InterMoE: Individual-Specific 3D Human Interaction Generation via Dynamic Temporal-Selective MoE

Lipeng Wang[1], Hongxing Fan[2], Haohua Chen[1], Zehuan Huang[1], Lu Sheng[1]*

[1]School of Software, Beihang University
[2]School of Computer Science and Engineering, Beihang University
wanglipeng@buaa.edu.cn

Abstract

Generating high-quality human interactions holds significant value for applications like virtual reality and robotics. However, existing methods often fail to preserve unique individual characteristics or fully adhere to textual descriptions. To address these challenges, we introduce InterMoE, a novel framework built on a Dynamic Temporal-Selective Mixture of Experts. The core of InterMoE is a routing mechanism that synergistically uses both high-level text semantics and low-level motion context to dispatch temporal motion features to specialized experts. This allows experts to dynamically determine the selection capacity and focus on critical temporal features, thereby preserving specific individual characteristic identities while ensuring high semantic fidelity. Extensive experiments show that InterMoE achieves state-of-the-art performance in individual-specific high-fidelity 3D human interaction generation, reducing FID scores by 9% on the InterHuman dataset and 22% on InterX.

Code — https://github.com/Lighten001/InterMoE
Extended version — https://arxiv.org/abs/2511.13488

1 Introduction

The generation of realistic and expressive Human Interaction has emerged as an important research area, propelled by rapid advancements in motion synthesis techniques. Faithfully modeling the complex joint motion between two individuals is crucial for a multitude of downstream applications, including but not limited to computer animation, virtual reality, and game development.

Recent works have made significant progress in generative human interaction. Some work (Liang et al. 2024; Javed, Li et al. 2025) uses cross-attention to fuse features between interacting individuals. However, the subsequent uniform processing of these fused features by standard feed-forward networks (FFNs) tends to suppress individual characteristics, resulting in homogenized motions. An alternative approach (Li et al. 2024; Wang et al. 2025) concatenates individual features to jointly generate motion for both individuals; however, this method often suffers from identity confusion due to the absence of explicit identity constraints, which can cause characters to swap roles or positions during interaction erroneously.

To address the core challenge of preserving distinct identities, we argue that the inherent complexity of simultaneously modeling individual-specific characteristics and the joint motion between persons necessitates a modular approach handled by specialized sub-modules. We identify the Mixture of Experts (MoE) architecture (Shazeer et al. 2017; Lepikhin et al. 2020) as a promising paradigm for this purpose. By design, MoE enables differentiated expert allocation by routing inputs based on their distinct motion characteristics, allowing for the development of specialists for each individual's unique motion patterns. This approach naturally mitigates identity confusion and homogenization. Prevailing MoE routing strategies fall into two main categories. In the Token-Choice paradigm (Fei et al. 2024), each token selects a fixed number of experts for processing. However, this uniform assignment fails to account for the varying importance across temporal features. The second category is Expert-Choice (Sun et al. 2024), where experts select a fixed number of the most salient tokens. Yet, this fixed-capacity approach can limit expert utility, especially when modeling complex interactions.

In this work, we propose InterMoE, a novel framework that introduces a Dynamic Temporal-Selective MoE to generate high-fidelity, individual-specific 3D human interactions. Our framework is built upon two key innovations: a Synergistic Router that directs information based on both semantic and kinematic cues, and a Dynamic Temporal-Selection mechanism that empowers experts to focus on critical temporal features. Specifically, the Synergistic Router leverages both high-level semantics from text and low-level kinematic features to guide routing decisions. This dual guidance ensures information is dispatched to the most appropriate experts, strengthening the alignment between the textual description and the generated motion. Building on this, the Dynamic Temporal-Selection mechanism enables each expert to dynamically determine the selection capacity and proactively identify the most salient temporal features, which effectively addresses the non-uniform temporal importance of the interactive motion sequence.

In summary, our contributions are as follows:

- We propose InterMoE, a novel diffusion-based MoE framework for text-driven 3D human interaction gen-

*Corresponding author.

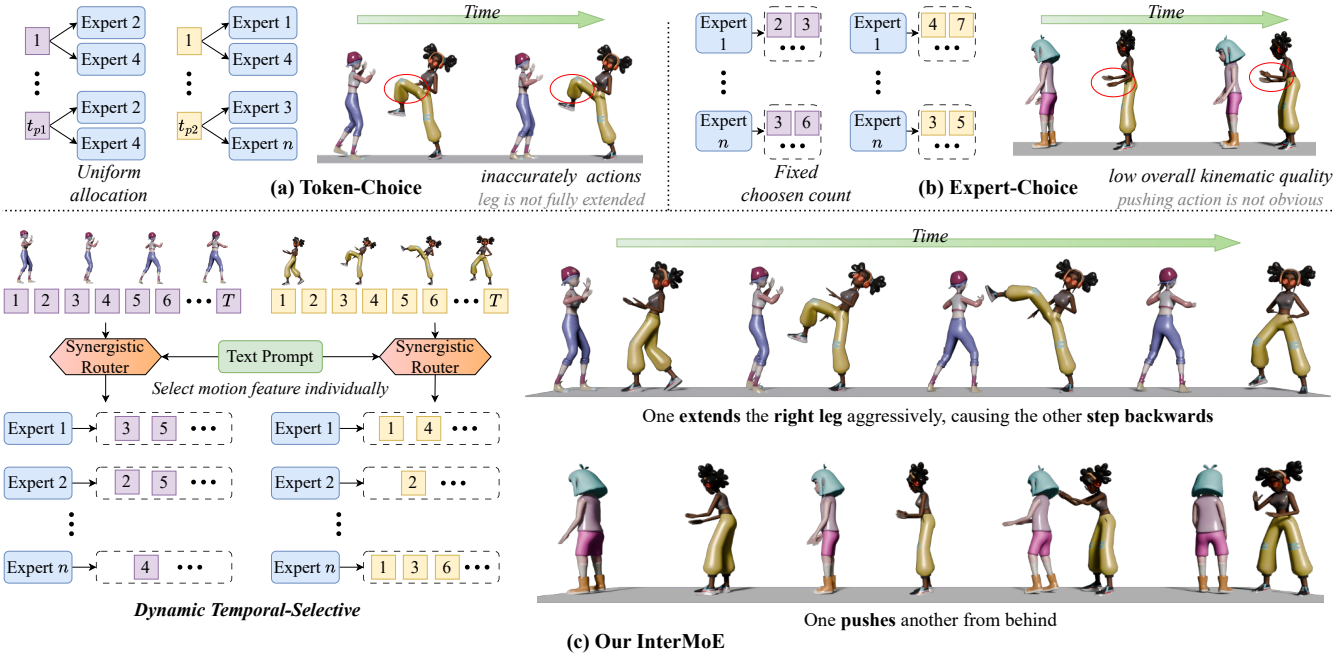

Figure 1: Compared with conventional MoE mechanisms, Token-Choice inaccurately generates the "*extends*" action, and Expert-Choice has low overall kinematic quality. Our framework leverages the Synergistic Router and Dynamic Temporal Selection mechanism to generate 3D human interactions that exhibit both high semantic fidelity and robust preservation of individual-specific characteristics.

eration, achieving notable improvements in individual-specific characteristics, semantic fidelity, and overall quality of interaction.

- We introduce the Dynamic Temporal-Selective MoE, a new paradigm tailored for generating 3D human interactions. It leverages a Synergistic Router that leverages both semantic and kinematic features for precise routing, and a Dynamic Temporal-Selection mechanism that empowers experts to dynamically focus on critical temporal features across varying noise levels.
- We conduct extensive experiments to demonstrate the effectiveness of our proposed framework. Furthermore, competitive results on the single-human motion generation task validate its strong generalization.

2 Related Works

Human Motion Generation Synthesizing single-person motion has gained interest, driven by large motion capture datasets and advancements in generative modeling techniques like Diffusions (Tevet et al. 2023; Tseng, Castellon, and Liu 2023; Zhang et al. 2024a; Chen et al. 2023a; Kong et al. 2023; Lou et al. 2023; Zhang et al. 2024b) and Autoregressive models (Guo et al. 2022b; Zhang et al. 2024c; Jiang et al. 2023; Lucas et al. 2022; Gong et al. 2023; Pinyoanuntapong et al. 2024). Recent works have explored various motion representations. TM2T (Guo et al. 2022b) applies VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) to human motion data. MoMask (Guo et al. 2024) reduced quantization errors via residual quantization. While MotionStreamer (Xiao et al. 2025) introduces a causal convolution to enforce temporal causality. SALAD (Hong et al. 2025) utilizes skeletal graph convolution to capture the spatial structure of human movement. Our work is motivated by these prior studies.

Human Interaction Generation Compared to single-human motion generation, human interaction generation is more challenging, as it requires accurately modeling interactions between individuals. Recently, ComMDM (Shafir et al. 2024) trained a small neural network to bridge two single-person motion diffusion model copies on a limited interaction dataset. InterGen (Liang et al. 2024) introduced a large-scale text-annotated two-person interaction dataset. And it proposed an interaction diffusion model that simultaneously denoises both individuals' motions. The in2IN (Ruiz-Ponce et al. 2024) further advances the field by introducing a diffusion model that conditions motion generation not only on overall interaction descriptions but also on individual actions. InterMask (Javed, Li et al. 2025) employs a spatial-temporal transformer to autoregressively generate interactions. TIMotion (Wang et al. 2025) models temporal and interactive dynamics in interactions. Although the above-mentioned methods have achieved promising results, they still show limitations in differentiating the identity-specific characteristics of each individual and remain semantically faithful.

Mixture-of-Experts Mixture-of-Experts (Shazeer et al. 2017; Lepikhin et al. 2020) paradigm has become a powerful and efficient strategy for scaling models while main-

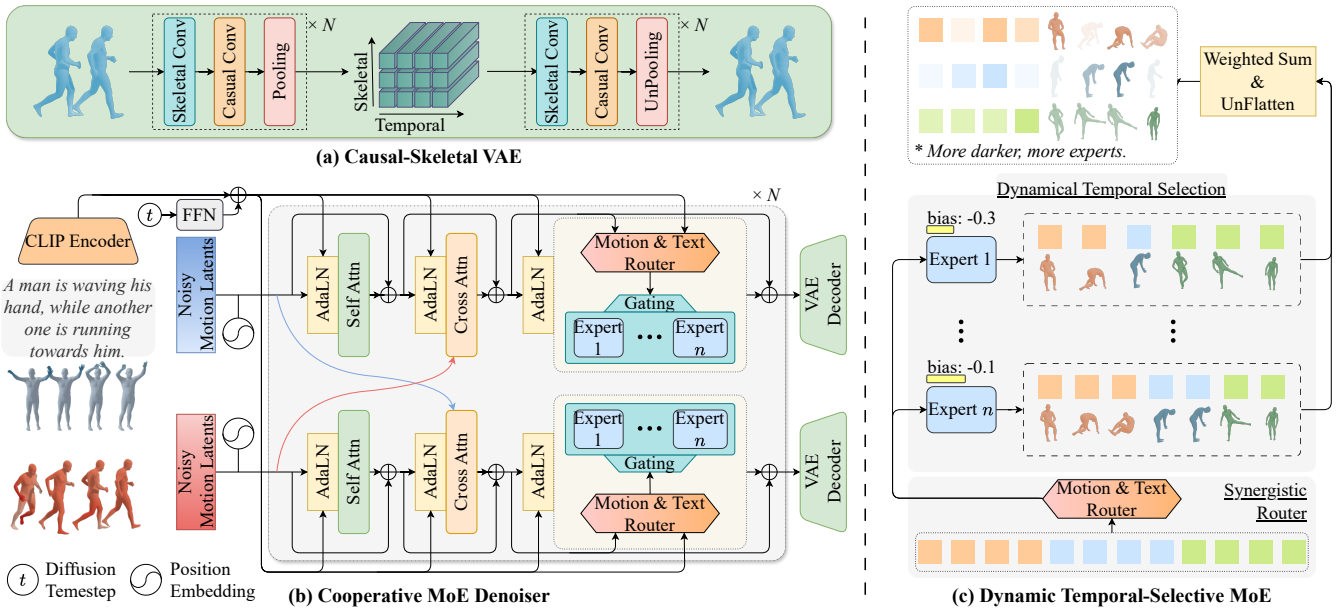

Figure 2: The overall framework of the InterMoE. (a) Causal-Skeletal VAE to encode/decode individual motions; (b) Two Cooperative MoE Denoisers to interactively perform denoising; (c) Our proposed Synergistic Router and Dynamic Temporal-Selective Expert mechanism. The router guides multiple experts to select and process critical temporal features of the motion sequence dynamically.

taining manageable computational costs by selectively activating expert subsets. This approach has made remarkable success in Large Language Models and multimodal large models, such as (DeepSeek-AI et al. 2024; MiniMax et al. 2025; Chen et al. 2023b; Muennighoff et al. 2024). Recent works have explored incorporating MoE architectures into diffusion models. MEME (Lee et al. 2023), eDiff-I (Balaji et al. 2022), and ERNIE-ViLG 2.0 (Feng et al. 2023) restrict expert selection to a specific timestep. SegMoE (Yatharth Gupta and Jaddipal 2024) and DiT-MoE (Fei et al. 2024) suffer from significant expert utilization imbalance due to isolated token processing. While EC-DiT (Sehwag et al. 2024; Sun et al. 2024) dispatch consequential tokens to each expert. Su (Su 2025) proposes a loss-free dynamical routing method. DiffMoE (Shi et al. 2025) utilizes a batch-level global token pool and dynamically adapting computation. Our work is informed by the foundation laid by these prior studies.

3 Methods

We target the task of text-driven synthesis of 3D human interaction. Given a textual description, our model generates a set of motion sequences $\mathbf{m}_i, i \in \{1, 2\}$, where $\mathbf{m}_i$ represents the 3D motion for an individual i. Each 3D motion $\mathbf{m}_i \in \mathbb{R}^{T \times J \times d}$ is composed of T frames, with each frame describing a pose via J joints, each represented by a d-dimensional feature vector. The overall pipeline of our framework is illustrated in Figure 2. In this section, we first briefly introduce the Causal-Skeletal VAE and the Cooperative MoE Denoiser used in the diffusion model in section 3.1. Then we give a detailed explanation of the core MoE design in section 3.2.

3.1 Interaction Latent Diffusion

Causal-Skeletal VAE Recent works have demonstrated the efficacy of applying the graph convolution to the human joint topology for extracting skeletal features (Hong et al. 2025), while the causal convolutions preserve temporal causality and enable efficient encoding of sequential data (Yu et al. 2023). Building upon these findings, we devise a hierarchical encoder-decoder architecture to embed the single-person motion. As shown in Figure 2(a), we first use skeletal convolutions to capture the complex intra-frame human joint dependencies. The resulting then fed into a causal convolution to model the inter-frame temporal dynamics and causal relationships. Finally, a pooling layer compresses these skeletal-temporal features into a compact representation. This design yields a lightweight yet highly efficient representation for our motion VAE.

Specifically, when feeding motion embeddings into the denoising network, we flatten the joint dimensions the same as InterGen (Liang et al. 2024) (i.e., reshape $\mathbf{m}_i^{\text{ori}} \in \mathbb{R}^{T \times J \times d}$ to $\mathbf{m}_i \in \mathbb{R}^{T \times D_m}$, where $D_m = J \times d$).

Cooperative MoE Denoiser The architecture of our denoising network inherits the design of InterGen (Liang et al. 2024). It utilizes two share-weight cooperative Denoisers to process the human interaction. Each denoiser is composed of a series of transformer blocks. As shown in Figure 2(b), each block contains three core components: (1) a Self-Attention Layer to model intra-individual temporal relationships; (2) a Cross-Attention Layer that conditions on the motion features of the interaction partner to model inter-individual dy-

namics; (3) and our proposed MoE Block. Furthermore, we integrate the denoising timestep and text-conditional information into the network via Adaptive Layer Normalization before all attention layers and the MoE Block.

3.2 InterMoE

Prior work (Wang et al. 2025) has shown that relying solely on cross-attention mechanisms is often insufficient for preserving distinct identities in human interaction synthesis. To overcome this limitation and to enhance the overall quality and semantic fidelity of generated sequences. Here, we introduce the core of InterMoE, which consists of two key components: the Synergistic Router and the Dynamic Temporal Selection. These components enable the generation of high-fidelity identity-preserving 3D human interaction.

Synergistic Router As shown in Figure 2(b) and 2(c), our synergistic router operates on the motion features $\mathbf{m}_i$ and the text condition c_t. We employ a motion router to generate routing logits based on the unique kinematic signatures of each individual. Concurrently, a parallel text router takes the text features as input and computes the semantic routing logits.

Furthermore, we identify that an instance-centric routing scope prevents routers from perceiving the heterogeneity of noise levels across different samples in a batch and hinders the discovery of global motion patterns. To overcome this, we introduce a batch-level routing strategy. Specifically, we flatten the input motion feature along its batch dimension (i.e., reshape $\mathbf{m}_i \in \mathbb{R}^{B\times T\times D_m}$ to $\mathbf{m}_i^{\text{flat}} \in \mathbb{R}^{S\times D_m}$, where $S = B \times T$) to create a batch-level temporal feature pool. Then calculate the routing logits $\mathbf{R}_e$ for each expert e. Note that $\mathbf{m}_i^{\text{flat}} = [m_{s,i}^{\text{flat}}],\ s \in \{1, \dots, S\}$, we have

$$\mathbf{R}_{e,s,i}^{\text{motion}} = \mathbf{Router}_e^{\text{motion}}(m_{s,i}^{\text{flat}}),\ m_{s,i}^{\text{flat}} \in \mathbb{R}^{1\times D_m} \tag{1}$$

$$\mathbf{R}_e^{\text{text}} = \mathbf{Router}_e^{\text{text}}(c_t),\ c_t \in \mathbb{R}^{1\times D_t} \tag{2}$$

These two sets of logits are subsequently fused via a weighted summation to produce the final logits $\mathbf{R}_{e,s,i}^{\text{comb}}$.

$$\mathbf{R}_{e,s,i}^{\text{comb}} = \alpha\mathbf{R}_{e,s,i}^{\text{motion}} + (1-\alpha)\mathbf{R}_e^{\text{text}} \tag{3}$$

In experiments, we set $\alpha = 0.5$. Through this design, routers are guided by both individual-specific dynamics and high-level semantic intent. By utilizing a batch-level temporal feature pool, the router can perform a more nuanced analysis of motion features and fully leverage the information inherent to different noise levels during allocation.

Dynamic Temporal Selection The conventional Token-Choice paradigm treats all tokens uniformly, which overlooks the non-uniform salience of temporal features in interactive motion sequences. To address this, we propose a Dynamic Temporal Selection mechanism, as shown in Figure 2(c). This mechanism empowers each expert to proactively select critical tokens from the entire batch-level temporal feature pool for processing. Furthermore, we remove constraints on each expert to a fixed capacity of selecting top-K features. Instead, we introduce a dynamic selection mechanism. Specifically, within each MoE module of our network, we associate every expert with a learnable bias parameter b_e. Notably, each Cooperative MoE Denoiser is dedicated to processing a single individual within the interaction, so omitting the annotation i for different individuals, we have:

$$\mathbf{M}_{e,s} = \texttt{sigmoid}(\mathbf{R}_{e,s}^{\text{comb}}) + b_e \tag{4}$$

$$\mathbf{A}_{e,s} = \texttt{softmax}(\mathbf{R}_{e,s}^{\text{comb}}) \tag{5}$$

$$\mathbf{G}_{e,s} = \begin{cases} \mathbf{A}_{e,s}, & \mathbf{M}_{e,s} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\mathbf{G}_{e,s}$ is the final gating weight, $\mathbf{M}_{e,s}$ determines to select this feature or not for experts. Since the `sigmoid` function maps values to $(0, 1)$, this bias b_e is constrained in $(-1, 0)$, which determines the number of features that the expert e will process, where a larger bias (i.e., a value closer to 0) corresponds to a higher capacity. Critically, these bias parameters b_e are optimized during training. With K_e^{exp} as a hyperparameter, in a training step, we count the number of selected features K_e^{select} for each expert:

$$K_e^{\text{select}} = \mathbf{Count}(\mathbf{M}_{e,s} > 0),\ s \in \{1, ..., S\}. \tag{7}$$

After this training step, we update bias b_e as:

$$\Delta b_e = \text{sign}(K_e^{\text{select}} - K_e^{\text{exp}}) \tag{8}$$

$$b_e \leftarrow b_e - \sigma\ \Delta b_e \tag{9}$$

In experiments, we set $\sigma = 1 \times e^{-4}$, As training converges, the Δb_e term is driven toward zero, ensuring that $\mathbf{E}(K_e^{\text{select}})$ approximates to K_e^{exp}, which facilitates a dynamic but stable allocation. Finally, we obtain the output feature m_s^{out} from the MoE block:

$$m_s^{\text{out}} = \sum_{e=1}^{N} \mathbf{G}_{e,s} E(m_s),\ m_s \in \mathbb{R}^{1\times D},\ s \in \{1, \dots, S\}. \tag{10}$$

This design enables each expert to dynamically determine the feature selection capacity and select their preferred temporal motions across all sequences within the entire batch. This global perspective yields a dual advantage: First, it endows the experts with noise-level awareness, enabling more robust feature selection at different denoising stages during inference. Second, it significantly enhances the experts' ability to identify universally critical temporal features through exposure to a more diverse set of motion samples.

4 Experiments

4.1 Experimental Setup

Datasets We adopt two datasets to evaluate our method for the text-conditioned human interaction generation task: InterHuman (Liang et al. 2024) and InterX (Xu et al. 2024). The InterHuman dataset contains 7,779 interaction sequences, and InterX contains 11,388, each paired with 3 distinct textual annotations. InterHuman follows the AMASS (Mahmood et al. 2019) skeleton representation with 22 joints, including the root joint. Each joint is represented by $\{pos, vec, rot\}$, where $pos \in \mathbb{R}^3$ is the global

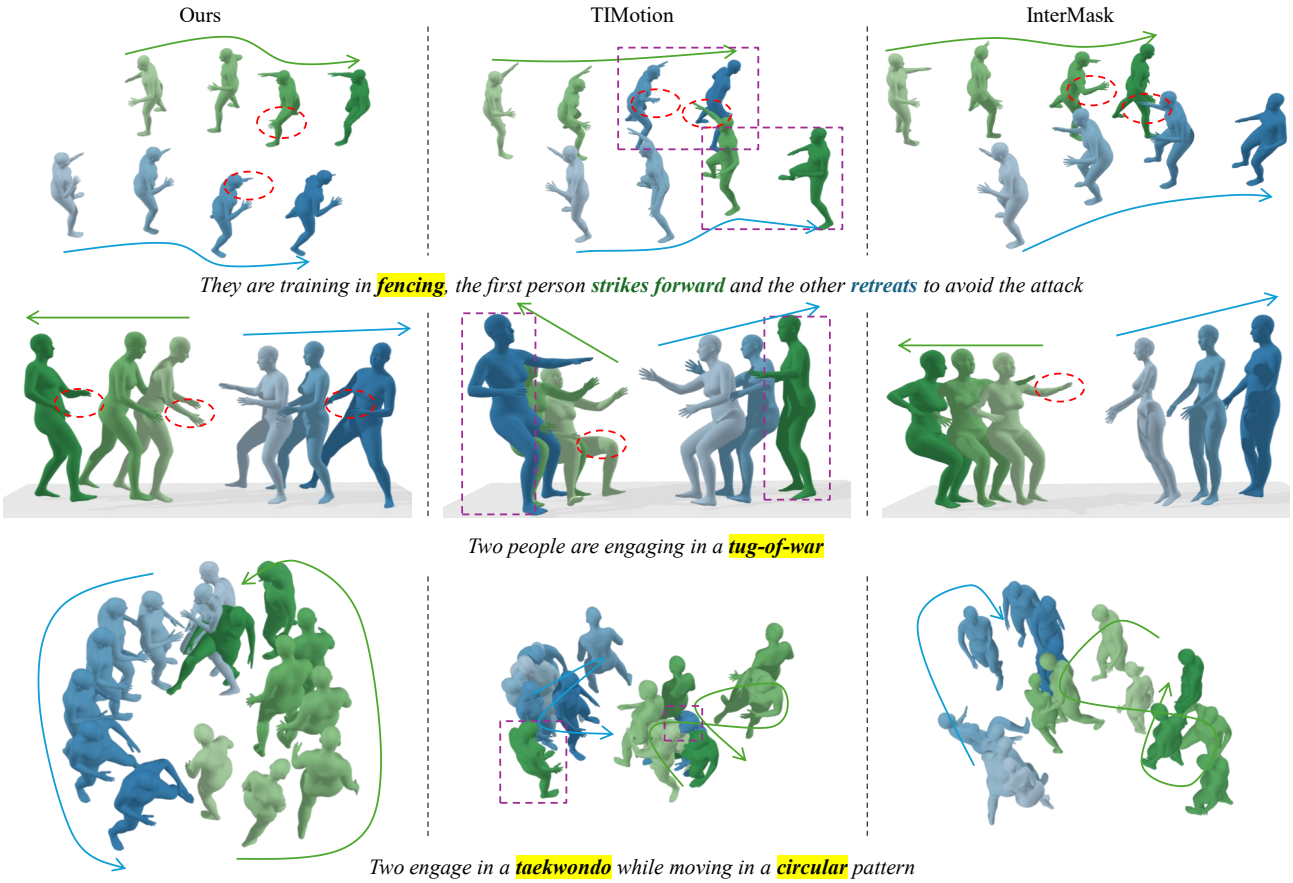

Figure 3: Qualitative comparisons with TIMotion (2025) and InterMask (2025). Arrowed lines mark the trajectories of motion, Red circles indicate key actions that align with the text, and Purple boxes highlight the identity confusion error.

position, $vec \in \mathbb{R}^3$ is the global velocity, and $rot \in \mathbb{R}^6$ is the local 6D rotation of each joint, rendering $\mathbf{m}_p \in \mathbb{R}^{N\times 22\times 12}$. InterX follows the SMPL-X (Pavlakos et al. 2019) skeleton representation, comprising 55 body, hands, and face joints. Each joint is represented by $\{pos, vec, rot\}$ where $pos \in \mathbb{R}^3$ is the global position, $vec \in \mathbb{R}^3$ is the global velocity, and $rot \in \mathbb{R}^6$ is the local 6D rotation of each joint, rendering $\mathbf{m}_p \in \mathbb{R}^{N\times 55\times 12}$.

Metrics We employ the same evaluation metrics as T2M (Guo et al. 2022a) and InterGen (Liang et al. 2024), which are as follows: (1) Frechet Inception Distance (FID). (2) R-Precision. (3) Diversity. (4) Multimodality (MModality). (5) Multi-modal distance (MM Dist).

Implementation Details Our framework was trained on two NVIDIA RTX3090 GPUs. We used the AdamW optimizer with betas of $(0.9, 0.999)$, a weight decay of 2×10^{-5}, a maximum learning rate of 1×10^{-4}, and a cosine LR schedule with 10 linear warm-up epochs. For InterHuman dataset, the VAE was trained for 100 epochs with batch size of 256, and the denoiser was trained for 1000 epochs with batch size of 64, respectively. For Inter-X dataset, the VAE was trained for 500 epochs with batch size of 256, and the denoiser was trained for 2000 epochs with batch size of 64, respectively. We trained the denoiser with 1000 diffusion steps, employing 50 steps for DDIM sampling during inference. For the CFG weight, we set $w = 3.5$ unless mentioned otherwise.

4.2 Comparative Experiments

Quantitative Results Table 1 shows quantitative comparisons of our InterMoE with previous methods on both InterHuman and InterX datasets. Following established practices (Liang et al. 2024; Zhang et al. 2023), each experiment is conducted 20 times, and the reported metric values represent the mean with a 95% statistical confidence interval. Our framework achieves state-of-the-art results on both InterHuman and InterX datasets. It records the lowest FID scores (4.677 on InterHuman and 0.297 on InterX), indicating superior realism and quality of generated interactions, and leads in R-Precision and MM-Dist, showing excellent semantic-fidelity. While our MultiModality is slightly lower than some methods, the high R-Precision and low MM-Dist emphasize that InterMoE prioritizes adherence to text over extreme diversity.

Qualitative Comparisons In Figure 3, we provide a qualitative comparison of interaction sequences generated by our InterMoE and prior state-of-the-art methods for the same

Datasets	Methods	R-Precision ↑ Top-1	Top-2	Top-3	FID ↓	MM-Dist ↓	Diversity →	Multi Modality ↑
Inter Human	Real motion	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
	T2M (2022a)	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
	MDM (2023)	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$7.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.350}^{\pm.080}$
	ComMDM (2024)	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$6.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
	InterGen (2024)	$0.371^{\pm.010}$	$0.515^{\pm.012}$	$0.624^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$	$7.387^{\pm.029}$	$\underline{2.141}^{\pm.063}$
	MoMat-MoGen (2024)	$0.449^{\pm.004}$	$0.591^{\pm.003}$	$0.666^{\pm.004}$	$5.674^{\pm.085}$	$3.790^{\pm.001}$	$8.021^{\pm.35}$	$1.295^{\pm.023}$
	in2IN (2024)	$0.425^{\pm.008}$	$0.576^{\pm.008}$	$0.662^{\pm.009}$	$5.535^{\pm.120}$	$3.803^{\pm.002}$	$\underline{7.953}^{\pm.047}$	$1.215^{\pm.023}$
	InterMask (2025)	$0.449^{\pm.004}$	$0.599^{\pm.005}$	$0.683^{\pm.004}$	$\underline{5.154}^{\pm.061}$	$3.790^{\pm.002}$	$\mathbf{7.944}^{\pm.033}$	$1.737^{\pm.020}$
	TIMotion (2025)	$\underline{0.491}^{\pm.005}$	$\underline{0.648}^{\pm.004}$	$\underline{0.724}^{\pm.004}$	$5.433^{\pm.080}$	$\underline{3.775}^{\pm.001}$	$8.032^{\pm.030}$	$0.952^{\pm.032}$
	Ours	$\mathbf{0.512}^{\pm.004}$	$\mathbf{0.671}^{\pm.004}$	$\mathbf{0.746}^{\pm.004}$	$\mathbf{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$	$7.990^{\pm.029}$	$0.964^{\pm.028}$
InterX	Real motion	$0.429^{\pm.004}$	$0.626^{\pm.003}$	$0.736^{\pm.003}$	$0.002^{\pm.0002}$	$3.536^{\pm.013}$	$9.734^{\pm.078}$	-
	T2M (2022a)	$0.184^{\pm.010}$	$0.298^{\pm.006}$	$0.396^{\pm.005}$	$5.481^{\pm.382}$	$9.576^{\pm.006}$	$2.771^{\pm.151}$	$2.761^{\pm.042}$
	MDM (2023)	$0.203^{\pm.009}$	$0.329^{\pm.007}$	$0.426^{\pm.005}$	$23.701^{\pm.057}$	$9.548^{\pm.014}$	$5.856^{\pm.077}$	$\underline{3.490}^{\pm.061}$
	ComMDM (2024)	$0.090^{\pm.002}$	$0.165^{\pm.004}$	$0.236^{\pm.004}$	$29.266^{\pm.067}$	$6.870^{\pm.017}$	$4.734^{\pm.067}$	$0.771^{\pm.053}$
	InterGen (2024)	$0.207^{\pm.004}$	$0.335^{\pm.005}$	$0.429^{\pm.005}$	$5.207^{\pm.216}$	$9.580^{\pm.011}$	$7.788^{\pm.208}$	$\mathbf{3.686}^{\pm.052}$
	InterMask (2025)	$0.403^{\pm.005}$	$0.595^{\pm.004}$	$0.705^{\pm.005}$	$0.399^{\pm.013}$	$\underline{3.705}^{\pm.017}$	$9.046^{\pm.073}$	$2.261^{\pm.081}$
	TIMotion (2025)	$\underline{0.412}^{\pm.004}$	$\underline{0.601}^{\pm.004}$	$\underline{0.714}^{\pm.003}$	$\underline{0.385}^{\pm.022}$	$3.706^{\pm.015}$	$\underline{9.191}^{\pm.092}$	$2.437^{\pm.069}$
	Ours	$\mathbf{0.427}^{\pm.003}$	$\mathbf{0.612}^{\pm.004}$	$\mathbf{0.721}^{\pm.004}$	$\mathbf{0.297}^{\pm.015}$	$\mathbf{3.605}^{\pm.016}$	$9.109^{\pm.083}$	$2.446^{\pm.069}$

Table 1: Quantitative evaluation results on the test sets of InterHuman and Inter-X datasets. ↑ and ↓ denote that higher and lower values are better, respectively, while → denotes that the values closer to the real motion are better. We run the evaluations 20 times. ± indicates a 95% confidence interval.

text descriptions. Given the prompt "*They are training in fencing, the first person strikes forward and the other retreats to avoid the attack*", existing methods like TIMotion exhibit significant identity confusion. Furthermore, both TIMotion and InterMask fail to render distinct offensive and defensive hand gestures or the precise forward and backward movements. In the "*Two people are engaging in a tug-of-war*" scenario, our InterMoE accurately synthesizes the hand-gripping-rope posture and the backward-leaning motion, whereas competitors fail to produce the correct kinematically plausible action. For a complex, 10-second interaction, "*Two engage in a taekwondo while moving in a circular pattern*", InterMoE not only generates coherent sparring motions but also strictly adheres to the specified circular movement pattern. In contrast, TIMotion and InterMask disregard this moving constraint, producing only stationary sparring. Collectively, these examples demonstrate that InterMoE generates higher-quality interactions with more precise semantic alignment and clearer individual-specific characteristics compared to prior methods.

4.3 Ablation Studies and Analysis

Main Ablation We begin by analyzing the impact of each key component. For a fair comparison, we define the baseline as the InterGen framework integrated with the Causal-Skeletal VAE (CS-VAE). As shown in Table 2, we first validate the contributions of our Synergistic Router design. Removing the parallel motion and text Router leads to a decline in both R-Precision and FID, demonstrating that our dual router design is crucial for enhancing semantic fidelity and generation quality. More notably, when batch-level routing is disabled and routing decisions are confined to the instance level, the FID score degrades significantly. This strongly substantiates that enabling the router to perceive and leverage global batch information is indispensable for learning a high-quality generative distribution. Next, we evaluate our temporal-selective expert mechanism. Disabling the dynamic selection and reverting to a fixed top-K features per expert also results in a substantial drop in FID, highlighting the superiority of learnable expert capacities. When removing the temporal-selective mechanism by uniformly assigning experts, its FID score is substantially worse. This performance gap highlights the critical importance of empowering experts to proactively select temporal features. Our full

Methods	R-Precision Top-1	FID ↓	MM-Dist ↓
Baseline (InterGen w/ CS-VAE)	$0.489^{\pm.006}$	$5.251^{\pm.086}$	$3.771^{\pm.001}$
Synergistic Router			
w/o Motion & Text Router	$0.503^{\pm.003}$	$4.782^{\pm.066}$	$3.766^{\pm.001}$
w/o Batch-level Routing	$0.492^{\pm.006}$	$6.036^{\pm.072}$	$3.774^{\pm.001}$
Dynamic Temporal Selection			
w/o Dynamic Selection	$0.498^{\pm.005}$	$6.242^{\pm.070}$	$3.772^{\pm.001}$
w/o Temporal-Selective	$0.505^{\pm.006}$	$5.195^{\pm.070}$	$3.767^{\pm.001}$
Ours Full	$\mathbf{0.512}^{\pm.004}$	$\mathbf{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$

Table 2: Ablation study results on the InterHuman test set to verify key components of our InterMoE.

MoE Type	R-Precision Top-1 ↑	FID ↓	MM-Dist ↓
Token Choice	$0.505^{\pm.006}$	$5.095^{\pm.070}$	$3.766^{\pm.001}$
Expert Choice	$0.441^{\pm.021}$	$8.699^{\pm.140}$	$3.796^{\pm.001}$
Ours	$\mathbf{0.512}^{\pm.004}$	$\mathbf{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$

Table 3: Ablation results of the MoE type.

Expert Number	Total Params	R-Precision Top-1 ↑	FID ↓	MM-Dist ↓
None (InterGen)	182M	$0.371^{\pm.010}$	$5.918^{\pm.079}$	$5.108^{\pm.014}$
4	164M	$0.494^{\pm.006}$	$5.114^{\pm.074}$	$3.773^{\pm.001}$
8	240M	$\mathbf{0.512}^{\pm.004}$	$\mathbf{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$
16	391M	$0.507^{\pm.005}$	$4.970^{\pm.090}$	$3.767^{\pm.003}$

Table 4: Ablation results of the expert number.

framework achieves the best performance across all metrics. These results indicate that our proposed Synergistic Router and Dynamic Temporal Selection mechanisms are not only effective individually but also work cooperatively to elevate the quality and fidelity of the generated interactions.

MoE Analysis We conduct a detailed analysis of the Mixture of Expert design, investigating the impact of different MoE paradigms, the number of experts, and different hyperparameters on InterHuman test dataset. The results are summarized in Table 3, 4, and 5.

We compare our proposed MoE against two other conventional MoE paradigms: Token-Choice (TC) and Expert-Choice (EC). As shown in 3, the results indicate that the standard EC paradigm struggles to effectively leverage the contextual information from the diffusion process during assignment and is limited in fully leveraging the capabilities of the experts. The TC paradigm, by assigning a fixed number of experts to each token, provides a stronger baseline. Our proposed method substantially outperforms both standard paradigms. The significant improvement in FID demonstrates the powerful capability of our Synergistic Router and Dynamic Temporal Selection mechanisms in enhancing the quality of the generated interaction.

We further investigate the trade-off of expert numbers as shown in Table 4. Our MoE paradigm with a small number of experts yields substantial improvements across all metrics compared to the dense baseline (InterGen). Increasing the number of experts to 8 further boosts performance, achieving the best results on all metrics. However, doubling the experts to 16 results in a slight degradation in performance across all metrics. This suggests that a larger number may introduce redundancy or require more extensive training.

We also investigate the impact of the hyperparameter K_e^{exp}, mentioned in Section 3.2. Annotating the expectation of the number of experts allocated per feature as C^{exp}, we use the value of C^{exp} to represent the distinction for different settings of K_e^{exp} following common practice by setting

$$K_e^{\text{exp}} = \frac{C^{\text{exp}} \times \text{Sequence Length}}{\text{Expert Number}}$$

C^{exp}	R-Precision Top-1 ↑	FID ↓	MM-Dist ↓
0.8	$0.510^{\pm.006}$	$4.933^{\pm.075}$	$3.766^{\pm.001}$
1	$0.512^{\pm.004}$	$\mathbf{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$
2	$\mathbf{0.517}^{\pm.007}$	$4.878^{\pm.073}$	$3.765^{\pm.001}$

Table 5: Ablation results of the expectation of the number of experts allocated per feature C^{exp}.

Methods	R-Precision Top 1 ↑	FID ↓	MM-Dist ↓
Real motion	$0.511^{\pm.003}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$
MDM (2023)	$0.320^{\pm.005}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$
+ Ours	$\mathbf{0.434}^{\pm.006}$	$\mathbf{0.483}^{\pm.031}$	$\mathbf{2.649}^{\pm.009}$
MLD (2023a)	$0.481^{\pm.003}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$
+ Ours	$\mathbf{0.493}^{\pm.002}$	$\mathbf{0.398}^{\pm.010}$	$\mathbf{3.138}^{\pm.011}$
SALAD (2025)	$0.581^{\pm.003}$	$0.076^{\pm.002}$	$2.649^{\pm.009}$
+ Ours	$\mathbf{0.586}^{\pm.003}$	$\mathbf{0.069}^{\pm.002}$	$\mathbf{2.632}^{\pm.008}$

Table 6: Quantitative results on the HumanML3D test set, demonstrate the generalization of our InterMoE framework.

as shown in Table 5, our model achieves the best performance on both FID and MM-Dist metrics when $C^{\text{exp}} = 1$, indicating optimal generation quality and diversity at this setting. Increasing C^{exp} to 2 yields a marginal improvement in R-Precision (text-motion alignment) but degrades the more critical FID score. This may be because forcing more experts to process the same feature can lead to redundancy, while also incurring higher computational costs. Conversely, decreasing C^{exp} to 0.8 also yields degradation, which suggests that a too sparse allocation provides insufficient modeling capacity to fully capture the complexity of the target interaction.

Single Human Motion Generation To investigate the generalizability of our framework, we integrated our Dynamic Temporal-Selective MoE paradigm into classic diffusion-based models for single-person motion generation while other hyperparameters remain unchanged. The experimental results, summarized in Table 6, strongly support this hypothesis. Upon incorporating our MoE, all baseline models exhibit a consistent and significant performance boost.

5 Conclusion

In this paper, we present InterMoE, a novel framework for generating 3D human interactions. The core of our work is a new dynamic temporal-selective MoE paradigm. By integrating the Synergistic Router and Dynamic Temporal Selection mechanism, InterMoE achieves significant improvements in both individual-specific characteristics and semantic fidelity. Comprehensive experiments demonstrate that InterMoE surpasses existing state-of-the-art models on several key metrics. Moreover, the excellent performance of our MoE paradigm on single-person tasks underscores the generalizability and broad potential of our framework.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62132001), and the Fundamental Research Funds for the Central Universities. And we would like to express our gratitude to our collaborators for their efforts.

References

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M.-Y. 2022. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*.

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

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023a. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Chen, Z.; Wang, Z.; Wang, Z.; Liu, H.; Yin, Z.; Liu, S.; Sheng, L.; Ouyang, W.; Qiao, Y.; and Shao, J. 2023b. Octavius: Mitigating Task Interference in MLLMs via MoE. arXiv:2311.02684.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Guo, D.; et al. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Fei, Z.; Fan, M.; Yu, C.; Li, D.; and Huang, J. 2024. Scaling Diffusion Transformers to 16 Billion Parameters. *arXiv preprint arXiv:2407.11633*.

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; Sun, Y.; Chen, L.; Tian, H.; Wu, H.; and Wang, H. 2023. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts. arXiv:2210.15257.

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

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

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

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

Hong, S.; Kim, C.; Yoon, S.; Nam, J.; Cha, S.; and Noh, J. 2025. SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7158–7168.

Javed, M. G.; Li, X.; et al. 2025. InterMask: 3D Human Interaction Generation via Collaborative Masked Modelling. In *The Thirteenth International Conference on Learning Representations*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36: 20067–20079.

Kong, H.; Gong, K.; Lian, D.; Mi, M. B.; and Wang, X. 2023. Priority-Centric Human Motion Generation in Discrete Latent Space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14806–14816.

Lee, Y.; Kim, J.-Y.; Go, H.; Jeong, M.; Oh, S.; and Choi, S. 2023. Multi-Architecture Multi-Expert Diffusion Models. arXiv:2306.04990.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, B.; Wang, X.; Song, R.; and Huang, W. 2024. Two-in-One: Unified Multi-Person Interactive Motion Generation by Latent Diffusion Transformer. arXiv:2412.16670.

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

Lou, Y.; Zhu, L.; Wang, Y.; Wang, X.; and Yang, Y. 2023. DiverseMotion: Towards Diverse Human Motion Generation via Discrete Diffusion. *arXiv preprint arXiv:2309.01372*.

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

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

MiniMax; et al. 2025. MiniMax-01: Scaling Foundation Models with Lightning Attention. arXiv:2501.08313.

Muennighoff, N.; Soldaini, L.; Groeneveld, D.; Lo, K.; Morrison, J.; Min, S.; Shi, W.; Walsh, P.; Tafjord, O.; Lambert, N.; Gu, Y.; Arora, S.; Bhagia, A.; Schwenk, D.; Wadden, D.; Wettig, A.; Hui, B.; Dettmers, T.; Kiela, D.; Farhadi, A.; Smith, N. A.; Koh, P. W.; Singh, A.; and Hajishirzi, H. 2024. OLMoE: Open Mixture-of-Experts Language Models. arXiv:2409.02060.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

Ruiz-Ponce, P.; Barquero, G.; Palmero, C.; Escalera, S.; and García-Rodríguez, J. 2024. in2IN: Leveraging individual Information to Generate Human INteractions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1941–1951.

Sehwag, V.; Kong, X.; Li, J.; Spranger, M.; and Lyu, L. 2024. Stretching Each Dollar: Diffusion Training from Scratch on a Micro-Budget. arXiv:2407.15811.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2024. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shi, M.; Yuan, Z.; Yang, H.; Wang, X.; Zheng, M.; Tao, X.; Zhao, W.; Zheng, W.; Zhou, J.; Lu, J.; Wan, P.; Zhang, D.; and Gai, K. 2025. DiffMoE: Dynamic Token Selection for Scalable Diffusion Transformers. *arXiv preprint arXiv:2503.14487*.

Su, J. 2025. A Journey of MoE: 4. Invest More in Difficulties. https://spaces.ac.cn/archives/10815. Accessed: 2025-07-15.

Sun, H.; Lei, T.; Zhang, B.; Li, Y.; Huang, H.; Pang, R.; Dai, B.; and Du, N. 2024. EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing. arXiv:2410.02098.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. ISBN 9781510860964.

Wang, Y.; Wang, S.; Zhang, J.; Fan, K.; Wu, J.; Xue, Z.; and Liu, Y. 2025. TIMotion: Temporal and Interactive Framework for Efficient Human-Human Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, L.; Lu, S.; Pi, H.; Fan, K.; Pan, L.; Zhou, Y.; Feng, Z.; Zhou, X.; Peng, S.; and Wang, J. 2025. MotionStreamer: Streaming Motion Generation via Diffusion-based Autoregressive Model in Causal Latent Space. *arXiv preprint arXiv:2503.15451*.

Xu, L.; Lv, X.; Yan, Y.; Jin, X.; Wu, S.; Xu, C.; Liu, Y.; Zhou, Y.; Rao, F.; Sheng, X.; et al. 2024. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, 22260–22271.

Yatharth Gupta, H. P.; and Jaddipal, V. V. 2024. SegMoE: Segmind Mixture of Diffusion Experts. https://huggingface.co/segmind/SegMoE-4x2-v0.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Gupta, A.; Gu, X.; Hauptmann, A. G.; Gong, B.; Yang, M.-H.; Essa, I.; Ross, D. A.; and Jiang, L. 2023. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. arXiv:2310.05737.

Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14730–14740.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024a. MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4115–4128.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024b. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y.; Huang, D.; Liu, B.; Tang, S.; Lu, Y.; Chen, L.; Bai, L.; Chu, Q.; Yu, N.; and Ouyang, W. 2024c. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7368–7376.