HHI-DYN: LEARNING INTERACTION DYNAMICS FOR HUMAN-HUMAN MOTION DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-driven human interaction generation demands a robust latent space structure, where each individual's motion patterns are jointly shaped by the textual description and the partner's behavior. Existing interaction models neglect the coupling between local joints and global motion representations. Moreover, they typically incorporate partner information solely through cross-attention module, failing to account for the varying influence of the partner across different interaction types. To address these limitations, we propose HHI-Dyn, a novel latent diffusion framework for human-human motion generation. HHI-Dyn first introduces a tailored VAE that effectively fuses global spatial features with local spatial details. At each time step, every local body part adaptively integrates with the global spatial context, yielding a more stable latent representation for subsequent diffusion. A latent diffusion model is then trained to capture this spatial structure conditioned on interaction text. Crucially, we design an interactive expert module for motion pattern learning, where each expert is responsible for capturing a specific type of motion pattern. These experts dynamically adapt to diverse input conditions and partner behaviors, effectively improving both the interactive quality and perceptual realism of the generated two-person interactions. Extensive experiments demonstrate that HHI-Dyn achieves state-of-the-art performance on InterHuman and InterX datasets, with significant improvements in text-motion alignment and motion realism. We will release code and models to facilitate reproducibility.

1 INTRODUCTION

Generating interactive motions from natural language descriptions is an important task in human motion generation, aiming to produce semantically faithful two-person interactions. Compared to single-person motion generation (Zhang et al., 2023a; Tevet et al., 2023; Guo et al., 2024; Pinyoanuntapong et al., 2024), human interaction generation poses additional challenges, which we attribute to two key aspects. *Challenging latent space construction:* in contrast to single-person motions that embed compactly, two-person interactions demand a more expressive spatial structure to capture interactional dependence, thereby increasing the difficulty of learning the latent space. *Interaction complexity:* generating two-person motions requires not only that each individual's motion aligns with the textual condition, but also that their mutual interactions are naturally and coherently modeled. These two challenges call for a new framework that jointly addresses the structural demands of latent space learning and the intricate modeling of interaction dynamics.

Motivated by these observations, we propose HHI-Dyn, a framework designed to generate realistic and natural human interactions from textual descriptions. We argue that dynamically fusing global motion features with spatio-temporal local features can yield a more comprehensive latent representation, producing more robust motion encodings. First, we introduce a Spatio-Temporal Hybrid Causal VAE (STC-VAE) for motion encoding. Existing motion encoding methods either directly encode global motion features to capture the overall body posture (Chen et al., 2023; Geng et al., 2026) or adopt spatio-temporal modeling to partition spatial joints into local regions for fine-grained representations (Javed et al., 2025; Hong et al., 2025; Wang et al., 2026). However, we contend that such partitioning inevitably disrupts the coupling between local joints and global motion representations. For example, in tug-of-war, local joint motions such as gripping and leg bracing must be coordinated with global body posture and the spatial relation between the two individuals. Modeling only local joints may produce pulling-like hands and legs but fail to capture the overall backward

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

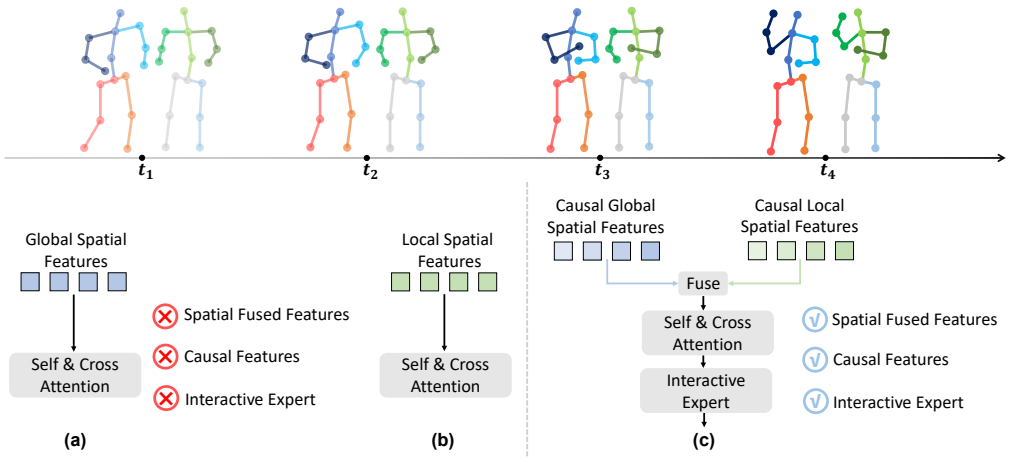


Figure 1: (a) Learning global spatial features with both self-attention and cross-attention modules. (b) Learning local spatial features with both self-attention and cross-attention modules. (c) Learning fused spatial features with self-attention, cross-attention, and interactive expert modules.

Table 1: Comparison between our model and the state-of-the-art model InterMoE Wang et al. (2026) on the text-conditioned human interaction dataset InterHuman Liang et al. (2024).

Methods	FID ↓	R-Precision Top 1 ↑	Dynamic Global Local Fusion	Learning With Partner Router	Capturing Interaction Types
InterMoE	4.677	0.512	✗	✗	✗
HHI-Dyn	**4.658**	**0.515**	✓	✓	✓

lean or the adversarial interaction, showing the need to jointly model local and global spatial information. To address this, we simultaneously construct causal global spatial representations and causal spatio-temporal partitioned features. The causal temporal design ensures that the feature ordering adheres to natural motion dynamics, thereby providing a more robust foundation for motion encoding. Furthermore, we introduce a spatial gating mechanism that performs gated interaction computation between each local part from the spatio-temporal representation and the global spatial representation, yielding dynamically fused features. As a result, each local spatial part at every timestep can be adaptively integrated with global spatial information, ultimately leading to a more stable latent space that lays a solid groundwork for subsequent diffusion-based generation. The advantages of our method are shown in Figure 1.

Second, building upon the aforementioned motion representations, we propose an Interactive Diffusion Transformer (Inter-DiT) tailored for two-person scenarios. Unlike previous approaches (Liang et al., 2024; Jiao et al., 2026; Wang et al., 2026), which rely solely on cross-attention to enhance pairwise interactions, we argue that partner information is crucial for learning motion patterns, and that different partners' motions exert varying degrees of influence on the agent. For tightly coupled motions like handshaking or Latin dancing, the influence is generally strong, whereas for loosely related motions, such as one person taking a photo while the other poses remotely, it is weak. We therefore further introduce an Interactive Expert Module to capture such cooperative motion patterns. Specifically, each expert within this module is responsible for capturing key temporal motion patterns learned from the entire motion sequence. Meanwhile, a Cooperative Router leverages high-level semantic information derived from textual descriptions, the agent's own kinematic features, and the partner's kinematic features to jointly guide routing decisions. As a result, the motion segments selected by each expert exhibit stronger interactional consistency with the partner. A comparison with the current state-of-the-art method, InterMoE Wang et al. (2026), is presented in Table 1.

The main contributions of this work are threefold. First, we propose a novel VAE with dynamic spatial fusion, which to the best of our knowledge is the first attempt to construct a more robust latent motion space via adaptive spatial gating, preserving natural dynamics for stable diffusion-based generation. Second, we introduce a diffusion-based generation model with an Interactive Expert Module, where multiple experts capture key temporal patterns and a Cooperative Router jointly leverages text and partner kinematics to enhance interactional coordination and semantic plausibility.

Finally, our HHI-Dyn framework achieves state-of-the-art performance on InterHuman and InterX, with ablation studies and user evaluations confirming its effectiveness and generalization.

2 Related Work

2.1 Human Motion Generation

Recent years have witnessed rapid progress in motion generation, enabling high-quality synthesis from text (Zhang et al., 2023a; Yao et al., 2026), audio (Li et al., 2023; Xie et al., 2026), music (Aristidou et al., 2022; Zhang et al., 2025), and motion prefixes (Mao et al., 2019; Liu et al., 2022). Early methods used GANs and sequence-to-sequence models (Ahn et al., 2018; Ahuja & Morency, 2019), followed by VAE-based approaches (Guo et al., 2022; Petrovich et al., 2022) that improved diversity. Diffusion models such as MDM (Tevet et al., 2023) and MotionDiffuse (Zhang et al., 2024a) further enhanced realism, while MLD (Chen et al., 2023) and MotionLCM (Dai et al., 2024) improved sampling efficiency. ReMoDiffuse (Zhang et al., 2023b) and SALAD (Hong et al., 2025) refined generation quality via retrieval and graph convolution, respectively. T2M-GPT (Zhang et al., 2023a) quantized motions into discrete indices and generated fine-grained motions autoregressively under textual guidance. MoMask (Guo et al., 2024) and MMM (Pinyoanuntapong et al., 2024) adopted masked generation to improve fidelity, while subsequent works introduced finer control over spatial trajectories (Xie et al., 2024; Pinyoanuntapong et al., 2025) and linguistic constraints (Wan et al., 2024; Huang et al., 2024). Despite these achievements, existing methods rely on either global spatial information or local spatial partitions, overlooking their inherent spatial correlations. To address this, we are the first to integrate global and local spatial information to improve motion encoding.

2.2 Human-Human Interaction Generation

Current research on human interaction generation mainly focuses on two-person scenarios. Early work ComMDM (Shafir et al., 2024) combined two single-person motion diffusion models via bridge networks for human interactive motion generation. RIG (Tanaka & Fujiwara, 2023) and InterGen (Liang et al., 2024) pioneered dedicated networks for directly modeling human-human interactions, while in2IN (Ruiz-Ponce et al., 2024) jointly used individual and interaction descriptions to improve text alignment and generation quality. MoMat-MoGen (Cai et al., 2024) retrieved motion data from databases to simulate interactive behaviors between agents. InterMask (Javed et al., 2025) and HiTMM (Jiao et al., 2026) adopted generative masked approaches for human-human interactions, while TIMotion (Wang et al., 2025) formulated the interaction process as a causal sequence. Text2Interact (Wu et al., 2026) further constructed high-quality and diverse two-person interaction scenes using textual information from large language models. InterMoE (Wang et al., 2026) partitioned body parts via graph convolution and employed expert networks to improve motion generation accuracy. Despite this progress, existing methods typically use cross-attention to incorporate partner information and overlook diverse interaction types. To address this, we propose a novel Interactive Expert Module to capture them.

3 Method

Problem Formulation. Given an interaction text description $\mathbf{c}$, our human-human motion generation task aims to produce interaction sequences $\{\mathbf{M}_h\}_{h\in\{a,b\}}$ that are semantically and spatially faithful to the provided text, where a and b denote the two individuals involved in the interaction. Here, $\mathbf{M}_h \in \mathbb{R}^{T\times J\times D}$ denotes the motion sequences for person h, where T is the number of frames, J is the number of joints per frame, and D is the dimensionality of each joint feature vector.

3.1 Spatio-Temporal Hybrid Causal VAE (STC-VAE)

3.1.1 Spatio-Temporal Hybrid Encoding

To address the limitation of existing methods in that latent features neglect the coupling between local joints and global motion representations, we propose a novel Spatio-Temporal Hybrid Causal VAE. This stage aims to learn latent representations that achieve better reconstruction quality and are

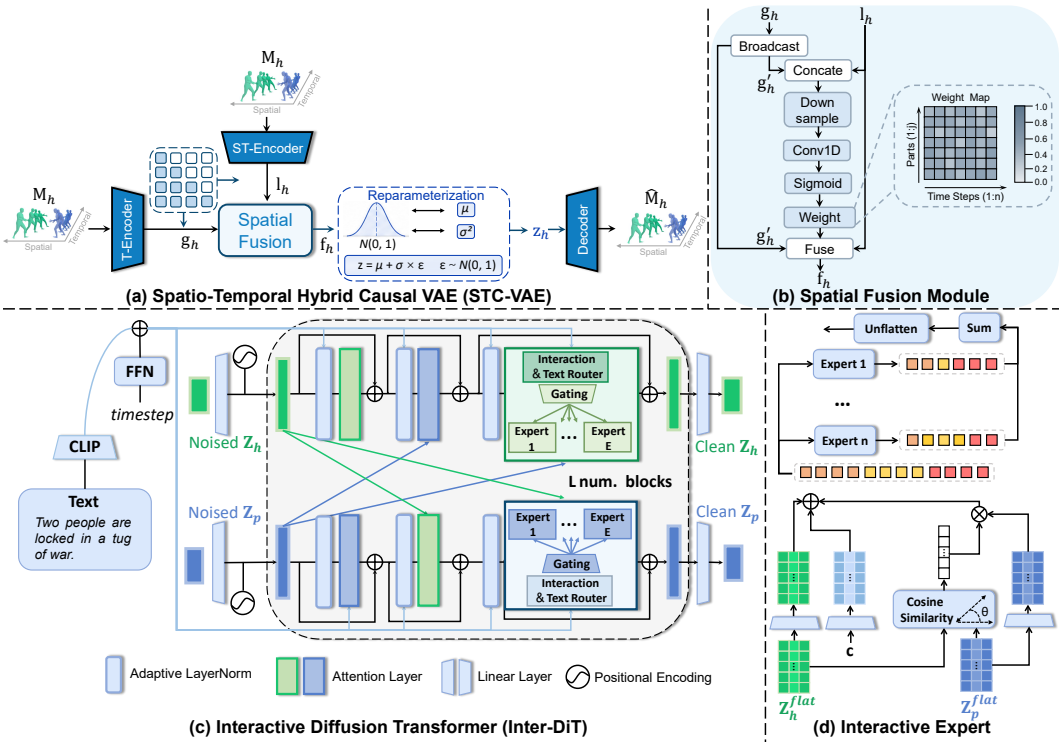


Figure 2: **Overview of the proposed framework**. (a) The latent representation of each individual's motion is obtained through dynamic spatial fusion and reparameterization sampling to produce $\mathbf{z}_h$. (b) The spatial fusion module integrates global and local spatial information. (c) Inter-DiT performs denoising on the noised latent variable $\mathbf{Z}$ conditioned on the text description $\mathbf{c}$. (d) The interactive expert dynamically selects features based on the person's own motion, the partner's motion, and the textual condition.

more expressive for diffusion-based generation. As illustrated in Figure 2(a), during reconstruction training, we treat each individual's motions as the same data type, allowing a unified VAE to model the interactive motions between two persons. Specifically, we first employ a temporal encoder and a spatio-temporal encoder to obtain latent features. The Temporal Encoder (T-Encoder) downsamples the temporal dimension of motions. It maps the motion sequences $\mathbf{M}_h \in \mathbb{R}^{T\times J\times D}$ into global spatial latent vector sequences $\mathbf{g}_h \in \mathbb{R}^{t\times d}$ with a downsampling ratio of t/T and a latent dimension of d. In this process, the global spatial information of each frame (originally of size $J \times D$) is directly encoded as a whole into the latent space of dimension d, thereby preserving the overall spatial structure.

Meanwhile, the Spatio-Temporal Encoder (ST-Encoder) maps the motion sequences $\mathbf{M}_h \in \mathbb{R}^{T\times J\times D}$ into local spatial latent vector sequences $\mathbf{l}_h \in \mathbb{R}^{t\times j\times d}$ with a temporal downsampling ratio of t/T, a spatial downsampling ratio of j/J, and a latent dimension d. In contrast to the T-Encoder, the ST-Encoder additionally downsamples the spatial dimension from J joints to j parts, where each of the j parts retains different local spatial information of the overall motions. It is worth noting that both the T-Encoder and ST-Encoder adopt causal convolution during encoding, ensuring that each downsampled token can only attend to preceding motion information rather than future frames. This design preserves the temporal causality of the latent features and enables stable encoding of sequential data.

3.1.2 Spatial Fusion Module

Given the global spatial representation $\mathbf{g}_h$ and the local spatial representation $\mathbf{l}_h$, we perform spatial fusion to ensure that each local spatial part can attend to the overall spatial context while preserving its own motion details, as illustrated in Figure 2(b). Specifically, we first broadcast g_h by a factor of j to match the dimension of $\mathbf{l}_h$, enabling each of the j parts to fuse with the global spatial information $\mathbf{g}_h$. The expanded global spatial features $\mathbf{g}'_h$ are then concatenated with the local spatial features along the latent dimension, followed by a downsampling operation to restore the original dimension

of $\mathbf{l}_h$. Subsequently, a 1D convolution and a Sigmoid activation are applied to generate the spatial fusion weight map $\mathbf{W}_s$. Each entry in this map indicates, for a given timestep, the importance weight assigned to the fusion between the global spatial features and a specific local spatial part. The final fused representation f_h is obtained via a weighted combination of the global and local spatial representations:

$$\mathbf{f}_h = \mathbf{l}_h \odot \mathbf{W}_s + (\mathbf{1} - \mathbf{W}_s) \odot \mathbf{g}'_h. \tag{1}$$

This dynamic weight map adaptively adjusts according to the input motion and the characteristics of each spatial part, enabling context-aware fusion. Next, we learn the mean and variance of f_h for reparameterization sampling to obtain the latent variable $\mathbf{z}_h$, which serves as the preparation for the subsequent diffusion denoising process. Finally, the reparameterized latent variable $\mathbf{z}_h$ is passed through a decoder to reconstruct the motion sequences $\hat{\mathbf{M}}_h$. The training loss details can be found in Appendix A.

3.2 Interactive Diffusion Transformer (Inter-DiT)

3.2.1 Transformer Block Design

Existing two-person motion denoisers (Liang et al., 2024; Ruiz-Ponce et al., 2024; Javed et al., 2025; Wang et al., 2026) rely only on cross-attention to incorporate partner information, which fails to capture dynamic interaction patterns. To bridge this gap, we introduce an interactive expert module within each denoising block. As illustrated in Figure 2(c), our denoiser adopts a block design with weights shared across both persons. Given the input feature $\mathbf{z}_h \in \mathbb{R}^{t\times j\times d}$, we first reshape it into $\mathbf{z}_h \in \mathbb{R}^{t\times D_z}$, where $D_z = j \times d$. The reshaped feature then passes through self-attention within each block to capture intra-person dynamics, followed by cross-attention to enable inter-person information exchange. Subsequently, the resulting feature is fed into our dynamic interactive expert to learn the interaction dynamics between the two individuals.

3.2.2 Interactive Expert Module

As illustrated in Figure 2(d), the interactive expert module is designed to learn motion patterns, where different experts specialize in distinct types of motion patterns. To facilitate efficient learning, we flatten the input motion features along the batch dimension. Specifically, we reshape $\mathbf{Z}_h \in \mathbb{R}^{B\times t\times D_z}$ into $\mathbf{Z}_h^{\text{flat}} \in \mathbb{R}^{N\times D_z}$, where $N = B \times t$ and B denotes the batch size, yielding a batch-level temporal feature pool. This allows each expert to simultaneously learn all relevant motion patterns within a batch, thereby improving learning efficiency. Our cooperative routing strategy takes three sources as input, including the person's own motion features $\mathbf{Z}_h^{\text{flat}}$, the text condition $\mathbf{c}$, and the partner's motion features $\mathbf{Z}_p^{\text{flat}}$. Specifically, we employ three routing modules. The motion router processes the person's own features to produce motion routing logits, the text router operates on the text condition to produce text routing logits, and the second motion router processes the partner's features to produce partner routing logits. For each expert e, let $\mathbf{Z}_h^{\text{flat}} = [\mathbf{z}_{n,h}^{\text{flat}}]$, where $n \in \{1, \ldots, N\}$. The motion routing logits $\mathbf{R}_{e,n,h}^{\text{motion}}$, text routing logits $\mathbf{R}_e^{\text{text}}$, and partner routing logits $\mathbf{R}_{e,n,p}^{\text{partner}}$ are computed as follows:

$$\mathbf{R}_{e,n,h}^{\text{motion}} = \text{Router}_e^{\text{motion}}(\mathbf{z}_{n,h}^{\text{flat}}), \quad \mathbf{z}_{n,h}^{\text{flat}} \in \mathbb{R}^{1\times D_z} \tag{2}$$

$$\mathbf{R}_e^{\text{text}} = \text{Router}_e^{\text{text}}(\mathbf{c}), \quad \mathbf{c} \in \mathbb{R}^{1\times D_z} \tag{3}$$

$$\mathbf{R}_{e,n,p}^{\text{partner}} = \text{Router}_e^{\text{partner}}(\mathbf{z}_{n,p}^{\text{flat}}), \quad \mathbf{z}_{n,p}^{\text{flat}} \in \mathbb{R}^{1\times D_z} \tag{4}$$

The degree of influence exerted by the partner inherently depends on the nature of the interaction. For tightly coupled motions such as pushing or pulling an object together, the partner's motion strongly constrains the agent's movement, whereas for loosely related motions such as two people reading in the same room, the influence is much weaker. To capture different interaction types, we modulate the partner routing logits by the cosine similarity between the person's own features and the partner's features:

$$\mathbf{R}_{e,n,p}^{\text{partner_cos}} = \mathbf{R}_{e,n,p}^{\text{partner}} \cdot \cos\big(\mathbf{Z}_h^{\text{flat}}, \mathbf{Z}_p^{\text{flat}}\big), \tag{5}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. The final routing logits are then computed as:

$$\mathbf{R}_{e,n,h}^{\text{final}} = \alpha\mathbf{R}_{e,n,h}^{\text{motion}} + (1-\alpha)\mathbf{R}_e^{\text{text}} + \beta\mathbf{R}_{e,n,p}^{\text{partner_cos}}, \tag{6}$$

where α and β are balancing hyper-parameters. With this design, the routing decision is jointly guided by individual-specific motion dynamics, high-level semantic cues, and interactive dynamics from the partner. Furthermore, we adopt a dynamic selection mechanism that allows each expert to actively select key tokens from the batch-level temporal feature pool, which includes a total of N features. In each expert module, we associate each expert with a learnable bias parameter. Specifically, we have:

$$\mathbf{S}_{e,n,h} = \texttt{sigmoid}(\mathbf{R}^{\text{final}}_{e,n,h}) + b_e, \tag{7}$$

$$\mathbf{A}_{e,n,h} = \texttt{softmax}(\mathbf{R}^{\text{final}}_{e,n,h}), \tag{8}$$

$$\mathbf{W}_{e,n,h} = \begin{cases} \mathbf{A}_{e,n,h}, & \mathbf{S}_{e,n,h} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here, $\mathbf{W}_{e,n,h}$ denotes the gating weight, and $\mathbf{S}_{e,n,h}$ determines whether the feature is selected for expert e. Since the sigmoid function maps values to $(0, 1)$, the bias b_e is constrained to $(-1, 0)$, which controls the number of features that expert e will process. These bias parameters b_e are gradually optimized during training. With k_e^{exp} as a predefined hyperparameter, we count the number of selected features k_e^{sel} for each expert at each training step:

$$H_e^{\text{select}} = \texttt{Count}(\mathbf{S}_{e,n,h} > 0), \quad n \in \{1, \ldots, N\}. \tag{10}$$

The target capacity k_e^{exp} is set to $k_e^{\text{exp}} = \lfloor N/E \rfloor$, where E is the number of experts. This formulation ensures that each expert processes a balanced and controllable number of tokens on average. After this training step, we update the bias b_e as follows:

$$\Delta b_e = \text{sign}(k_e^{\text{select}} - k_e^{\text{exp}}), \tag{11}$$

$$b_e \leftarrow b_e - \sigma\, \Delta b_e, \tag{12}$$

where σ is a gradual coefficient, and $\text{sign}(\cdot)$ returns 1, 0, or -1 for positive, zero, or negative inputs, respectively. As training converges, Δb_e approaches zero, ensuring that the number of selected features $\mathbb{E}(k_e^{\text{select}})$ approximates the target k_e^{exp}, which facilitates a dynamic yet stable allocation. Finally, the output feature $\mathbf{z}_l^{\text{out}}$ from the interactive expert module is obtained as:

$$\mathbf{z}^{\text{out}}_{n,h} = \sum_{e=1}^{E} \mathbf{W}_{e,n,h} \cdot \mathit{Expert}(\mathbf{z}^{\text{flat}}_{n,h}), \quad \mathbf{z}^{\text{flat}}_{n,h} \in \mathbb{R}^{1 \times D_z},\ n \in \{1, \ldots, N\}. \tag{13}$$

This design allows each expert to dynamically determine its feature selection capacity and select its preferred temporal motions across all sequences within the batch, conditioned on high-level semantic information and partner kinematics.

4 Experiments

4.1 Experimental Setup

Dataset. We employ two widely used datasets, InterHuman (Liang et al., 2024) and InterX (Xu et al., 2024), for training and evaluation. InterHuman contains 7,779 interactive motion sequences, each with three textual descriptions. It adopts the AMASS (Mahmood et al., 2019) skeleton representation with 22 joints. Each single-person motion is encoded as sequences of $\mathbf{m}_h = [\mathbf{j}^p, \mathbf{j}^p, \mathbf{j}^r, \mathbf{c}^f]$, where $\mathbf{j}^p \in \mathbb{R}^{3J}$ denotes global joint positions, $\mathbf{j}^p \in \mathbb{R}^{3J}$ global joint velocities, $\mathbf{j}^r \in \mathbb{R}^{6J}$ local joint rotations, and $\mathbf{c}^f \in \mathbb{R}^4$ binary foot contact features, yielding a 262-dimensional representation for each frame. InterX comprises 11,388 interactive motion sequences, each also paired with three textual annotations. It is built upon the SMPL-X (Pavlakos et al., 2019) skeleton, which includes 56 joints covering the body, hands, and face, along with root orientation and translation, resulting in a 336-dimensional motion feature $\mathbf{m}_h$ for each frame.

Metrics. Following standard evaluation protocols in human-human interaction generation (Liang et al., 2024; Ruiz-Ponce et al., 2024; Javed et al., 2025), we adopt R-Precision (Top-1, 2, 3), Frechet Inception Distance (FID), Multimodal Distance (MM Dist), Diversity, and Multimodality (MModality) to evaluate our models. For a comprehensive description of each metric, we refer readers to the Appendix B.

Table 2: **Quantitative evaluation** on the **InterHuman** and **InterX** test sets. $\pm$ indicates a 95% confidence interval and $\rightarrow$ means the closer to ground truth the better. **Bold** face indicates the best result, while underscore refers to the second best.

Dataset	Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
		Top 1	Top 2	Top 3				
Inter Human	Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
	T2M (Guo et al., 2022)	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$5.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
	MDM (Tevet et al., 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 (Shafir et al., 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 (Liang et al., 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 (Cai et al., 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 (Ruiz-Ponce et al., 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 (Javed et al., 2025)	$0.449^{\pm.004}$	$0.599^{\pm.005}$	$0.683^{\pm004}$	$5.154^{\pm.061}$	$3.790^{\pm.002}$	$\mathbf{7.944}^{\pm.033}$	$1.737^{\pm.020}$
	TIMotion (Wang et al., 2025)	$0.491^{\pm.005}$	$0.648^{\pm.004}$	$0.724^{\pm.004}$	$5.433^{\pm.080}$	$3.775^{\pm.001}$	$8.032^{\pm.030}$	$0.952^{\pm.032}$
	Text2Interact (Wu et al., 2026)	$0.483^{\pm.005}$	$0.638^{\pm.005}$	$0.717^{\pm.005}$	$5.191^{\pm.055}$	$3.778^{\pm.001}$	$7.900^{\pm.030}$	$1.051^{\pm.031}$
	HiTMM (Jiao et al., 2026)	$0.452^{\pm.004}$	$0.613^{\pm.003}$	$0.697^{\pm.003}$	$5.017^{\pm.054}$	$3.789^{\pm.001}$	$7.971^{\pm.030}$	$0.916^{\pm.034}$
	InterMoE (Wang et al., 2026)	$\underline{0.512}^{\pm.004}$	$\mathbf{0.671}^{\pm.004}$	$\mathbf{0.746}^{\pm.004}$	$\underline{4.677}^{\pm.069}$	$\mathbf{3.762}^{\pm.001}$	$7.990^{\pm.029}$	$0.964^{\pm.028}$
	HHI-Dyn	$\mathbf{0.515}^{\pm.005}$	$\underline{0.669}^{\pm.003}$	$\underline{0.744}^{\pm004}$	$\mathbf{4.658}^{\pm.053}$	$\underline{3.765}^{\pm.001}$	$7.964^{\pm.029}$	$0.882^{\pm.029}$
InterX	Ground Truth	$0.429^{\pm.004}$	$0.626^{\pm.003}$	$0.736^{\pm.003}$	$0.002^{\pm.0002}$	$3.536^{\pm.013}$	$9.734^{\pm.078}$	-
	T2M (Guo et al., 2022)	$0.184^{\pm.010}$	$0.298^{\pm.006}$	$0.396^{\pm.005}$	$5.481^{\pm.382}$	$9.576^{\pm.006}$	$2.771^{\pm.151}$	$2.761^{\pm.042}$
	MDM (Tevet et al., 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 (Shafir et al., 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 (Liang et al., 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 (Javed et al., 2025)	$0.403^{\pm.005}$	$0.595^{\pm.004}$	$0.705^{\pm005}$	$0.399^{\pm.013}$	$3.705^{\pm.017}$	$9.046^{\pm.073}$	$2.261^{\pm.081}$
	TIMotion (Wang et al., 2025)	$0.412^{\pm.004}$	$0.601^{\pm.004}$	$0.714^{\pm.003}$	$0.385^{\pm.022}$	$3.706^{\pm.015}$	$\underline{9.191}^{\pm.092}$	$2.437^{\pm.069}$
	HiTMM (Jiao et al., 2026)	$0.425^{\pm.003}$	$\underline{0.614}^{\pm.004}$	$\underline{0.727}^{\pm.003}$	$0.373^{\pm.012}$	$3.677^{\pm.013}$	$\mathbf{9.371}^{\pm.064}$	$2.316^{\pm.065}$
	InterMoE (Wang et al., 2026)	$\underline{0.427}^{\pm.003}$	$0.612^{\pm.004}$	$0.721^{\pm.004}$	$\mathbf{0.297}^{\pm.015}$	$\underline{3.605}^{\pm.016}$	$9.109^{\pm.083}$	$2.446^{\pm.069}$
	HHI-Dyn	$\mathbf{0.428}^{\pm.004}$	$\mathbf{0.619}^{\pm.004}$	$\mathbf{0.729}^{\pm.004}$	$\underline{0.314}^{\pm.016}$	$\mathbf{3.602}^{\pm.014}$	$9.113^{\pm.072}$	$2.226^{\pm.066}$

4.2 Comparison with the State-of-the-art

Quantitative Comparison. Table 2 presents a quantitative comparison between HHI-Dyn and state-of-the-art methods. Each experiment is repeated 20 times, and we report both the mean and the 95% confidence interval for each metric. HHI-Dyn achieves state-of-the-art performance on both InterHuman and InterX. In terms of motion quality, our method achieves the best FID on InterHuman (4.658) and the second-best FID on InterX (0.314), demonstrating superior realism in the generated interactions. Moreover, its leading R-Precision and MM-Dist scores reflect strong semantic alignment with the input text. Given the differences in body representations across the two datasets, these results underscore the robustness of our approach and its independence from specific skeletal configurations. Although the MultiModality score is slightly lower than that of some competing methods, the high R-Precision and low MM-Dist suggest that HHI-Dyn prioritizes text adherence over extreme diversity, while still maintaining a sufficient level of diversity to generate a wide range of two-person interactions.

Table 3: Ablation study on the InterHuman test set to verify key components of **HHI-Dyn**.

Method	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.948^{\pm.064}$	-
w/o Interactive Expert	$0.475^{\pm.004}$	$0.627^{\pm.004}$	$0.704^{\pm004}$	$5.697^{\pm.091}$	$3.781^{\pm.001}$	$7.883^{\pm.029}$	$\mathbf{0.970}^{\pm.018}$
w/o Casual Encoding	$0.493^{\pm.003}$	$0.634^{\pm.004}$	$0.716^{\pm003}$	$5.361^{\pm.058}$	$3.774^{\pm.001}$	$7.923^{\pm.026}$	$0.944^{\pm.027}$
HHI-Dyn	$\mathbf{0.515}^{\pm.005}$	$\mathbf{0.669}^{\pm.003}$	$\mathbf{0.744}^{\pm004}$	$\mathbf{4.658}^{\pm.053}$	$\mathbf{3.765}^{\pm.001}$	$\mathbf{7.964}^{\pm.029}$	$0.882^{\pm.029}$

Qualitative Comparison. We qualitatively compare our model with InterMask, HiTMM, and InterMoE, all trained on the InterHuman dataset, under the same text descriptions. As shown in Figure 3, for the first text prompt, InterMask produces unnatural hand and leg movements, the second person in the interactions generated by HiTMM does not step back, and InterMoE exhibits interpenetration. In contrast, HHI-Dyn generates motions that are more consistent with real-world actions. For the second text prompt, InterMask, HiTMM, and InterMoE fail to vividly depict two people pulling with force, whereas HHI-Dyn successfully depicts two people engaged in a tug of war. For the third interaction prompt, InterMask, HiTMM, and InterMoE depict the other person as failing to

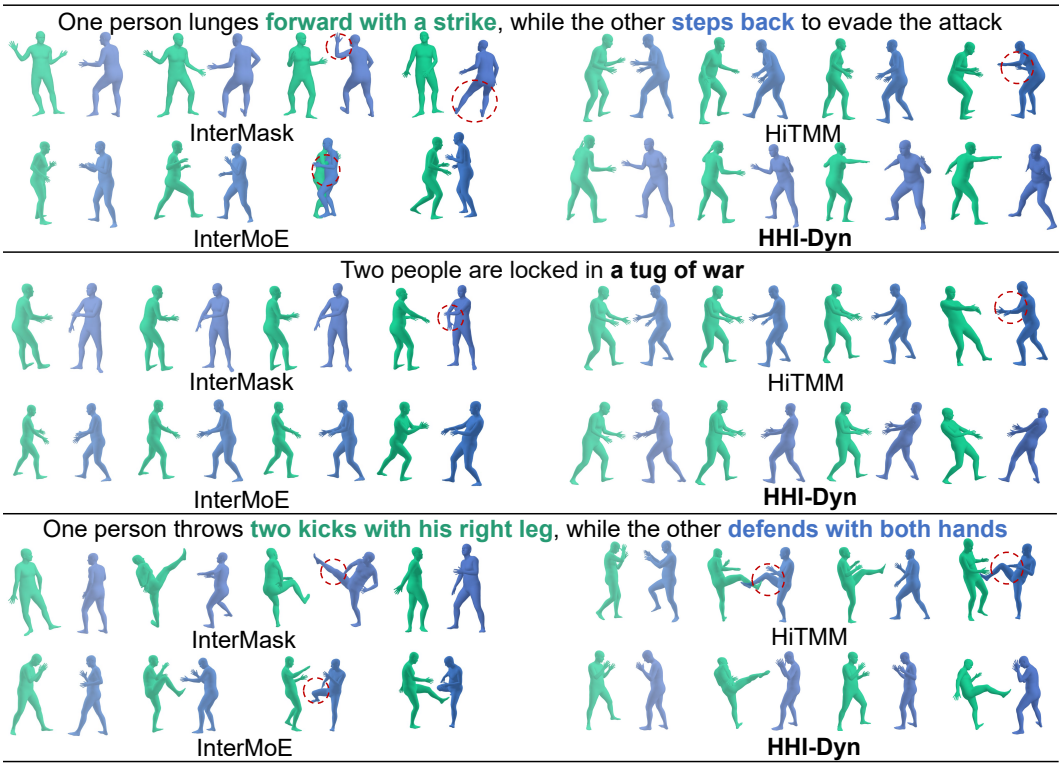


Figure 3: Qualitative comparison among InterMask, HiTMM, InterMoE and our HHI-Dyn, demonstrating that HHI-Dyn achieves superior interaction quality, stronger text adherence, and reduced implicit bias.

defend and instead throwing the same right-leg kick. In contrast, HHI-Dyn faithfully generates the described scene. These results highlight that HHI-Dyn accurately captures these spatial dynamics and interactions while also outperforming state-of-the-art methods in terms of pose robustness, an aspect that is not adequately measured by evaluation metrics. More qualitative comparisons can be found in Appendix D.

User Study. We conducted a user preference study on Interaction Quality and Adherence to Text across InterMask, HiTMM, InterMoE and our HHI-Dyn. The study involved 20 users and 30 out-of-distribution text prompts. For each prompt, every model generated the corresponding interactive motion sequence, and each user rated the generated motion sequence of each model on a scale from 1 to 5, where a higher score indicates better performance. As shown in Figure 4, HHI-Dyn achieves the best performance on both Interaction Quality (4.06) and Adherence to Text (4.30), highlighting the effectiveness of our model in producing high-quality, text-aligned interactive motions. More details are provided in Appendix E.

4.3 Further Evaluation

Main Ablation. Table 3 reports an ablation study on the individual contributions of the Interactive Expert module and the Causal Features mechanism in HHI-Dyn. The Causal Encoding preserve temporal causality in latent representations, enabling effective encoding of sequential motion, while the Interactive Expert leverages text and motion dynamics to enhance inter-person coordination and semantic plausibility. The full model consistently outperforms all ablated variants across various metrics, with notable gains in R-Precision and FID.

Ablation Study Results on Spatial Configurations. Table 4 presents an ablation study evaluating key spatial configurations of our proposed STC-VAE. Specifically, we compare three variants: Global Spatial, which relies solely on global spatial cues; Local Spatial, which uses only local spatial information; and Global-Local Fusion, which adaptively integrates both. The experimental results demonstrate that the Global-Local Fusion configuration consistently outperforms the other two vari-

Table 4: Ablation study on the InterHuman test set comparing different **spatial information configurations** for both reconstruction and generation tasks.

	Reconstruction					Generation				
	R-Precision ↑			FID ↓	MM-Dist ↓	R-Precision ↑			FID ↓	MM-Dist ↓
	Top-1	Top-2	Top-3			Top-1	Top-2	Top-3		
Ground Truth	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$
Global Spatial	$0.406^{\pm.005}$	$0.581^{\pm.004}$	$0.682^{\pm.004}$	$0.765^{\pm.013}$	$3.792^{\pm.001}$	$0.501^{\pm.006}$	$0.654^{\pm.005}$	$0.730^{\pm.004}$	$5.898^{\pm.068}$	$3.766^{\pm.001}$
Local Spatial	$0.405^{\pm.006}$	$0.585^{\pm.003}$	$0.687^{\pm.004}$	$0.667^{\pm.013}$	$3.791^{\pm.001}$	$0.494^{\pm.004}$	$0.648^{\pm.003}$	$0.726^{\pm.003}$	$4.814^{\pm.057}$	$3.772^{\pm.001}$
Global-Local Fusion	$\mathbf{0.409}^{\pm.005}$	$\mathbf{0.587}^{\pm.003}$	$\mathbf{0.688}^{\pm.004}$	$\mathbf{0.641}^{\pm.015}$	$\mathbf{3.790}^{\pm.001}$	$\mathbf{0.515}^{\pm.005}$	$\mathbf{0.669}^{\pm.003}$	$\mathbf{0.744}^{\pm.004}$	$\mathbf{4.658}^{\pm.053}$	$\mathbf{3.765}^{\pm.001}$

Table 5: Ablation study results on the InterHuman test set to verify key components of our **Inter-DiT**.

Methods	R-Precision Top-1 ↑	FID ↓	MM-Dist ↓
Interactive Expert			
w/o Motion Router	$0.312^{\pm.004}$	$5.285^{\pm.085}$	$3.846^{\pm.001}$
w/o Text Router	$0.509^{\pm.004}$	$5.000^{\pm.056}$	$3.766^{\pm.001}$
w/o Partner Router	$0.497^{\pm.004}$	$5.146^{\pm.090}$	$3.771^{\pm.001}$
w/o Cosine Similarity	$0.506^{\pm.005}$	$4.886^{\pm.069}$	$3.768^{\pm.001}$
Number of Experts			
4	$0.508^{\pm.004}$	$4.947^{\pm.064}$	$3.769^{\pm.001}$
16	$0.481^{\pm.004}$	$4.923^{\pm.063}$	$3.774^{\pm.001}$
Ours Full (8 Experts)	$\mathbf{0.515}^{\pm.005}$	$\mathbf{4.658}^{\pm.053}$	$\mathbf{3.765}^{\pm.001}$

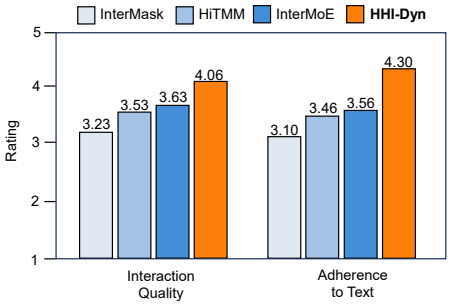


Figure 4: User Study comparing our **HHI-Dyn** with sota methods.

ants in both reconstruction and generation tasks. These findings indicate that incorporating both global and local spatial information not only enables more accurate motion reconstruction but also yields a more robust latent representation, facilitating the generation of high-fidelity motions.

Ablation Study Results on Inter-DiT. We conduct an ablation study on Inter-DiT regarding the Interactive Expert and the number of experts, as reported in Table 5. The full HHI-Dyn model with the complete Interactive Expert outperforms all simplified versions, particularly in terms of R-Precision and FID. Notably, removing any key component of the Interactive Expert leads to performance degradation, validating our hypothesis that partner information and interaction type are crucial for human interactive motion generation. As for the number of experts, too few experts fail to capture the full range of motion patterns, while an excessive number may introduce redundancy.

5 Conclusion

We propose HHI-Dyn, a diffusion-based framework comprising two core components: (1) STC-VAE, a novel and effective architecture that dynamically fuses local and global spatial information to learn more robust latent representations; (2) Inter-DiT, a text-conditioned diffusion transformer that generates motions using interactive experts as its core. The effectiveness of HHI-Dyn is validated through ablation study, user study, quantitative and qualitative results. HHI-Dyn achieves substantial improvements across key metrics, establishing a new state-of-the-art for human-human motion generation.

Limitations and Future Work. Nevertheless, several challenges remain. The current model does not yet ensure physical plausibility over long time horizons, and it also lacks support for the seamless composition of multi-person interactions. Future work will focus on physically grounded objectives (Peng et al., 2018; Dou et al., 2023; Jiang et al., 2023; Luo et al., 2024; Zhang et al., 2024b), finer-grained interaction editing, and multimodal scenarios involving visual, textual, and audio information.

6 AI USE STATEMENT

LLMs were used solely for grammar checking in the preparation of this manuscript. They were not employed in any research activities or in drafting the scientific content.

7 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research is based exclusively on publicly accessible datasets and does not include human subjects, personally identifiable information, or sensitive data. The methods proposed herein are designed solely for the advancement of fundamental machine learning research.

8 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in both the main paper and the supplementary material.

REFERENCES

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5915–5920. IEEE, 2018.

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*, pp. 719–728. IEEE, 2019.

Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *IEEE transactions on visualization and computer graphics*, 29(8):3519–3534, 2022.

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*, pp. 582–592, 2024.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010. IEEE, 2023.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408. Springer, 2024.

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Zichen Geng, Zeeshan Hayder, Wei Liu, Hesheng Wang, and Ajmal Saeed Mian. Armflow: Autoregressive meanflow for online 3d human reaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30718–30728, 2026.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151. IEEE, 2022.

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*, pp. 1900–1910, 2024.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7158–7168. IEEE, 2025.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pp. 180–196. Springer, 2024.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZAyuwJYN8N.

Yifeng Jiang, Jungdam Won, Yuting Ye, and C Karen Liu. Drop: Dynamics responses from human motion prior and projective dynamics. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Zicheng Jiao, Yunlian Sun, Hongwen Zhang, Jinhui Tang, and Massimo Tistarelli. Hitmm: Generative temporal masked modeling of human interactive motions. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Linchao Bao, and Zhenyu He. Audio2gestures: Generating diverse gestures from audio. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

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*, pp. 1–21, 2024.

Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng. Investigating pose representations and motion contexts modeling for 3d motion prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):681–697, 2022.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations*, volume 2024, pp. 56766–56782, 2024.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9489–9497, 2019.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael 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*, pp. 10975–10985, 2019.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

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

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*, pp. 1546–1555, 2024.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9955–9965, 2025.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2in: Leveraging individual information to generate human interactions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1941–1951. IEEE, 2024.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024.

Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15953–15963. IEEE, 2023.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Lipeng Wang, Hongxing Fan, Haohua Chen, Zehuan Huang, and Lu Sheng. Intermoe: Individual-specific 3d human interaction generation via dynamic temporal-selective moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2137–2145, 2026.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7169–7178. IEEE, 2025.

Qingxuan Wu, Zhiyang Dou, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, Lingjie Liu, et al. Text2interact: High-fidelity and diverse text-to-two-person interaction generation. In *International Conference on Learning Representations*, volume 2026, pp. 103856–103875, 2026.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, volume 2024, pp. 28176–28194, 2024.

Yong Xie, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. Recom: Realistic co-speech motion generation with recurrent embedded transformer. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13162–13166. IEEE, 2026.

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*, pp. 22260–22271, 2024.

Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11901–11909, 2026.

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*, pp. 14730–14740, 2023a.

Jinming Zhang, Yunlian Sun, Hongwen Zhang, and Jinhui Tang. Edmg: Towards efficient long dance motion generation with fundamental movements from dance genres. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10447–10456, 2025.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 364–373. IEEE, 2023b.

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*, 46(6):4115–4128, 2024a.

Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2305–2317. IEEE, 2024b.

A Training Loss Detail

A.1 VAE Training Objective

Our VAE is trained with a composite objective comprising reconstruction loss $\mathcal{L}_{\text{rec}}$, joint position loss $\mathcal{L}_{\text{pos}}$, velocity loss $\mathcal{L}_{\text{vel}}$, and KL divergence $\mathcal{L}_{\text{kl}}$:

$$\mathcal{L}_{\text{VAE}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}}, \tag{14}$$

where λ_{rec}, λ_{pos}, λ_{vel}, and λ_{kl} are weighting coefficients.

A.2 Diffusion Training Objective

Our diffusion model is trained with a denoising objective. Given the motion latents $\mathbf{z}_a$ and $\mathbf{z}_b$, we add noise at timestep t to obtain $\mathbf{z}_a^t$ and $\mathbf{z}_b^t$. The denoiser G jointly denoises them conditioned on the timestep t and the text condition c, predicting the clean latents $[\mathbf{z}_a^0, \mathbf{z}_b^0]$. The training loss is computed over the concatenated latents:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{z}_a,\mathbf{z}_b,\epsilon}\left[\left\|[\mathbf{z}_a^0, \mathbf{z}_b^0] - G([\mathbf{z}_a^t, \mathbf{z}_b^t], t, c)\right\|_2^2\right], \tag{15}$$

where $[\cdot,\cdot]$ denotes concatenation, $\mathbf{z}_a^t$ and $\mathbf{z}_b^t$ are the noisy latents at timestep t, and c is the text condition.

B Metrics Detail

B.1 Frechet Inception Distance (FID)

Feature vectors are first extracted from both generated and real motion sequences. The FID is then computed by comparing the distribution of these features between the two sets. This metric is widely adopted as a key indicator for evaluating the overall fidelity and realism of synthesized motions.

B.2 R-Precision

For each generated motion sample, a candidate description pool is constructed, comprising its ground-truth textual label along with 31 randomly selected mismatched descriptions drawn from the test set. For every description in this pool, we compute the Euclidean distance between the motion feature and the corresponding text feature, and subsequently rank these distances in ascending order. The evaluation is performed by measuring the average retrieval accuracy at the top-1, top-2, and top-3 positions. A retrieval is counted as successful if the ground-truth description appears among the top-k ranked candidates; otherwise, it is considered a failure.

B.3 Multimodal Distance (MM-Dist)

The MM-Dist is defined as the mean Euclidean distance computed between the motion feature of each generated sample and the text feature of its associated ground-truth description from the test set. This metric reflects the semantic alignment between motion and language modalities.

B.4 Diversity

To quantify the variability among generated motions, we randomly sample two disjoint subsets of equal size S_d from the entire generated set. Let their corresponding motion feature vectors be denoted as $\{v_1, \ldots, v_{S_d}\}$ and $\{v'_1, \ldots, v'_{S_d}\}$. The diversity score is then calculated as

$$\text{Diversity} = \frac{1}{S_d}\sum_{i=1}^{S_d} \|v_i - v'_i\|_2.$$

In our experiments, the subset size is set to $S_d = 300$.

C Implementation Details

C.1 Model Architecture

In the second stage, both datasets share the same architecture configuration: 8 transformer blocks with 8 attention heads each, a latent space dimension of 768, and the number of experts set to 8. For the second-stage input, the motion representation is reshaped prior to feeding into the transformer. For both the InterHuman and InterX datasets, the input dimensions are reshaped from $B \times T \times 7 \times 64$ to $B \times T \times 448$.

Table 6: Model Architecture Parameters

Parameter	Value
Transformer blocks (stage 2)	8
Attention heads (stage 2)	8
Latent space dimension (stage 2)	768
Number of experts (stage 2)	8
Window size (VAE stage 1)	64
Window stride (VAE stage 1)	10

C.2 Training Details

Our framework is trained on a single NVIDIA RTX 3090 GPU. We employ the AdamW optimizer with betas of $(0.9, 0.999)$ and a weight decay of 2×10^{-5}. The learning rate is scheduled separately for each stage: the first stage (VAE training) uses a maximum learning rate of 2×10^{-4}, while the second stage (denoiser training) uses a maximum learning rate of 1×10^{-4}, both decayed via a cosine schedule with 10 linear warm-up epochs.

Table 7: Dataset-Specific Training Hyperparameters for the VAE and Denoiser

Parameter	InterHuman	InterX
VAE training epochs	100	500
VAE batch size	256	256
VAE maximum learning rate	2×10^{-4}	2×10^{-4}
Denoiser training epochs	1,000	2,000
Denoiser batch size	64	64
Denoiser maximum learning rate	1×10^{-4}	1×10^{-4}

Table 8: General Training Hyperparameters

Parameter	Value
GPU	Single NVIDIA RTX 3090
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight decay	2×10^{-5}
Learning rate schedule	Cosine decay with 10 linear warm-up epochs
Diffusion steps (denoiser)	1,000
DDIM sampling steps	50
CFG weight w	3.5
σ	1×10^{-4}
α	0.5
β	0.3

For the InterHuman dataset, the VAE is trained for 100 epochs with a batch size of 256, while the denoiser is trained for 1,000 epochs with a batch size of 64. For the InterX dataset, the VAE is trained for 500 epochs with a batch size of 256, and the denoiser is trained for 2,000 epochs with a batch

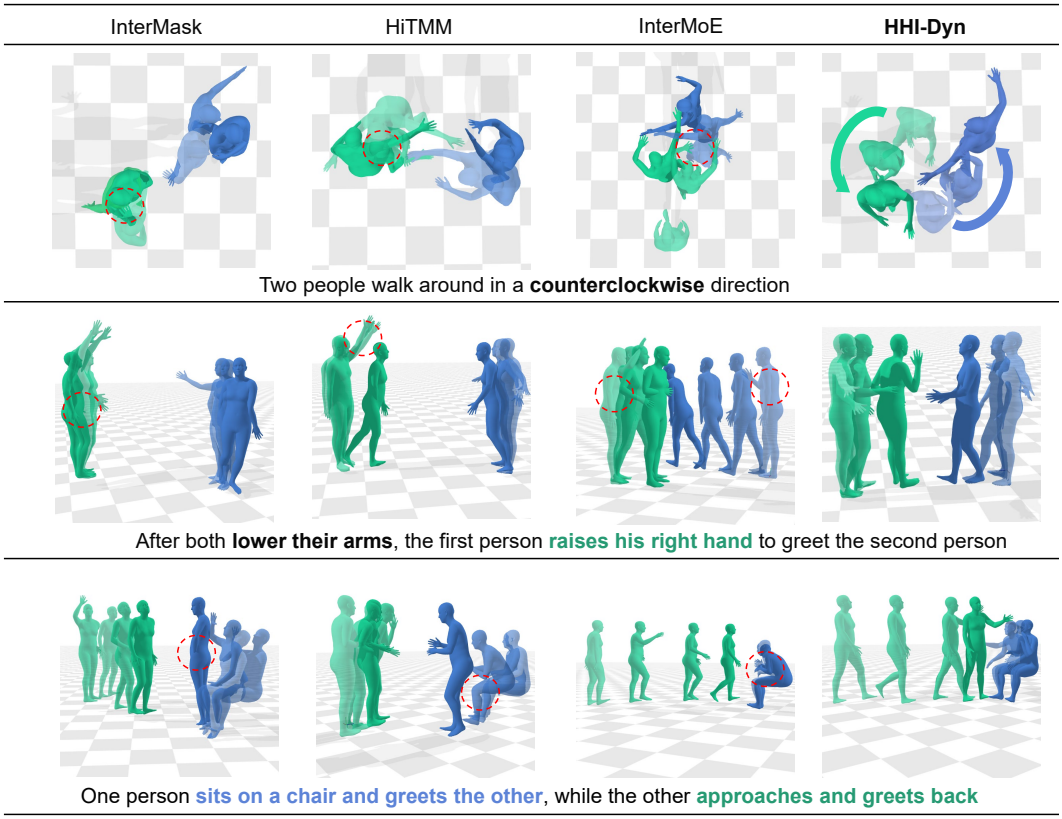


Figure 5: Qualitative results comparing InterMask, HiTMM, InterMoE, and HHI-Dyn.

size of 64. The denoiser is trained with 1,000 diffusion steps, and we employ 50 DDIM sampling steps during inference. Unless otherwise specified, the classifier-free guidance (CFG) weight is set to $w = 3.5$. For other hyper-parameters, we set $\sigma = 1 \times 10^{-4}$, $\alpha = 0.5$, and $\beta = 0.3$.

D DIVERSITY DEMONSTRATION

We conduct a qualitative comparison among InterMask, HiTMM, InterMoE, and HHI-Dyn, as shown in Figure 5. For the prompt "*Two people walk around in a counterclockwise direction*", InterMask, HiTMM, and InterMoE all fail to accurately and smoothly express the counterclockwise direction. For "*After both lower their arms, the first person raises his right hand to greet the second person*", the characters in InterMask and InterMoE fail to lower their arms synchronously at the beginning, while HiTMM incorrectly raises the left hand at the end. Lastly, for "*One person sits on a chair and greets the other, while the other approaches and greets back*", the characters in InterMask and HiTMM incorrectly stand up at the end, while the seated character in InterMoE fails to greet the other person. In contrast, HHI-Dyn correctly generates the corresponding interactive motions across all these scenarios.

We further verify the contribution of the key components through the qualitative comparison illustrated in Figure 6. For the first text prompt, the model trained with global spatial features, the model trained with local spatial features, and the model without the Interactive Expert all generate incorrect kicking orders. They fail to fully capture the dynamics of the interaction, leading to an incorrect text-motion order between the two individuals. Similarly, in the circling scenario, the model trained with global spatial features and the model trained with local spatial features fail to move synchronously, while the model without the Interactive Expert fails to express circling. In contrast, the model trained with global-local fused features and the Interactive Expert accurately captures these spatial dynamics and interactions.

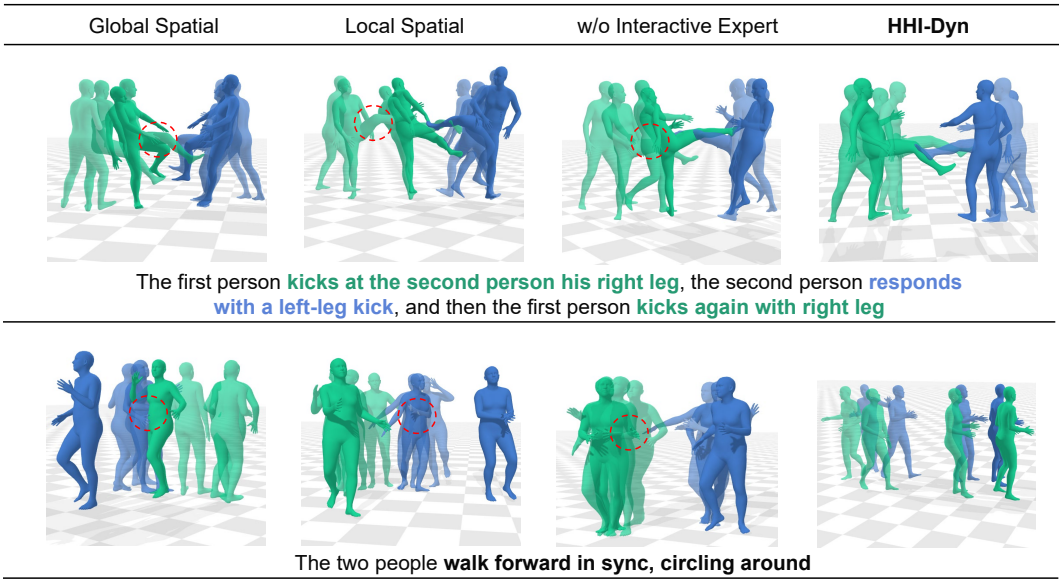


Figure 6: Qualitative results to verify key components of our HHI-Dyn.

E USER STUDY DETAILS

We conducted a user study to evaluate our HHI-Dyn against state-of-the-art methods. For each sample, participants were presented with four animated sequences generated from the same text description by InterGen, InterMask, HiTMM, and HHI-Dyn, respectively. The order of the animations was randomized to avoid potential bias. Participants were asked to rate each animation on a scale from 1 to 5 in terms of interaction quality and text adherence. In addition, they were required to select the better animation based on their overall impression. A total of 30 samples were evaluated, each rated by 20 users. An illustration of the user study procedure is shown in Figure 7.

F HUMAN-HUMAN INTERACTION ATTENTION

F.1 SELF-ATTENTION

The Self-Attention module, illustrated in Figure 8, is designed to process the motion features of a single interacting person. Given the motion features as input, the module first applies an Adaptive Layer Normalization (AdaLN) block, which incorporates conditional information by modulating the features through scaling and shifting operations based on the mean and variance conditioned on the current diffusion timestep as well as the textual information. Subsequently, the normalized features are projected into query, key, and value tensors via three separate linear projection layers. These tensors are then fed into the scaled dot-product attention mechanism proposed by (Vaswani et al., 2017) to compute the attention output. Finally, the resulting attention output is passed through a linear output projection layer to produce the final representation. Formally, given the input motion features $\mathbf{X} \in \mathbb{R}^{T \times D}$, this process can be summarized as:

$$\mathbf{X}' = \text{AdaLN}(\mathbf{X}, t, c), \tag{16}$$

$$\mathbf{Q} = \mathbf{X}'\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}'\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}'\mathbf{W}_V, \tag{17}$$

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{18}$$

$$\mathbf{O} = \mathbf{A}\mathbf{W}_O, \tag{19}$$

where t denotes the current diffusion timestep, c is the textual condition, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$ are learnable projection matrices, d_k is the dimension of the key vectors, and $\mathbf{O}$ is the final output representation.

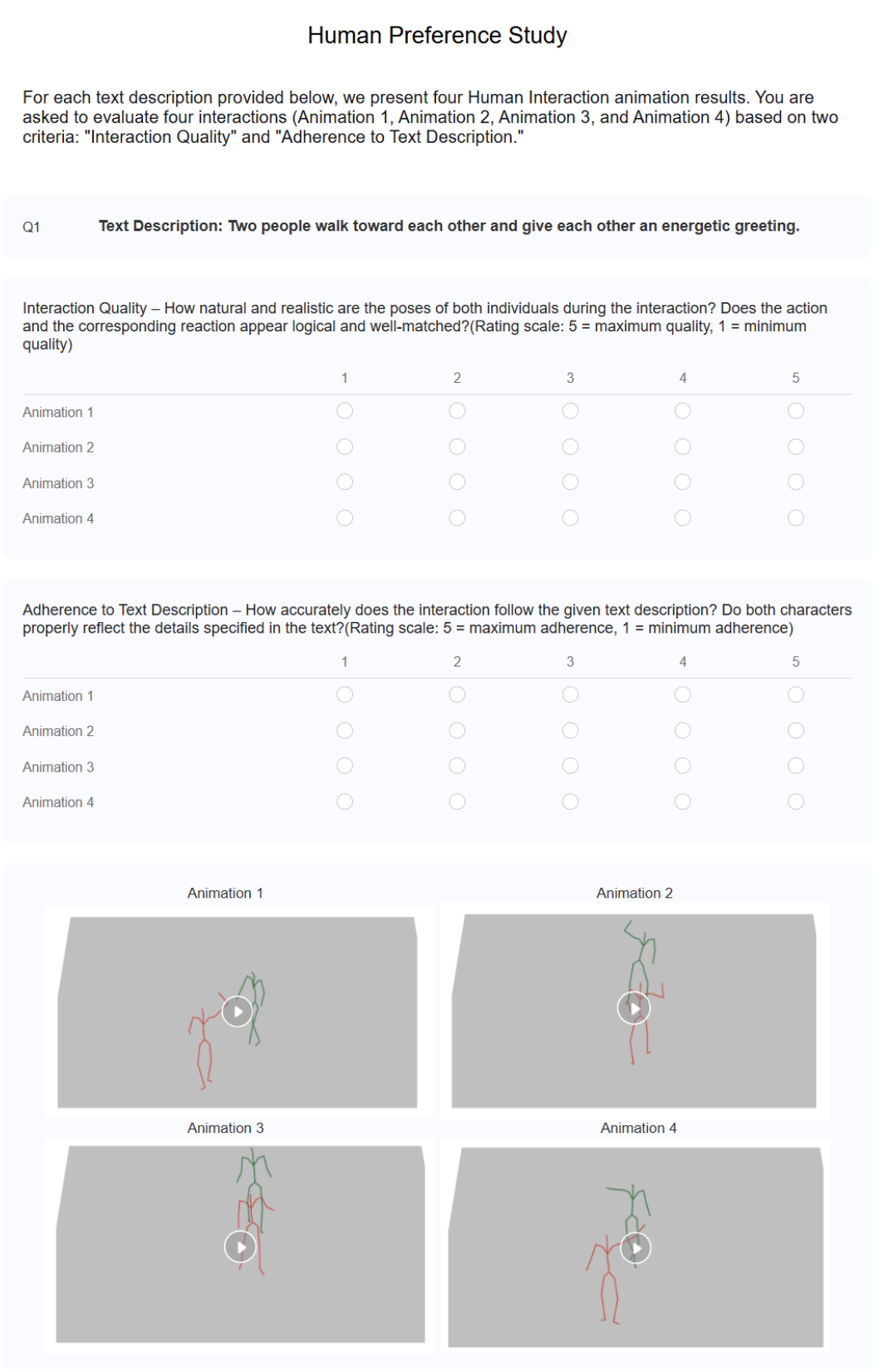


Figure 7: A screenshot of the user study.

F.2 CROSS-ATTENTION

The Cross-Attention module, depicted in Figure 9, is responsible for capturing inter-person interactions. In this module, the motion features from person A and person B are processed in an asymmetric manner to enable bidirectional information exchange. Specifically, the features of both persons are first independently passed through separate AdaLN blocks to incorporate the current diffusion timestep as well as the textual information. The normalized features of person A are then projected via a query projection layer to generate the query tensor, while the normalized features of person B are projected through key and value projection layers to produce the key and value tensors, respectively. These query, key, and value tensors are subsequently fed into the attention mechanism to compute the cross-attention output, which captures the influence of person B on person A. The output is then linearly projected to yield the final cross-attended features. The reverse direction (person B attending to person A) is computed symmetrically in practice. Formally, the cross-attention from B to A can be expressed as:

$$\mathbf{X}'_A = \text{AdaLN}(\mathbf{X}_A, t, c), \quad \mathbf{X}'_B = \text{AdaLN}(\mathbf{X}_B, t, c), \tag{20}$$

$$\mathbf{Q}_A = \mathbf{X}'_A \mathbf{W}_Q^A, \quad \mathbf{K}_B = \mathbf{X}'_B \mathbf{W}_K^B, \quad \mathbf{V}_B = \mathbf{X}'_B \mathbf{W}_V^B, \tag{21}$$

$$\mathbf{A}_{A \leftarrow B} = \text{Softmax}\left(\frac{\mathbf{Q}_A \mathbf{K}_B^\top}{\sqrt{d_k}}\right)\mathbf{V}_B, \tag{22}$$

$$\mathbf{O}_{A \leftarrow B} = \mathbf{A}_{A \leftarrow B} \mathbf{W}_O^A, \tag{23}$$

where t denotes the current diffusion timestep, c is the textual condition, $\mathbf{W}_Q^A$ is the query projection matrix for person A, $\mathbf{W}_K^B$ and $\mathbf{W}_V^B$ are the key and value projection matrices for person B, $\mathbf{W}_O^A$ is the output projection matrix, d_k is the key dimension, and $\mathbf{O}_{A \leftarrow B}$ is the final cross-attended feature for person A.

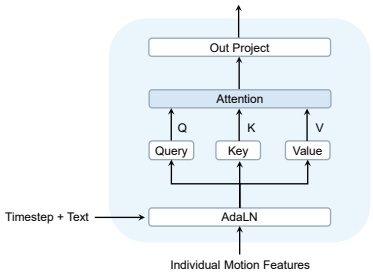


Figure 8: Illustration of the **Self-Attention** module in the Human-Human Interaction Attention Block.

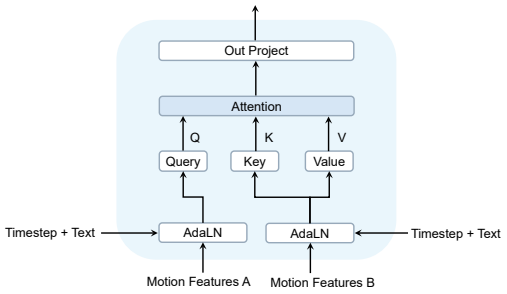


Figure 9: Illustration of the **Cross-Attention** module in the Human-Human Interaction Attention Block.