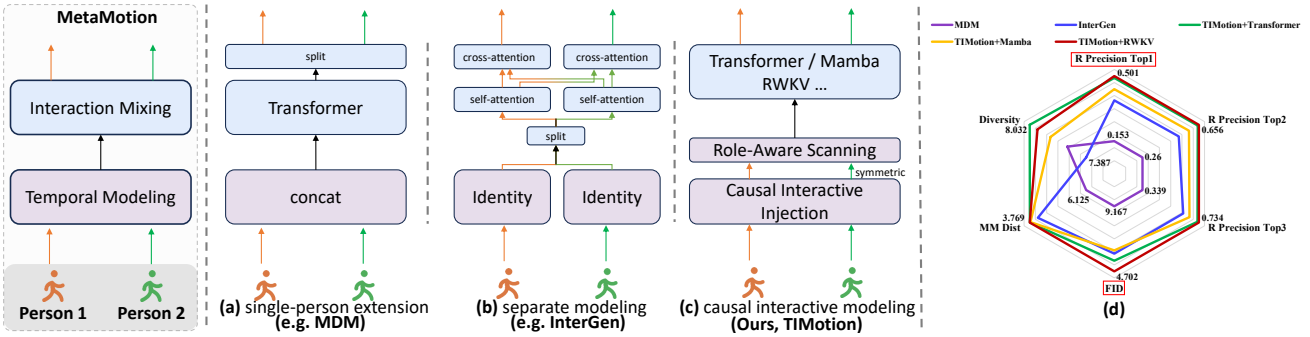

Figure 1. **MetaMotion and performance of MetaMotion-based models on InterHuman validation set.** We abstract the MetaMotion concept that illustrates the intrinsic properties of human-human motion generation in the interaction process. (a) and (b) show the two types of methods currently, and (c) shows our method TIMotion, LPA refers to the Localized Pattern Amplification. In (d) we compare the performance of the different methods on the InterHuman dataset.

erties of motion sequences. Second, since active and passive roles are not static in interactions, we propose Role-Evolving Mixing to make two humans act in both active and passive roles. Then the network can dynamically adjust the roles of the two humans based on the text's semantics and the motion's context. Finally, we propose Localized Pattern Amplification, which captures short-term motion patterns for each individual separately, resulting in smoother and more logical motion generation. Our proposed framework can be well adapted to different interaction-mixing modules, including Transformer, RWKV, and Mamba.

Our contributions can be summarized as follows:

- We conceptualize human-human motion generation within a general framework MetaMotion and design an innovative method. Our proposed framework, TIMotion, is versatile enough to integrate with various interaction-mixing modules (*e.g.* Transformer, RWKV, Mamba) and reduces the number of parameters of these modules.
- To utilize the temporal and causal properties, we propose Causal Interactive Injection to model two separate motion sequences as a unified causal sequence. In addition, we introduce Role-Evolving Scanning to accommodate shifts between the active and passive roles during interactions. Moreover, we also design Localized Pattern Amplification to capture short-term motion patterns effectively.
- We perform extensive experiments on the benchmark human-human motion generation datasets: InterHuman and Inter-X. The results demonstrate the effectiveness and generalizability of our proposed methods.

2. Related Work

Single-Person Human Motion Generation. Creating human motion is vital for applications such as 3D modeling and robot manipulation. The primary approach, known as the Text-to-Motion task, involves learning a unified latent space for both language and motion.

Autoencoders have been widely adopted in human motion generation. MotionCLIP [30] effectively integrates semantic knowledge from CLIP [25] into the human motion manifold. TEMOS [23] and T2M [10] combine a Transformer-based VAE with a text encoder to generate motion sequences based on text descriptions. AttT2M [44] and TM2D [7] integrate a spatial-temporal body-part encoder into VQ-VAE [34] to improve the learning of discrete latent space. T2M-GPT [40] redefines text-driven motion generation as a next-index prediction task and proposes a framework based on VQ-VAE and Generative Pretrained Transformer (GPT) for motion generation.

Recently, diffusion-based generative modeling has been gaining attention. MotionDiffuse [42] exhibits several desirable properties, such as probabilistic mapping, realistic synthesis, and multi-level manipulation. MDM [31] aims to predict motion directly and incorporates a geometric loss to boost the model's performance. Instead of employing a diffusion model to link raw motion sequences with conditional inputs, MLD [5] further utilizes the latent diffusion model to reduce the training and inference costs substantially. ReMoDiffuse [41] introduces an improvement mechanism based on dataset retrieval to enhance the denoising process of Diffusion.

Multi-Person Human Motion Generation. As multi-person motion synthesis involves the interactive dynamics between multiple individuals, it is more challenging than single-person motion generation. Early work typically relied on motion graphs and momentum-based inverse kinematics to model human joints. Recently, ComMDM [27] finetunes a pre-trained text-to-motion diffusion model using a small scale of two human motions. RIG [29] converts the text of asymmetric interactions into both active and passive voice to maintain consistent textual context for each individual. To complete the interaction process by

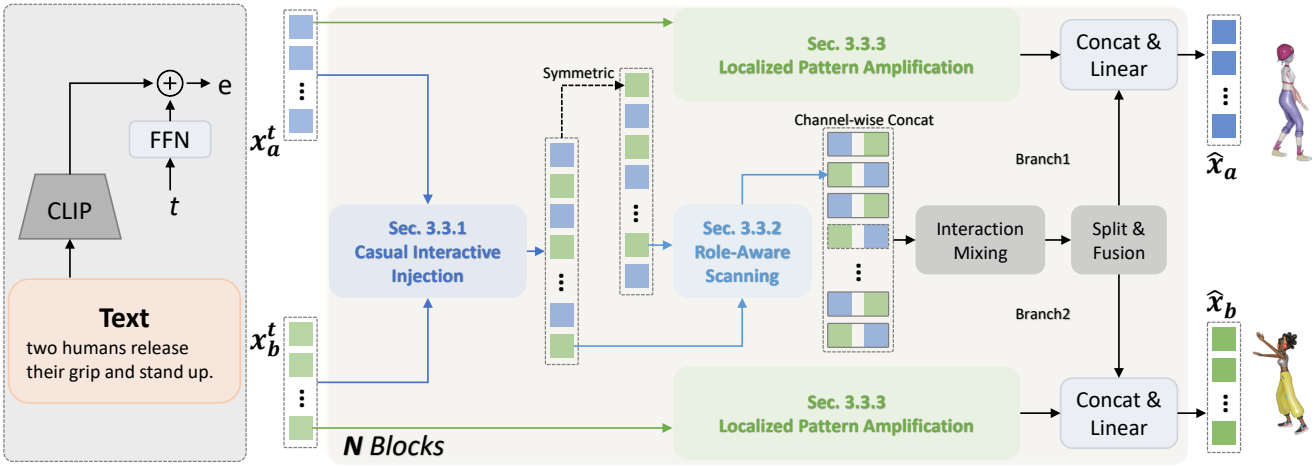

Figure 2. **The overall framework of our TIMotion.** We contribute three primary technical designs. First, we propose Causal Interactive Injection to utilize the temporal properties of motion sequences. Then we present Role-Evolving Mixing to adjust to the ever-evolving roles during interaction. Finally, we design Localized Pattern Amplification to capture short-term motion patterns.

manipulating the joints of two individuals to specific positions, InterControl [38] employs a Large Language Model to generate movement plans for the two individuals by designing prompts. InterGen [16] introduces a large-scale, text-annotated two-person motion dataset. Building on this dataset, it proposes a diffusion model with shared weights and multiple regularization losses. FreeMotion [6] proposes to decouple the process of conditional motion generation and support the number-free motion synthesis.

3. Method

3.1. Motion Diffusion Model

Current methods usually use diffusion models [15] for motion generation. For the Interhuman dataset, a non-canonical representation [16] is typically used. The representation is formulated as: $x^i = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f]$, where the i-th motion state x^i is defined as a collection of global joint positions $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$, velocities $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ in the world frame, 6D representation of local rotations $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ in the root frame, and binary foot-ground contact $\mathbf{c}^f \in \mathbb{R}^4$. Our goal is to train a model parameterized by θ to approximate the human interactive motion data distribution $p(\mathbf{x_0})$.

Diffusion Process. Following a specified schedule $\beta_t \in (0,1)$, models incrementally degrade input data $\mathbf{x}_0 \sim p(\mathbf{x}_0)$, ultimately transforming the data distribution into an isotropic Gaussian over T steps. Each diffusion transition can be considered as

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \quad (1)$$

where the full diffusion process can be written as

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) = \prod_{1 \le t \le T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right). \quad (2)$$

Denoising Process. In the denoising process, models are trained to reverse the diffusion procedure, enabling them to transform random noise into real data distribution during inference. The denoising process can be written as

$$\begin{aligned} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) &= \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right) \\ &= \mathcal{N}\left(\mathbf{x}_{t-1}; \tfrac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \tfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right), \tfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\right), \end{aligned} \quad (3)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and θ denotes parameters of the network learning to denoise. The training objective is to maximize the likelihood of observed data $p_\theta\left(\mathbf{x}_0\right) = \int p_\theta\left(\mathbf{x}_{0:T}\right) d\mathbf{x}_{1:T}$, by maximizing its evidence lower bound (ELBO), which effectively aligns the true denoising model $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ with the parameterized $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$. During training, the goal of the denoising network $\epsilon_\theta(.)$ is to reconstruct $\mathbf{x}_0$ given any noised input $\mathbf{x}_t$, by predicting the added noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via minimizing the noise prediction error

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right]. \quad (4)$$

To condition the model on additional context information $\mathbf{c}$, *e.g.*, text, we inject $\mathbf{c}$ into $\epsilon_\theta(.)$ by replacing $\mu_\theta\left(\mathbf{x}_t, t\right)$ and $\Sigma_\theta\left(\mathbf{x}_t, t\right)$ with $\mu_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right)$ and $\Sigma_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right)$.

3.2. MetaMotion

We present the core concept "MetaMotion" for human-human motion generation at first. As shown in Fig. 1,

we abstract the process of human-human motion generation into two phases: temporal modeling and interaction mixing.

Specifically, two single-person sequences go through the temporal modeling module to get the input sequence X. Then, the input sequence is fed to the interaction-mixing module and this process can be expressed as

$$Y = \text{InteractionMixing}(X), \tag{5}$$

where InteractionMixing is usually the transformer structure including self-attention and cross-attention. Notably, InteractionMixing can also be some emerging structures, *e.g.*, Mamba [8], RWKV [21].

3.3. TIMotion

For human-human motion generation, most current methods are based on extensions of single-person motion generation. Some design-specific approach (*e.g.* InterGen [16]) considers the interaction between two persons in the interaction mixing module. However, current methods ignore the importance of temporal modeling, resulting in suboptimal performance in generating long-sequence motions and multi-person interactions.

In this work, we propose TIMotion, as illustrated in Fig. 2, an effective temporal modeling approach that can be applied to different interaction-mixing structures, enabling better handling the human-human interactions. TIMotion consists of three key technical designs: (1) Causal Interactive Injection, which identifies the causal relationships between human-human interactions; (2) Role-Evolving Scanning, which adapts to shifts between active and passive roles during interactions; (3) Localized Pattern Amplification, designed to handle short-term sequences better.

3.3.1. Causal Interactive Injection

Perception of ego-motion and interaction between two persons are two important elements of human-human motion generation. Considering the causal properties of motion, we propose Causal Interactive Injection, a temporal modeling approach to simultaneously achieve both perceptions of ego-motion and interactions with each other.

Specifically, we denote two single-person motion sequences as $\{x_a,\ x_b\}$, where $x_a = \{x_a^j\}_{j=1}^{L}$ and $x_b = \{x_b^j\}_{j=1}^{L}$ are respective sequences of motion, and L is the length of the sequence. Since the motions of two individuals at the current time step are jointly determined by their motions at previous time steps, we model the two single-person motion sequences as a causal interaction sequence $x_{cii} = \{x_k^{j//2}\}_{j=1}^{2L}$, the symbol // denotes division followed by rounding up and k can be acquired as follows:

$$k = \begin{cases} a, & j \,\%\, 2 = 1. \\ b, & j \,\%\, 2 = 0. \end{cases} \tag{6}$$

Then we can inject them into the interaction-mixing module and separate the motion embeddings of the two individuals from the outputs according to Eq. (6).

Figure 3. Illustration of changing active and passive roles. The first person acts as the active role in the early stages, and as time progresses, the other person becomes the active role of the motion.

3.3.2. Role-Evolving Scanning

As noted in [4], human interactions inherently follow a certain order, *e.g.*, "shaking hands" typically begins with one person extending a hand first, which means that interactive motion can be categorized as active or passive. Some methods [29] split the text description into active and passive voices. However, the "active" and "passive" roles are not static in a text description. As shown in Fig. 3, these roles constantly swap between characters as the interaction progresses. To minimize redundant text preprocessing and adapt to the ongoing shifts in roles, we design Role-Evolving Scanning, which is both efficient and effective.

For the causal interaction sequence x defined in Causal Interactive Injection (Sec. 3.3.1), it is obvious that x_a and x_b represent the active sequence and passive sequence, respectively. However, the above assumption of active and passive motions is not always in line with the real order. To cope with the changing roles, we re-model the interactive motion sequences as a symmetric causal interaction sequence $x_{sym\_cii} = \{x_{k'}^{j//2}\}_{j=1}^{2L}$, where the symbol // is defined as division followed by rounding up and k' is obtained by exchanging a and b in Eq. (6).

Given that the causal interaction sequence $x_{cii} \in \mathbb{R}^{2L\times C}$ and the symmetric causal interaction sequence $x_{sym\_cii} \in \mathbb{R}^{2L\times C}$, where L is the length of the sequence and C is the dimension of motion embedding, we can acquire the final interaction sequence $X \in \mathbb{R}^{2L\times 2C}$ through Role-Evolving Scanning as:

$$X = \text{Concat}(x_{cii},\ x_{sym\_cii}). \tag{7}$$

Then the sequence X is fed to the interaction-mixing module. After obtaining the output $Y \in \mathbb{R}^{2L\times 2C}$ through Eq. (5), we first divide it at the channel level into causal interaction embeddings and symmetric causal interaction embeddings. According to x_{cii} andx_{sym_cii}, we can obtain the two individuals' splitting embeddings twice. Then we merge the embeddings of the two individuals to obtain the final global motion embeddings $y_a^g \in \mathbb{R}^{L\times C}$ and $y_b^g \in \mathbb{R}^{L\times C}$. The overall process is as follows:

$$\begin{aligned} y_{a1},\ y_{b1} &= \text{Split}(Y[:,\ :C]), \\ y_{b2},\ y_{a2} &= \text{Split}(Y[:,\ C:]), \\ y_a^g &= y_{a1} \oplus y_{a2}, \\ y_b^g &= y_{b1} \oplus y_{b2}, \end{aligned} \tag{8}$$

where $\oplus$ denotes element-wise sum.

By utilizing Role-Evolving Mixing to make two humans act as both active and passive roles, the network can dynamically adjust the roles of the two humans based on the semantics of the text and the context of the motion.

3.3.3. Localized Pattern Amplification

Transformers and RNNs excel at global modeling and capturing long-range dependencies but tend to overlook local semantic information [13]. Additionally, Causal Interactive Injection and Role-Evolving Mixing mainly model the overall motion based on the causality of the interactions, but neglect to focus on the localized motion patterns of the single person. To address this issue, we propose Localized Pattern Amplification, which summarizes short-term motion patterns for each person individually while generating smoother and more rational motion.

Specifically, we utilize 1-D convolution layers and the residual structure to realize Localized Pattern Amplification. Given that the condition embedding e and two single-person motion sequences x_a and x_b, the local motion embedding for x_a can be expressed as:

$$\begin{aligned} x_a^l &= \text{Conv}_3(\text{AdaLN}(x_a,\ e)), \\ y_a &= \text{Conv}_1(\text{AdaLN}(x_a^l,\ e)), \\ y_a^l &= x_a + y_a, \end{aligned} \tag{9}$$

where Conv_k denotes the 1-D convolution with kernel size k and AdaLN is the adaptive layer normalization used in [42]. The process for x_b is the same as for x_a and they share network weights.

Once we obtain the outputs $\{y_a^g,\ y_b^g\}$ through Eq. (8) and the output $\{y_a^l,\ y_b^l\}$ from the convolution block, global embeddings and local embeddings are aggregated through concatenation along the channel dimension, followed by a linear layer to restore the original channels, resulting in the final outputs $\{y_a^{final},\ y_b^{final}\}$:

$$\begin{aligned} y_a^{final} &= \text{Linear}(\text{Concat}(y_a^g,\ y_a^l)), \\ y_b^{final} &= \text{Linear}(\text{Concat}(y_b^g,\ y_b^l)). \end{aligned} \tag{10}$$

Then the final outputs are fed into the next encoder block or the final decoder layer.

3.3.4. Objective Function

We use the same loss functions as InterGen [16], including the diffusion loss $\mathcal{L}_{simple}$, the foot contact loss $\mathcal{L}_{foot}$, joint velocity loss $\mathcal{L}_{vel}$, the bone length loss $\mathcal{L}_{BL}$, the masked joint distance map loss $\mathcal{L}_{DM}$, and the relative orientation loss $\mathcal{L}_{RO}$. For more details about the losses, we refer readers to InterGen. Finally, the overall loss is defined as:

$$\begin{aligned} \mathcal{L}_{motion} = \mathcal{L}_{simple} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{foot}\mathcal{L}_{foot} \\ +\lambda_{BL}\mathcal{L}_{BL} + \lambda_{DM}\mathcal{L}_{DM} + \lambda_{RO}\mathcal{L}_{RO}, \end{aligned} \tag{11}$$

where the hyper-parameters $\lambda_{vel}, \lambda_{foot}, \lambda_{BL}, \lambda_{DM}, \lambda_{RO}$ are the same as InterGen.

4. Experiments

4.1. Experimental Setup

Datasets. We assess our proposed framework using the InterHuman [16] and Inter-X [39] dataset. InterHuman is the first dataset to incorporate text annotations for two-person motions. This dataset includes 6,022 motions spanning various categories of human actions and is labeled with 16,756 unique descriptions made up of 5,656 distinct words. Inter-X is the largest human-human interaction dataset with diverse interaction patterns. It includes about 11k interaction sequences and more than 8.1M frames.

Metrics. We employ the same evaluation metrics as InterGen [16], which are as follows: (1) *Frechet Inception Distance (FID)*. (2) *R-Precision*. (3) *Diversity*. (4) *Multimodality (MModality)*. (5) *Multi-modal distance (MM Dist)*. Detailed explanations and calculations of the metrics can be found in the **supplementary materials**.

4.1.1. Implementation Details.

We use a frozen CLIP-ViT-L/14 model as the text encoder. The dimension of motion embedding is set to 512. During training, the number of diffusion timesteps is set to 1000, and we employ the DDIM [28] sampling strategy with 50 timesteps and $\eta = 0$. The cosine noise level schedule [19] and classifier-free guidance [14] are adopted, with 10% of random CLIP embeddings set to zero during training and a guidance coefficient of 3.5 during sampling. All the models are trained using the AdamW [17] optimizer with betas of (0.9, 0.999), a weight decay of 2×10^{-5}, a maximum learning rate of 10^{-4}, and a cosine learning rate schedule with 10 linear warm-up epochs. To balance the contributions of different loss terms, we set $\lambda_{vel} = 30$, $\lambda_{foot} = 30$, $\lambda_{BL} = 10$,

Methods	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
TEMOS [23]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$5.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M [10]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$4.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM [31]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$6.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.355}^{\pm.080}$
ComMDM* [27]	$0.067^{\pm.013}$	$0.125^{\pm.018}$	$0.184^{\pm.015}$	$38.643^{\pm.098}$	$13.211^{\pm.013}$	$3.520^{\pm.058}$	$0.217^{\pm.018}$
ComMDM [27]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$5.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
RIG [29]	$0.285^{\pm.010}$	$0.409^{\pm.014}$	$0.521^{\pm.013}$	$6.775^{\pm.069}$	$4.876^{\pm.018}$	$7.311^{\pm.043}$	$2.096^{\pm.065}$
InterGen [16]	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$\underline{2.141}^{\pm.063}$
TIMotion+Transformer(ours)	$\underline{0.491}^{\pm.005}$	$\underline{0.648}^{\pm.004}$	$\underline{0.724}^{\pm.004}$	$\underline{5.433}^{\pm.080}$	$\underline{3.775}^{\pm.001}$	$\underline{8.032}^{\pm.030}$	$0.952^{\pm.032}$
TIMotion+Mamba(ours)	$0.431^{\pm.004}$	$0.586^{\pm.005}$	$0.668^{\pm.004}$	$6.142^{\pm.059}$	$3.800^{\pm.001}$	$7.793^{\pm.024}$	$0.837^{\pm.036}$
TIMotion+RWKV(ours)	$\mathbf{0.501}^{\pm.005}$	$\mathbf{0.656}^{\pm.006}$	$\mathbf{0.734}^{\pm.006}$	$\mathbf{4.702}^{\pm.069}$	$\mathbf{3.769}^{\pm.001}$	$\mathbf{7.943}^{\pm.034}$	$1.005^{\pm.020}$

Table 1. **Quantitative evaluation on the InterHuman [16] test set.** We run all the evaluations 20 times. ± indicates a 95% confidence interval. **Bold** indicates the best result, while underline refers to the second best. ComMDM* indicates the ComMDM model fine-tuned in the original few-shot setting with 10 training samples and ComMDM indicates fine-tuned on the entire InterHuman training set.

Interaction Mixing	Temporal Modeling	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
		Top 1	Top 2	Top 3				
Transformer [35]	single-person extension	$0.395^{\pm.004}$	$0.535^{\pm.004}$	$0.615^{\pm.004}$	$8.028^{\pm.099}$	$3.820^{\pm.001}$	$7.687^{\pm.022}$	$0.852^{\pm.024}$
	separate modeling	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$\mathbf{2.141}^{\pm.063}$
	TIMotion (Ours)	$\mathbf{0.491}^{\pm.005}$	$\mathbf{0.648}^{\pm.004}$	$\mathbf{0.724}^{\pm.004}$	$\mathbf{5.433}^{\pm.080}$	$\mathbf{3.775}^{\pm.001}$	$\mathbf{8.032}^{\pm.030}$	$0.952^{\pm.032}$
Mamba [8]	single-person extension	$0.368^{\pm.005}$	$0.513^{\pm.005}$	$0.597^{\pm.006}$	$7.232^{\pm.081}$	$3.825^{\pm.001}$	$7.864^{\pm.024}$	$\mathbf{0.870}^{\pm.020}$
	separate modeling	$0.420^{\pm.005}$	$0.569^{\pm.004}$	$0.650^{\pm.003}$	$7.221^{\pm.097}$	$3.803^{\pm.001}$	$\mathbf{7.932}^{\pm.035}$	$0.855^{\pm.020}$
	TIMotion (Ours)	$\mathbf{0.431}^{\pm.004}$	$\mathbf{0.586}^{\pm.005}$	$\mathbf{0.668}^{\pm.004}$	$\mathbf{6.142}^{\pm.059}$	$\mathbf{3.800}^{\pm.001}$	$7.793^{\pm.024}$	$0.837^{\pm.036}$
RWKV [21]	single-person extension	$0.425^{\pm.004}$	$0.576^{\pm.004}$	$0.656^{\pm.003}$	$9.181^{\pm.096}$	$3.801^{\pm.001}$	$7.679^{\pm.022}$	$0.846^{\pm.026}$
	separate modeling	$0.465^{\pm.005}$	$0.603^{\pm.005}$	$0.689^{\pm.004}$	$5.943^{\pm.079}$	$3.790^{\pm.001}$	$7.787^{\pm.032}$	$0.859^{\pm.036}$
	TIMotion (Ours)	$\mathbf{0.501}^{\pm.005}$	$\mathbf{0.656}^{\pm.006}$	$\mathbf{0.734}^{\pm.006}$	$\mathbf{4.702}^{\pm.069}$	$\mathbf{3.769}^{\pm.001}$	$\mathbf{7.943}^{\pm.034}$	$\mathbf{1.005}^{\pm.020}$

Table 2. **Comparison of different temporal modeling approaches on different interaction mixing structures.** Our proposed causal interactive modeling is able to adapt to different interaction mixing architectures and outperforms the other two ways.

$\lambda_{DM} = 3$, $\lambda_{RO} = 0.01$ and $\lambda_{reg} = 1$ in all experiments, as same as InterGen. We train our diffusion denoisers with a batch size of 256 for 1500 epochs on 8 Nvidia L40S GPUs.

4.2. Comparisons with State-of-the-arts

4.2.1. Quantitative Results.

Following established practices [16], each experiment is conducted 20 times, and the reported metric values represent the mean with a 95% statistical confidence interval. The results on InterHuman are shown in Tab. 1. In comparison to existing methods, our TIMotion achieves competitive results on three different interaction mixing architectures: transformer, mamba, and RWKV. Equipped with RWKV, TIMotion achieves 4.702 FID and 0.501 Top1 R precision, setting new state-of-the-art (SoTA) for the competitive InterHuman benchmark. The results on Inter-X [39] can be found in the **supplementary materials**.

4.2.2. Qualitative Comparisons.

In Fig. 4, we qualitatively compare the InterGen and our TIMotion. It can be seen that sequences generated by TIMotion are more consistent with the description.

4.3. Effectiveness of TIMotion

To validate the effectiveness of our proposed causal interactive modeling (TIMotion), we compare different temporal modeling approaches on three different structures, respectively. As shown in Tab. 2, single-person extension performs poorly, indicating that it does not model two-person interactions well. In contrast, separate modeling leverages the ability of the interaction mixing module to model the two-person interactions but achieves sub-optimal performance on the Transformer and mamba architectures. Our proposed approach is able to adapt to different interaction mixing architectures and outperforms the other two approaches.

4.4. Ablation Study and Analysis

In this section we conduct comprehensive ablation studies to investigate the effectiveness of the key components in TIMotion, thereby providing a deeper insight into our approach. All experiments are performed on the InterHuman dataset. Unless otherwise stated, we use RWKV as the interaction mixing structure for the following ablations. More theoretical analyses are in the **supplementary materials**.

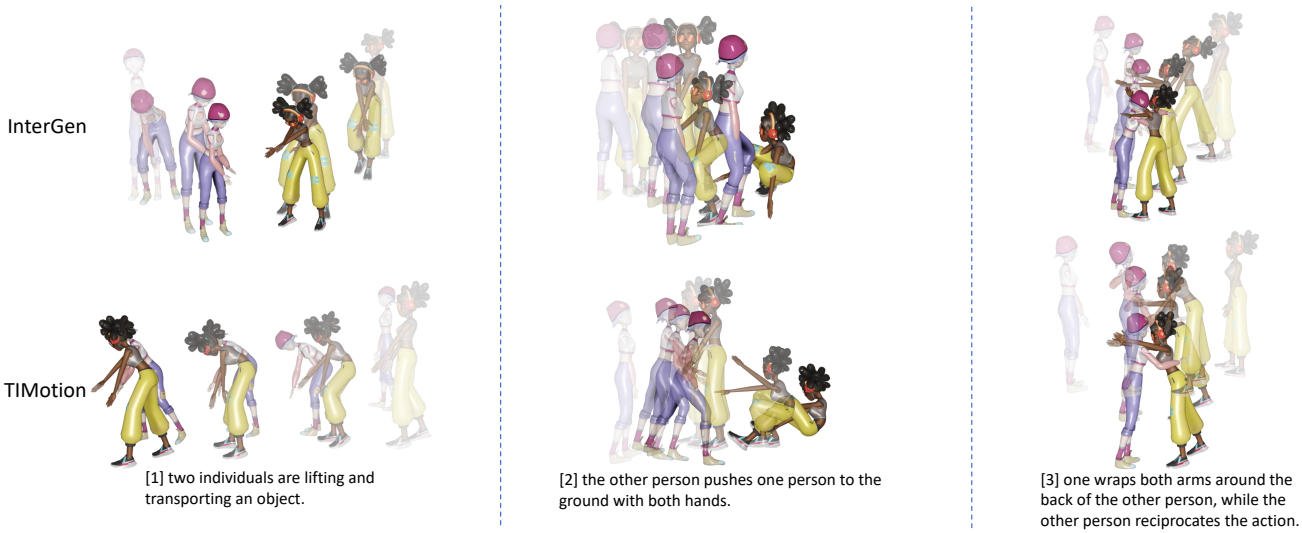

Figure 4. Qualitative comparison with Intergen on human-human motion generation. Darker color indicates later frames. The sequences generated by TIMotion are more consistent with the text description.

CII	RES	LPA	R Precision Top 1↑	FID ↓	Params (M)
			$0.371^{\pm.010}$	$5.918^{\pm.079}$	182
✔			$0.478^{\pm.005}$	$5.410^{\pm.069}$	144
✔	✔		$0.494^{\pm.006}$	$5.019^{\pm.077}$	115
✔	✔	✔	$\mathbf{0.501}^{\pm.005}$	$\mathbf{4.702}^{\pm.069}$	127

Table 3. **Ablation studies on the effectiveness of each component in TIMotion.** "CII" denotes Causal Interactive Injection, "RES" denotes Role-Evolving Scanning, and "LPA" denotes Localized Pattern Amplification.

Kernel Size	Norm	R Precision Top 1↑	FID ↓
k=3,1	BN	$0.494^{\pm.005}$	$5.443^{\pm.093}$
k=3,1	LN	$0.488^{\pm.005}$	$5.339^{\pm.080}$
k=3,1	AdaLN	$\mathbf{0.501}^{\pm.005}$	$\mathbf{4.702}^{\pm.069}$
k=3,3	AdaLN	$0.484^{\pm.005}$	$7.120^{\pm.090}$
k=5,1	AdaLN	$0.497^{\pm.005}$	$5.102^{\pm.064}$

Table 4. **Ablation studies on LPA.** "BN" denotes batch normalization, "LN" denotes layer normalization and "AdaLN" denotes adaptive layer normalization. "k=3,1" means that the first kernel size of the convolution is 3 and the second kernel size is 1.

4.4.1. Main Ablations

To understand the contribution of each component to the final performance, we incrementally add the proposed modules based on the baseline InterGen [16] and present the results in Tab. 3. Initially, we replace self-attention and cross-attention in the transformer with Causal Interactive Injection (CII), which effectively improves both R-Precision and FID while reducing the number of parameters. Next, we apply Role-Evolving Scanning (RES) and yield gains in both R-Precision and FID. It is worth noting that to keep the input feature dimension of RWKV constant, we reduce the dimension of the motion embedding to half of the original, thus also effectively reducing the number of parameters. Finally, when all three methods are applied together, the R-Precision achieves 0.501 and FID achieves 4.702.

4.4.2. Design of Localized Pattern Amplification.

In this section, we conduct ablations on the design of Localized Pattern Amplification (LPA), and the results are shown in Tab. 4. First, we explore the effect of different normalizations on model performance. Using batch normalization (BN) and layer normalization (LN) does not bring gains for text-to-motion tasks. In contrast, AdaLN, which integrates text information into the extraction process of the local motion pattern, proves to be the optimal normalization method. Additionally, we examine how the kernel size of the convolutional layers impacts the model's performance and select the setting of k=3,1.

To further demonstrate the effectiveness of LPA, we visualize the spectrum of TIMotion w/o and w/ LPA in Fig. 6, respectively. The magnitude has been normalized. As can be seen in Fig. 6, adding LPA reduces the high-frequency component of the feature, making the motion less likely to change drastically and making it smoother. Moreover, we randomly sampled 200 instances, and w/ LPA, the average proportion of the amplitude of high-frequency components is 0.3729, whereas w/o LPA, the average proportion of the amplitude of high-frequency components is 0.9063.

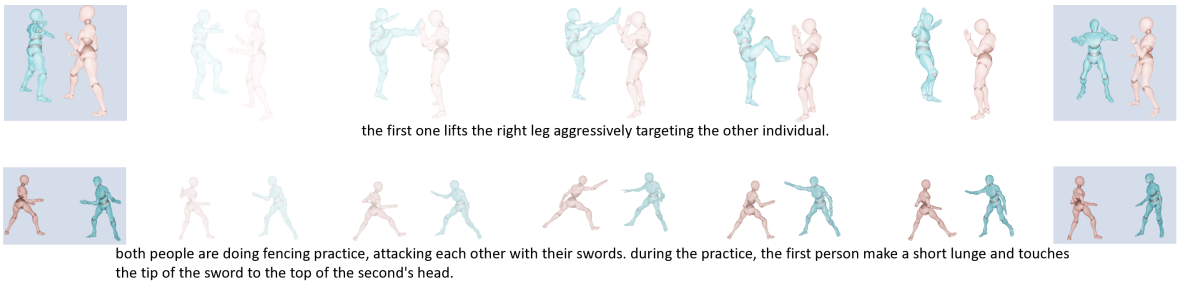

Figure 5. **Qualitative results on the motion in-betweening task.** The first and last frames are fixed. Darker colors indicate later frames. Our method achieves smooth and natural transitions between the conditioned motions.

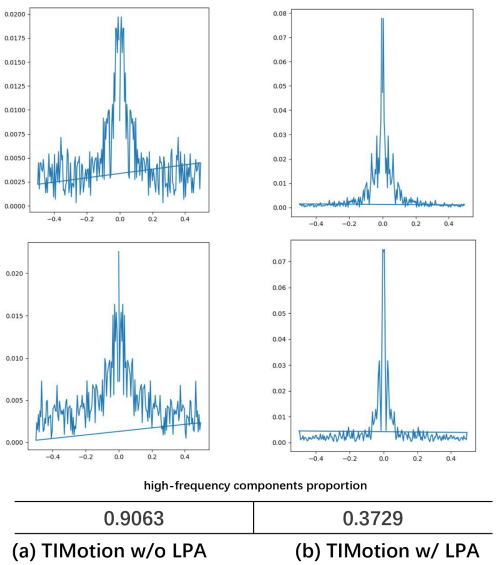

Figure 6. **Spectrum of motion features.** (a) and (b) show the spectrum of TIMotion w/o and w/ LPA, respectively. The horizontal axis denotes the frequency and the vertical axis represents the normalized magnitude. TIMotion w/ LPA contains fewer high-frequency components and therefore generates smoother motion.

4.5. Computational Complexity and Editability.

4.5.1. Computational Complexity.

In Tab. 5, we compare our approach TIMotion with the SOTA method InterGen [16] in terms of computational complexity. TIMotion requires fewer parameters and FLOPs than InterGen but outperforms it on the comprehensive metric FID and R precision. Notably, using a similar transformer architecture to InterGen, TIMotion's average inference time per sample is only 0.632 s while InterGen requires 1.991 s.

4.5.2. Editability.

Tab. 6 illustrates that TIMotion surpasses InterGen in the task of motion in-betweening editing. We perform experiments on the test set of InterHuman and evaluate by generating 80% of the sequences based on the first and last 10% of the sequences. The quantitative results are shown in Fig. 5 and our method achieves smooth and natural transitions between the conditioned motions.

Methods	R Precision Top 1↑	FID ↓	Params (M)	Flops (G)	Inference time (s)
InterGen (Transformer)	$0.371^{\pm.010}$	$5.918^{\pm.079}$	182	80.5	1.991
TIMotion+Transformer	$\underline{0.491}^{\pm.005}$	$\underline{5.433}^{\pm.080}$	**65**	**40.4**	**0.632**
TIMotion+RWKV	$\mathbf{0.501}^{\pm.005}$	$\mathbf{4.702}^{\pm.069}$	$\underline{127}$	$\underline{58.0}$	$\underline{1.733}$

Table 5. **Comparison of computational complexity.**

Methods	R Precision Top 1↑	FID ↓	MM Dist↓	Diversity→
InterGen	$0.461^{\pm.006}$	$4.700^{\pm.066}$	$3.780^{\pm.001}$	$7.682^{\pm.029}$
TIMotion	$\mathbf{0.516}^{\pm.006}$	$\mathbf{3.590}^{\pm.049}$	$\mathbf{3.760}^{\pm.001}$	$\mathbf{7.795}^{\pm.031}$

Table 6. Evaluation of motion in-betweening editing task.

5. Conclusion

In this paper, we abstract the overall human-human motion generation process into a general framework MetaMotion, which consists of two phases: temporal modeling and interaction mixing. We find that the current methods lead to sub-optimal results and redundancy of model parameters due to inadequate modeling. Based on this, we introduce TIMotion (Temporal and Interactive Modeling), an efficient and effective approach for human-human motion generation. Specifically, we first propose Causal Interactive Injection to model two separate sequences as a causal sequence leveraging the temporal and causal properties. Additionally, we proposed Role-Evolving Mixing to adapt to the dynamic roles throughout interactions and designed Localized Pattern Amplification to capture short-term motion patterns for generating smoother and more rational motion. Extensive experiments on the InterHuman and Inter-X datasets demonstrate that TIMotion significantly outperforms existing methods, achieving state-of-the-art results.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 1

[3] Erik Bethke. *Game development and production*. Wordware Publishing, Inc., 2003. 1

[4] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi 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*, pages 582–592, 2024. 4

[5] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[6] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 3

[7] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 2

[8] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 4, 6

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 6

[11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1

[12] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 1

[13] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564*, 2024. 5

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[16] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 3, 4, 5, 6, 7, 8

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[18] Nadia Magnenat-Thalmann, Daniel Thalmann, Nadia Magnenat-Thalmann, and Daniel Thalmann. *Computer animation*. Springer, 1985. 1

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 5

[20] Rick Parent. *Computer animation: algorithms and techniques*. Newnes, 2012. 1

[21] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023. 4, 6

[22] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 1

[23] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2, 6

[24] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 1

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[26] George Saridis. Intelligent robotic control. *IEEE Transactions on Automatic Control*, 28(5):547–557, 1983. 1

[27] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2, 6

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[29] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15999–16009, 2023. 2, 4, 6

[30] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[31] G Tevet, S Raab, B Gordon, Y Shafir, D Cohen-Or, and AH Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 6

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1

[33] Jay Urbain. Introduction to game development. *Cell*, 414: 745–5102, 2010. 1

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 6

[36] Shuo Wang, Xinhai Zhao, Hai-Ming Xu, Zehui Chen, Dameng Yu, Jiahao Chang, Zhen Yang, and Feng Zhao. Towards domain generalization for multi-view 3d object detection in bird-eye-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13333–13342, 2023. 1

[37] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao. Stream query denoising for vectorized hd-map construction. In *European Conference on Computer Vision*, pages 203–220. Springer, 2024. 1

[38] Zhenzhi Wang, Jingbo Wang, Dahua Lin, and Bo Dai. Intercontrol: Generate human motion interactions by controlling every joint. *arXiv preprint arXiv:2311.15864*, 2023. 3

[39] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 5, 6

[40] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[41] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2

[42] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 5

[43] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024. 1

[44] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 2