GLoRe: Detail-Preserving Text-Driven Two-Person Motion In-Betweening via Global Guidance and Local Refinement

Anonymous authors
Paper under double-blind review

Abstract

Motion in-betweening is fundamental to controllable character animation for virtual humans, games, and embodied interaction. While recent two-person motion in-betweening approaches have achieved good performance for generating coordinated interaction between two persons, they often result in degraded individual motion details. We introduce *GLoRe*, a novel *three-stage detail-preserving framework* containing an individual motion initialization stage that leverages a single-person motion in-betweening model to generate detailed individual motions, a global guidance stage that provides global context for motion refinement, and a controlled local refinement stage that adaptively applies residual corrections. Specifically, in the global guidance stage, we propose a new *global interaction condition* that models sequence-level interaction demands from the initialized motions and input constraints; in the controlled local refinement stage, we introduce a new *local residual gate* that regulates the person- and frame-level corrections based on cross-person evidence. Extensive experiments on InterHuman and Inter-X demonstrate that our method achieves superior performance on text-driven two-person motion in-betweening, effectively balancing individual motion quality and interaction coordination.

1 Introduction

Human motion in-betweening aims to generate complete motion sequences from sparse observations and has become a fundamental problem for controllable character animation, virtual humans, games, and embodied interaction (Cohan et al., 2024). Existing single-person motion in-betweening methods have made promising progress in recovering detailed and temporally coherent human motions from sparse constraints (Harvey et al., 2020; Qin et al., 2022; Zhang et al., 2025). However, extending these capabilities to two-person interactions remains challenging.

Recent advances in text-driven two-person motion generation have further enabled the synthesis of complex human interactions from language descriptions (Liang et al., 2024; Wang et al., 2025; Ruiz-Ponce et al., 2026; Jiang et al., 2023). Nevertheless, directly adapting these methods to the motion in-betweening setting may sacrifice fine-grained individual motion details, making it challenging to balance interaction coordination and individual motion quality.

To address this challenge, we introduce GLoRe, a novel three-stage detail-preserving framework that decomposes the task into single-person motion initialization, global interaction condition construction, and local residual refinement stages as illustrated in Figure 1. Our approach first synthesizes individual motions for each character independently using a single-person motion in-betweening model. The initialized motions then undergo a global-to-local refinement process. Specifically, we propose a Global Interaction Condition (GIC) that provides sequence-level interaction guidance, and introduce a Local Residual Gate (LRG) that adaptively controls the refinement strength while preserving reliable motion details. GIC integrates initialized pairwise motions with text and keyframe constraints to model the interaction adjustments required by the current motion state. Based on this global guidance, LRG leverages cross-person evidence to regulate person- and frame-level residual corrections, enabling stronger refinement in poorly coordinated regions while suppressing unneces-

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.

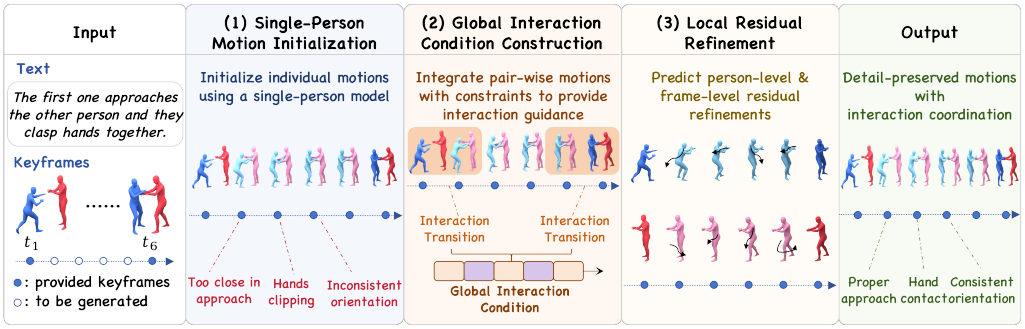


Figure 1: We propose *GLoRe*, a novel three-stage framework for text-driven two-person motion in-betweening. Given text descriptions and keyframes, it initializes individual motions, constructs global guidance, and performs local residual refinement to improve interaction coordination while preserving motion details.

sary modifications elsewhere. Through this global-to-local refinement strategy, GLoRe effectively maintains individual motion quality while improving interaction coordination.

Extensive experiments on InterHuman and Inter-X demonstrate that GLoRe effectively resolves the trade-off between individual motion quality and interaction coordination, achieving superior performance over existing approaches while preserving temporal stability and constraint consistency.

The contributions of our work are summarized as follows:

- We introduce *GLoRe*, a novel three-stage detail-preserving framework for text-driven two-person motion in-betweening that decomposes the task into single-person motion initialization, global interaction condition construction, and local residual refinement, enabling high-quality individual motions and coordinated interactions under constraints.
- We propose an effective *Global Interaction Condition* and a new *Local Residual Gate*. Global interaction condition provides sequence-level interaction guidance and local residual gate adaptively controls residual corrections for improving interaction coordination while preserving individual motion details.
- We conduct extensive experiments on two public datasets and demonstrate that our method outperforms existing approaches, achieving a better balance between individual motion quality and interaction coordination.

2 Related Work

2.1 Two-Person Motion Synthesis

Two-person motion synthesis aims to generate coordinated motions between two characters under various interaction conditions. Recent approaches mainly formulate this task as a direct generation problem, where individual movements and interpersonal coordination are jointly learned. InterGen (Liang et al., 2024) introduced a diffusion-based framework with mutual attention between two character streams, while InterMask (Javed et al., 2025) explored collaborative masked modeling for interaction generation. Recent methods further investigated temporal interaction modeling, pair-level representations, and autoregressive generation strategies (Wang et al., 2025; Li et al., 2025; Wu et al., 2025; Ruiz-Ponce et al., 2026). Motion-conditioned approaches such as InterFormer (Chopin et al., 2023) and ReMoS (Ghosh et al., 2024) further exploited one character's motion to guide the response of another. Although these approaches improved the quality of synthesized interactions, they focused primarily on modeling pairwise coordination and did not explicitly preserve fine-grained individual motion details. Unlike existing methods that directly synthesize interacting motions, our method explores a different paradigm that preserves individual motion details while refining interpersonal coordination through global guidance and local refinement.

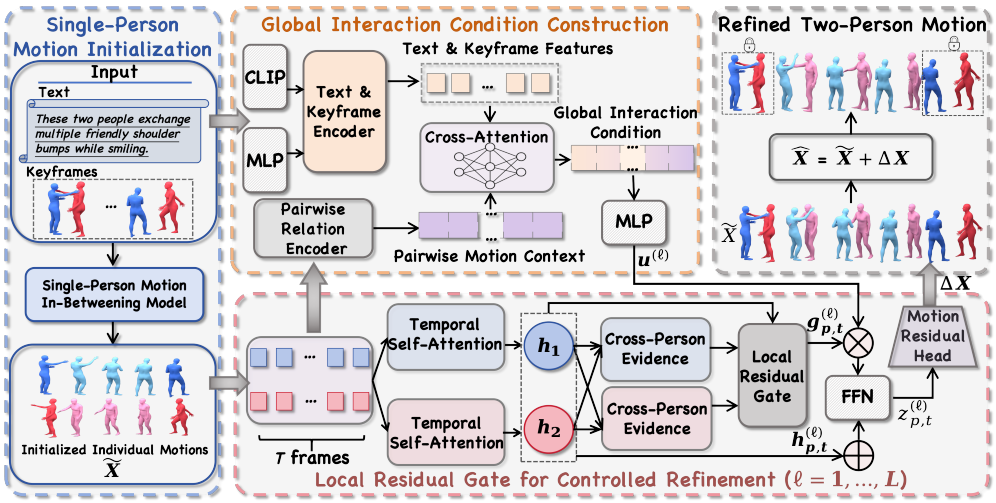


Figure 2: Overview of GLoRe. It first initializes two individual motion sequences, then constructs a Global Interaction Condition from their pairwise motion context and the text-keyframe features, and finally performs local-gated residual refinement. The resulting motion residual is added to the initialized motions, with the input keyframes preserved exactly.

2.2 Motion In-Betweening

Motion in-betweening aims to generate plausible transitions from sparse keyframe constraints while maintaining natural dynamics. Early learning-based methods improved temporal consistency and controllability through phase-based, Transformer-based, and diffusion-based models (Holden et al., 2017; Harvey et al., 2020; Starke et al., 2023; Qin et al., 2022; Yun et al., 2025). CondMDI (Cohan et al., 2024) further enabled text-conditioned motion completion with flexible keyframe constraints. However, extending motion in-betweening to two-person interactions remains challenging. Cross-Space In-Betweening (Zhang et al., 2025) explored long-horizon two-person interaction in-betweening from paired keyposes, while existing methods mainly focused on interaction modeling rather than preserving individual motion details. In contrast, our method adopts a three-stage detail-preserving framework to balance individual motion quality and interaction coordination.

2.3 Motion Refinement

Motion refinement methods improve existing motion sequences by predicting corrections rather than generating motions from scratch. Coarse-to-fine approaches such as MEVA (Luo et al., 2020) and MoMask (Guo et al., 2024), as well as motion editing methods such as MotionFix (Athanasiou et al., 2024) and OmniME (Shi et al., 2026), demonstrate the effectiveness of residual correction for preserving reliable motion information. These studies highlighted the effectiveness of correction-based refinement for maintaining existing motion quality while introducing desired modifications. Inspired by this idea, our method treats two-person motion in-betweening as an interaction refinement problem, where initialized individual motions are preserved and refined through interaction-aware residual corrections.

3 Method

3.1 Problem Formulation

We study text-driven two-person motion in-betweening. Given an interaction description s, a target sequence length T, and sparse paired keyframes, the goal is to generate a complete two-person motion sequence that satisfies both individual motion quality and interpersonal coordination requirements. Let

$$\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2] \in \mathbb{R}^{2\times T\times D} \tag{1}$$

denote a two-person motion sequence, where $\mathbf{X}^p = [\mathbf{x}_1^p, \ldots, \mathbf{x}_T^p] \in \mathbb{R}^{T \times D}$ represents the motion sequence of person $p \in \{1, 2\}$ and $\mathbf{x}_t^p \in \mathbb{R}^D$ denotes the motion feature at frame t.

The sparse paired keyframes are represented as

$$\mathcal{K} = \{(\tau_i, \mathbf{k}_i)\}_{i=1}^K, \qquad \mathbf{k}_i = [\mathbf{k}_i^1, \mathbf{k}_i^2] \in \mathbb{R}^{2 \times D}, \tag{2}$$

where $1 \leq \tau_1 < \cdots < \tau_K \leq T$. Each paired keyframe specifies the complete motion states of both characters at the same timestamp.

The task is to learn a conditional generator

$$\widehat{\mathbf{X}} = \mathcal{G}_\theta(s, \mathcal{K}, T) \in \mathbb{R}^{2 \times T \times D}, \tag{3}$$

where the observed keyframes are exactly preserved in the generated sequence. For unconstrained frames, the generated motion should be consistent with both the interaction description and the sparse keyframes, while maintaining fine-grained individual motion details and coordinated inter-person behaviors.

3.2 Method Overview

To address the formulated problem, we propose GLoRe, a novel framework for detail-preserving text-driven two-person motion in-betweening. As illustrated in Figure 2, GLoRe follows a three-stage pipeline. First, a text-driven single-person motion in-betweening model independently initializes the motion of each character, producing initialized individual motions with detailed body motion and plausible temporal dynamics. These motions provide reliable individual motion content to be preserved and serve as the basis for subsequent interaction refinement. Second, GLoRe constructs a Global Interaction Condition by combining the interaction information contained in the initialized motions with the target constraints from text and paired keyframes. GIC identifies the interaction-related adjustment required by the current motion state and provides global guidance for refinement. Third, GLoRe performs controlled local refinement, where Local Residual Gates adaptively regulate the correction strength across persons and frames, enabling stronger refinement in poorly coordinated regions while preserving reliable motion details. The final motion is obtained by adding the predicted residuals to the initialized motions, with the input keyframes exactly preserved.

3.3 Stage 1: Single-Person Motion Initialization

We first initialized each character motion using a single-person motion in-betweening model. Specifically, we retrained CondMDI (Cohan et al., 2024) on the person-wise motion streams extracted from the two-person training data since it can handle both text and keyframe constraints. Given the interaction description s and the corresponding keyframes, the shared model $\mathcal{P}_\phi$ independently generated an initialized motion for each person:

$$\overline{\mathbf{X}}^p = \mathcal{P}_\phi\left(s, \mathcal{K}^p, T\right), \qquad p \in \{1, 2\}, \tag{4}$$

where $\overline{\mathbf{X}}^p$ denotes the initialized motion before coordinate alignment. This stage provides detailed individual motions with plausible temporal dynamics while preserving the given keyframe constraints.

Following InterGen (Liang et al., 2024), we represented the two motions in a shared pair-centric coordinate frame to preserve their relative spatial configuration (Athanasiou et al., 2023). Person 1 is used as the reference, and Person 2 is aligned through the initial relative yaw and horizontal translation:

$$\widetilde{\mathbf{X}}^2 = \mathcal{T}_{2\to1}(\overline{\mathbf{X}}^2), \qquad \widetilde{\mathbf{X}} = [\overline{\mathbf{X}}^1, \widetilde{\mathbf{X}}^2]. \tag{5}$$

The same transformation is applied to the corresponding keyframes, while preserving relative translation and orientation between the two characters. The resulting Initialized Individual Motions retain fine-grained body details and serve as the motion basis for subsequent interaction refinement.

3.4 Stage 2: Global Interaction Condition Construction

Given the Initialized Individual Motions $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}^1, \widetilde{\mathbf{X}}^2]$, GLoRe constructed a global interaction condition to provide sequence-level guidance for interaction refinement. The key idea is to compare

the current pairwise motion state with the desired interaction constraints, and construct a global condition that describes the interaction adjustments required for subsequent refinement.

Pairwise Motion Context. The initialized motions provide the current motion states of both characters. We first embedded synchronized motion frames into a shared latent space:

$$\mathbf{m}_t^p = f_{\text{mot}}(\widetilde{\mathbf{x}}_t^p), \qquad p \in \{1, 2\}. \tag{6}$$

Since both persons are encoded in the same motion space, their latent representations can be directly compared to characterize their relative motion states. Following pairwise feature comparison designs in metric learning and multimodal representation learning (Wada et al., 2024), we constructed frame-level pairwise descriptors:

$$\boldsymbol{\rho}_t = [\mathbf{m}_t^1 + \mathbf{m}_t^2; \mathbf{m}_t^2 - \mathbf{m}_t^1; |\mathbf{m}_t^2 - \mathbf{m}_t^1|; \mathbf{m}_t^1 \odot \mathbf{m}_t^2; \phi(\xi_t)]. \tag{7}$$

The combined descriptors preserve the overall pair state, directional differences between characters, discrepancy magnitude, and feature-level compatibility between the two motions. A temporal encoder followed by attention pooling (Lee et al., 2019; Vaswani et al., 2017) aggregates these frame-level descriptors into a compact Pairwise Motion Context:

$$\mathbf{c}_{\text{pair}} = \text{Pool}\left(\{f_{\text{rel}}(\boldsymbol{\rho}_t)\}_{t=1}^T\right). \tag{8}$$

The resulting context summarizes the current pairwise motion state and serves as the reference for subsequent interaction refinement.

Text and Keyframe Features. The interaction text and the keyframes provide the target constraints that the refined motion should satisfy. We encode the text description into text features using CLIP (Radford et al., 2021) and embed each paired keyframe with its temporal position and person identity:

$$\mathbf{a}_{\text{text}} = f_{\text{text}}(E_{\text{CLIP}}(s)), \qquad \mathbf{a}_i^p = f_{\text{kf}}([\mathbf{k}_i^p; \xi_i; \pi_p]). \tag{9}$$

A lightweight Transformer-based Text and Keyframe Encoder (Vaswani et al., 2017) aggregates these features into a unified constraint representation:

$$\mathbf{A} = \text{Enc}_A([\mathbf{a}_{\text{text}}; \{\mathbf{a}_i^p\}]). \tag{10}$$

The resulting Text and Keyframe Features encode the semantic descriptions and prescribed motion states that define the desired interaction. Their combination allows GIC to distinguish between existing motion content that should be preserved and interaction aspects that require correction.

Global Interaction Condition. The Pairwise Motion Context describes the current initialized motions, while the Text and Keyframe Features specify the target interaction constraints. Considering only the initialized motions cannot determine whether the current coordination satisfies the intended interaction, while relying only on external constraints ignores the motion details that should be preserved. Therefore, we jointly consider the current motion state and target constraints to construct a global refinement condition. We fused them through cross-attention to construct the GIC:

$$\mathbf{S}_{\text{GIC}} = \text{LN}(\mathbf{c}_{\text{pair}} + \text{CA}(\mathbf{c}_{\text{pair}}, \mathbf{A})). \tag{11}$$

Using the current motion context as the query allows GIC to selectively retrieve constraint information relevant to the initialized interaction state. GIC provides a sample-specific, sequence-level condition for interaction refinement by integrating the current motion state with the target constraints. Determining what interaction-related adjustments should be considered by the refiner, while the following stage determines where and how strongly these corrections are applied.

3.5 Stage 3: Local Residual Gate for Controlled Refinement

Temporal Modeling and Cross-Person Evidence. The initialized individual motions are encoded into person-wise frame tokens with temporal, keyframe, and identity information. In each refinement block, temporal self-attention independently models the motion evolution of each character, while cross-person attention extracts complementary information from the other character to obtain Cross-Person Evidence. Instead of directly updating motion features, this evidence is only used to estimate local correction requirements, providing relational cues for refinement while preserving the initialized motion details.

Local Residual Gate. Although GIC provides sequence-level guidance, the required correction varies across persons and frames. LRG therefore predicts a scalar gate based on the current motion state and Cross-Person Evidence:

$$g_{p,t}^{(\ell)} = g_{\max}\sigma(f_g(\mathbf{h}_{p,t}^{(\ell)}, \mathbf{e}_{p,t}^{(\ell)})), \tag{12}$$

where $\mathbf{e}_{p,t}^{(\ell)}$ denotes the Cross-Person Evidence and $g_{p,t}^{(\ell)} \in [0, g_{\max}]$ controls the local correction strength. Different from uniform refinement, LRG adaptively allocates stronger corrections to poorly coordinated regions while preserving reliable individual motions. This separation prevents interaction evidence from overwriting character-specific motion details.

Controlled Motion Reconstruction. In each refinement block, the Global Interaction Condition is transformed by an MLP into a latent correction signal $\mathbf{u}^{(\ell)}$. This signal represents the global correction content, while LRG determines its local strength:

$$\mathbf{y}_{p,t}^{(\ell)} = \mathbf{h}_{p,t}^{(\ell)} + g_{p,t}^{(\ell)}\mathbf{u}^{(\ell)}. \tag{13}$$

Here, $\mathbf{u}^{(\ell)}$ specifies what interaction-related correction information is available, and $g_{p,t}^{(\ell)}$ determines where and how strongly it is applied. The updated representations are processed by the block-specific feed-forward network (FFN) layers. After the final refinement block, the hidden representations $\mathbf{Z}^{(L)}$ are decoded by a Motion Residual Head:

$$\Delta\mathbf{X} = f_{\Delta}(\mathbf{Z}^{(L)}), \qquad \widehat{\mathbf{X}} = \widetilde{\mathbf{X}} + \Delta\mathbf{X}. \tag{14}$$

The input keyframes are restored exactly after residual addition.

4 Experiments

4.1 Experimental Setup

Table 1: Quantitative comparison on InterHuman and the Inter-X in terms of motion in-betweening quality and interaction coordination quality. ↓/↑ indicate that lower/higher is better, respectively. We performed evaluations 10 times. ± indicates a 95% confidence interval. The metrics are computed only over the unobserved frames.

Dataset	Method	L2P ↓	L2Q ↓	NPSS ↓	FID ↓	CDA@5cm↑	ROA@15° ↑
Inter Human	Real	–	–	–	–	$1.000^{\pm.000}$	$1.000^{\pm.000}$
	CondMDI-Concat (Cohan et al., 2024)	$0.776^{\pm.001}$	$1.271^{\pm.011}$	$1.837^{\pm.011}$	$1.693^{\pm.008}$	$0.379^{\pm.103}$	$0.546^{\pm.102}$
	InterGen(Liang et al., 2024)	$\mathbf{0.688}^{\pm.002}$	$1.291^{\pm.002}$	$2.490^{\pm.016}$	$1.215^{\pm.021}$	$0.391^{\pm.164}$	$0.640^{\pm.099}$
	TIMotion(Wang et al., 2025)	$4.912^{\pm.012}$	$2.680^{\pm.006}$	$3.667^{\pm.013}$	$2.177^{\pm.021}$	$0.214^{\pm.091}$	$0.125^{\pm.063}$
	Interact2Ar(Ruiz-Ponce et al., 2026)	$1.400^{\pm.005}$	$1.397^{\pm.002}$	$2.206^{\pm.011}$	$1.944^{\pm.027}$	$0.315^{\pm.147}$	$0.444^{\pm.121}$
	Ours	$0.700^{\pm.004}$	$\mathbf{1.024}^{\pm.006}$	$\mathbf{0.634}^{\pm.007}$	$\mathbf{0.497}^{\pm.011}$	$\mathbf{0.491}^{\pm.110}$	$\mathbf{0.706}^{\pm.088}$
Inter-X	Real	–	–	–	–	$1.000^{\pm.000}$	$1.000^{\pm.000}$
	CondMDI-Concat(Cohan et al., 2024)	$1.699^{\pm.004}$	$1.186^{\pm.002}$	$0.921^{\pm.004}$	$1.508^{\pm.028}$	$0.273^{\pm.233}$	$0.380^{\pm.191}$
	InterGen(Liang et al., 2024)	$1.968^{\pm.006}$	$1.141^{\pm.002}$	$0.792^{\pm.007}$	$1.457^{\pm.011}$	$\mathbf{0.627}^{\pm.285}$	$0.605^{\pm.147}$
	TIMotion (Wang et al., 2025)	$3.485^{\pm.006}$	$1.969^{\pm.002}$	$1.968^{\pm.006}$	$1.492^{\pm.011}$	$0.306^{\pm.152}$	$0.277^{\pm.049}$
	Interact2Ar (Ruiz-Ponce et al., 2026)	$1.871^{\pm.005}$	$1.201^{\pm.003}$	$0.957^{\pm.009}$	$1.311^{\pm.021}$	$0.140^{\pm.030}$	$0.349^{\pm.141}$
	Ours	$\mathbf{1.401}^{\pm.001}$	$\mathbf{1.105}^{\pm.001}$	$\mathbf{0.582}^{\pm.006}$	$\mathbf{0.455}^{\pm.003}$	$0.519^{\pm.188}$	$\mathbf{0.740}^{\pm.103}$

Datasets. To cover a broader range of interaction types, we conducted experiments on InterHuman (Liang et al., 2024) and Inter-X (Xu et al., 2024). The InterHuman dataset contains 7,779 interaction sequences, and Inter-X contains 5793 sequences which we sampled from the whole dataset to complement more interaction categories, each paired with 3 distinct textual annotations. We randomly picked one sentence every epoch during training. Inter-X motions are downsampled from 120 to 30 fps and converted to the same 22-joint, 262-dimensional representation as InterHuman. Sequences longer than 300 frames were cropped into contiguous 300-frame clips, whereas shorter sequences were zero-padded to the same length. A validity mask is maintained throughout training and evaluation so that padded frames do not participate in keyframe sampling, loss computation, or metric evaluation. We trained a separate model on each dataset without cross-dataset mixing.

Table 2: Quantitative comparison on InterHuman and Inter-X in terms of text-conditioned generation quality. ↓/↑ indicate that lower/higher is better, respectively; → indicates that a value closer to real-motion diversity is better.

Dataset	Method	Diversity→	R-Precision ↑			MM-Dist↓	MModality↑
			Top-1	Top-2	Top-3		
Inter Human	Real motion	$7.778^{\pm.008}$	$0.098^{\pm.002}$	$0.173^{\pm.002}$	$0.214^{\pm.003}$	$3.789^{\pm.001}$	–
	CondMDI-Concat(Cohan et al., 2024)	$7.836^{\pm.005}$	$0.083^{\pm.003}$	$0.145^{\pm.003}$	$0.193^{\pm.003}$	$3.805^{\pm.000}$	$0.312^{\pm.009}$
	InterGen (Liang et al., 2024)	$7.819^{\pm.009}$	$0.089^{\pm.005}$	$0.148^{\pm.005}$	$0.206^{\pm.003}$	$3.798^{\pm.001}$	$0.355^{\pm.014}$
	TIMotion (Wang et al., 2025)	$7.693^{\pm.014}$	$\mathbf{0.102}^{\pm.004}$	$0.153^{\pm.003}$	$\mathbf{0.208}^{\pm.004}$	$3.789^{\pm.001}$	$\mathbf{0.405}^{\pm.017}$
	Interact2Ar (Ruiz-Ponce et al., 2026)	$7.742^{\pm.007}$	$0.091^{\pm.002}$	$0.144^{\pm.002}$	$0.196^{\pm.003}$	$3.798^{\pm.001}$	$0.313^{\pm.012}$
	Ours	$\mathbf{7.784}^{\pm.002}$	$0.092^{\pm.002}$	$\mathbf{0.159}^{\pm.002}$	$0.206^{\pm.001}$	$\mathbf{3.785}^{\pm.000}$	$0.326^{\pm.007}$
Inter-X	Real motion	$5.954^{\pm.006}$	$0.006^{\pm.002}$	$0.008^{\pm.001}$	$0.015^{\pm.001}$	$4.149^{\pm.001}$	–
	CondMDI-Concat(Cohan et al., 2024)	$5.653^{\pm.009}$	$0.007^{\pm.001}$	$0.012^{\pm.002}$	$0.016^{\pm.001}$	$4.146^{\pm.001}$	$0.196^{\pm.008}$
	InterGen (Liang et al., 2024)	$\mathbf{5.956}^{\pm.009}$	$0.004^{\pm.001}$	$0.009^{\pm.002}$	$0.014^{\pm.002}$	$4.141^{\pm.001}$	$0.250^{\pm.011}$
	TIMotion (Wang et al., 2025)	$5.812^{\pm.010}$	$0.005^{\pm.001}$	$0.012^{\pm.001}$	$0.018^{\pm.001}$	$4.152^{\pm.001}$	$\mathbf{0.256}^{\pm.011}$
	Interact2Ar (Ruiz-Ponce et al., 2026)	$5.828^{\pm.010}$	$0.005^{\pm.001}$	$0.012^{\pm.001}$	$0.020^{\pm.001}$	$4.143^{\pm.001}$	$0.241^{\pm.008}$
	Ours	$5.940^{\pm.002}$	$\mathbf{0.008}^{\pm.001}$	$\mathbf{0.015}^{\pm.002}$	$\mathbf{0.021}^{\pm.001}$	$\mathbf{4.139}^{\pm.001}$	$0.220^{\pm.006}$

Keyframe Sampling. For a motion sequence with valid length L, we randomly sampled $\lceil \rho L \rceil$ frames as keyframes where $\rho = 0.10$ during training. The sampled indices are renewed across training epochs, producing irregular observation intervals that cover both short and long range transitions. For each interaction pair, the two persons share the same temporal crop and keyframe indices. To improve robustness to missing observations, we independently removed the entire keyframe condition for 10% of the training samples, so that these samples retain only the text condition. During evaluation, we used different fixed numbers of keyframes.

Metrics. To evaluate both motion in-betweening quality and text-conditioned generation quality, we report two complementary groups of metrics. Following standard motion in-betweening protocols (Harvey et al., 2020; Gopalakrishnan et al., 2019), we use L2P, L2Q, NPSS, FID to measure global position error, quaternion-space rotation error, frequency-domain motion discrepancy and motion realism respectively. Additionally, to assess coordination in two-person interactions, we compute Contact Distance Accuracy (CDA) and Relative Orientation Accuracy (ROA), which measure the consistency of inter-person contact distances and relative pelvis orientations respectively. Following InterGen (Liang et al., 2024), we further report R-Precision (Top-1/2/3), Diversity, Multimodality (MModality), and Multimodal Distance (MM-Dist) to evaluate text-motion alignment, and generation diversity. Detailed explanations and calculations of the metrics can be found in the supplementary material.

Implementation Details. We trained our framework on four NVIDIA A800 GPUs. We retrained the single-person motion in-betweening model (Cohan et al., 2024) on the separated single-person motions extracted from each two-person dataset. Further discussion of initializer selection is provided in the supplementary material. For InterHuman dataset, the denoiser was trained for 3000 epochs with batch size of 48. For Inter-X dataset, the denoiser was trained for 1000 epochs with batch size of 256. Then we trained the residual refiner of our work from scratch using AdamW with a constant learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and gradient clipping at 1.0. For InterHuman and Inter-X, we used global batch sizes of 128 with 2×10^4 steps and 3×10^4 steps, respectively. The LRG was initialized to 0.1, bounded by $g_{\max} = 1.0$, and used a temperature of $\tau_g = 1.0$.

4.2 Comparative Experiments

We compared our method with two categories of baselines. First, we constructed CondMDI-Concat by independently completing the two individual motions using CondMDI (Cohan et al., 2024) and directly combining them without interaction refinement, which evaluates whether concatenating individual motion priors can achieve coherent two-person interactions. Second, we compared with three representative two-person text-to-motion methods, InterGen (Liang et al., 2024), TIMotion (Wang et al., 2025), and Interact2Ar (Ruiz-Ponce et al., 2026). Since these methods are originally

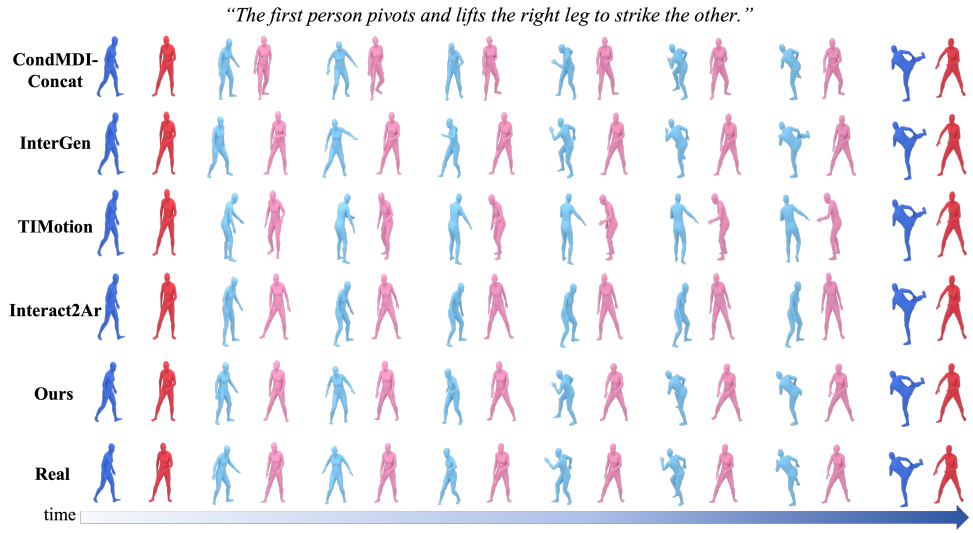


Figure 3: Qualitative comparison with prior state-of-the-art methods on a striking sample. Darker color indicates keyframes. Our method produces smooth and interaction-coordinated in-between keyposes.

designed for text-to-motion generation, we retrained them for text-driven motion in-betweening by providing randomly-sampled 10% keyframes together with the text condition while preserving their original text encoders and interaction modeling architectures.

Quantitative Results. Table 1 and Table 2 demonstrate that our work outperforms existing approaches on text-driven two-person motion in-betweening with five randomly sampled keyframes, achieving superior performance in both motion quality and text-conditioned generation quality evaluations. For motion in-betweening quality, our work achieves better reconstruction and distribution metrics, indicating that the refinement strategy improves motion accuracy and temporal consistency. For interaction coordination quality, the competitive CDA and ROA performance indicates that the refinement strategy enhances spatial coordination and relative orientation consistency between the two characters. For text-conditioned generation quality, our work maintains competitive text-motion alignment and generation diversity, with improved R-Precision and MM-Dist performance. These results validate that our method combines individual motion initialization with interaction refinement and remains robust across diverse interaction patterns.

Qualitative Results. Figure 3 presents a qualitative comparison under the same interaction prompt. Compared with independently generated motions and existing two-person methods, our work produces more coordinated interaction dynamics with clearer individual behaviors. For this case, our work preserves complete striking process, including the preparation, leg lifting and final attacking pose, while maintaining spatial relationship between two characters. In contrast, competing methods tend to generate less coordinated transitions or weaker interaction responses between two persons. These results demonstrate that our work preserves individual motion details while achieving coordinated interactions. Figure 4 further demonstrates the detail-preserving advantage of our work. Compared with existing approaches, it maintains more complete individual motion details during the handover process, such as arm extension and object-transfer posture, while satisfying the interaction requirement.

4.3 Ablation Studies

Ablation of Global Interaction Condition. Table 3 verifies the effectiveness of each component in GIC. Removing the Pairwise Motion Context degrades performance, demonstrating that initialized motion relations provide essential guidance for interaction refinement. Removing text or keyframe features further reduces performance, validating that GIC benefits from jointly modeling current motion relations and external interaction constraints.

Table 3: Ablation studies for Global Interaction Condition and Local Residual Gate on InterHuman.

Module	Method	L2P↓	L2Q↓	NPSS↓	FID↓	CDA@5cm↑	ROA@15° ↑
Global Interaction Condition	w/o GIC	0.917	1.045	0.646	0.522	0.470	0.678
	w/o pairwise motion context	0.761	1.065	0.673	0.515	0.459	0.665
	w/o text & keyframe features	0.841	1.087	0.738	1.253	0.412	0.618
	w/o text features	0.782	1.039	0.645	0.546	0.482	0.698
	w/o keyframe features	0.720	1.048	0.652	0.526	0.476	0.689
Local Residual Gate	w/o global-guidance ($g = 0$)	0.736	1.055	0.653	0.541	0.451	0.654
	w/o LRG ($g = 1$)	0.730	1.158	0.660	0.525	0.473	0.684
	Sequence-level adaptive g	0.718	1.043	0.649	0.513	0.481	0.695
	Self-conditioned local g	0.719	1.033	0.642	0.519	0.485	0.681
	Ours Full	**0.700**	**1.024**	**0.634**	**0.497**	**0.491**	**0.706**

Ablation of Local Residual Gate. We evaluate LRG by replacing it with fixed correction strategies and using sequence-level adaptive or self-conditioned local gates. These comparisons highlight the importance of both local modulation and cross-person conditioning. Uniform and sequence-level gates share correction strength across persons and frames, while self-conditioned gates omit explicit cross-person evidence when determining local strength. LRG adaptively modulates residual corrections according to local cross-person evidence, improving interaction coordination while avoiding unnecessary modifications to reliable individual motions.

Refinement Quality and Computational Cost. Table 4 compares initialization alone, a plain residual refiner, and GLoRe on InterHuman. Both refiners share the same frozen initializer, cached motions, conditioning inputs, losses, and training schedule. GLoRe further reduces FID at a comparable parameter count and training cost, with similar end-to-end inference time. Together with the component ablations, these results support the effectiveness of combining sequence-level interaction guidance with person- and frame-specific control of correction strength. Detailed architecture settings, FLOPs, and stage-wise computational costs are provided in the supplementary material.

Table 4: Quality and computational cost of different refinement strategies on InterHuman.

Method	FID ↓	Params (M)	Training Cost (GPU-hours)	Inference Time (s/pair)
Initialization Only	1.693	235.11	360	32.0
Initialization + Plain Refiner	0.948	266.73	385	33.1
Initialization + GLoRe	**0.497**	267.00	396	32.7

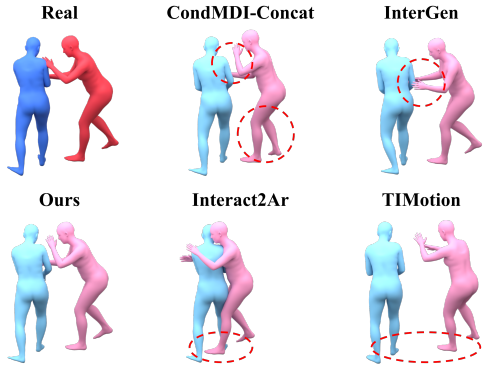


Figure 4: Qualitative comparison under the text prompt *"One person hands an object to the other."* Our method preserves finer individual motion details while generating coordinated interactions.

5 Conclusion

In this work, we proposed GLoRe, a novel three-stage framework for detail-preserving text-driven two-person motion in-betweening by decomposing the task into single-person motion initialization, global interaction condition construction, and local residual refinement. Specifically, we introduced Global Interaction Condition that provides sequence-level interaction guidance and Local Residual Gate to adaptively regulate local corrections, enabling interaction refinement while preserving high-quality individual motions. Extensive experiments on InterHuman and Inter-X demonstrated that GLoRe achieves superior performance over existing approaches, effectively balancing individual motion quality and interaction coordination under constraints.

AI USE STATEMENT

We used generative AI tools for grammar checking, LaTeX formatting, and English phrasing of author-provided submission statements. This assistance was limited to manuscript preparation; generative AI was not used to develop the research methodology, implement the proposed method, generate or analyze experimental data, or interpret the results. The authors are responsible for verifying all AI-assisted revisions and retain full responsibility for the accuracy, originality, and integrity of the final manuscript.

ETHICS STATEMENT

This work uses the publicly available InterHuman and Inter-X datasets. All user study participants provided informed consent, and their preference responses were collected anonymously without collecting personally identifiable information. We acknowledge that generated motions may inherit dataset biases or be misused to create misleading content, and encourage responsible use that respects privacy and dataset usage terms.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we include part of our implementation in the supplementary material as supporting evidence. Upon publication, we will publicly release the complete codebase required to reproduce the results reported in this paper.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9995, 2023.

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 25:8842–8854, 2023.

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–9, 2024.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European conference on computer vision*, pp. 418–437. Springer, 2024.

Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12116–12125, 2019.

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.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion inbetweening. *ACM Transactions on Graphics (TOG)*, 39(4):60:1–60:12, 2020.

Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

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.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Boyuan Li, Xihua Wang, Ruihua Song, and Wenbing Huang. Two-in-one: Unified multi-person interactive motion generation by latent diffusion transformer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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*, 132(9):3463–3483, 2024.

Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

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

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23559–23569, 2026.

Zhenwu Shi, Jingyu Gong, Peiwei Wang, Xingzan Wang, Tianwen Qian, Wenxi Li, Yuan Fang, Jiao Xie, Lizhuang Ma, and Shaohui Lin. Omni-supervised motion editing: Balancing change and invariance through positive-negative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30595–30606, 2026.

Paul Starke, Sebastian Starke, Taku Komura, and Frank Steinicke. Motion in-betweening with phase manifolds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–17, 2023.

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.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13559–13568, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7169–7178, 2025.

Zizhao Wu, Yingying Sun, Yiming Chen, Xiaoling Gu, Ruyu Liu, and Jiazhou Chen. Intermamba: Efficient human-human interaction generation with adaptive spatio-temporal mamba. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

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.

Kwan Yun, Seokhyeon Hong, Chaelin Kim, and Junyong Noh. Anymole: Any character motion in-betweening leveraging video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27838–27848, 2025.

Xiaotang Zhang, Ziyi Chang, Qianhui Men, and Hubert PH Shum. Motion in-betweening for densely interacting characters. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

APPENDIX

A ADDITIONAL METHOD DETAILS

A.1 MOTION REPRESENTATION DETAILS

Following the motion representation used in InterHuman (Liang et al., 2024), we represent each two-person motion sequence as a pair of synchronized motion streams:

$$\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2] \in \mathbb{R}^{2\times T\times D}, \tag{15}$$

where $D = 262$ denotes the motion feature dimension and $\mathbf{X}^p$ represents the motion sequence of person p. Each frame feature $\mathbf{x}_t^p$ contains global root motion, local body pose information, joint velocities, and foot-contact features.

The original motion representation contains both absolute spatial configuration and local articulation information. Specifically, the motion feature consists of global root translation and orientation, local joint rotations, joint positions relative to the root, joint velocities, and contact states. This representation provides sufficient information to reconstruct full-body motions while preserving detailed individual movements.

For two-person interaction modeling, we follow the pair-centric coordinate representation introduced in InterGen (Liang et al., 2024). Instead of representing each character independently in its own global coordinate system, the two motion streams are transformed into a shared coordinate frame. Person 1 is used as the reference character, while Person 2 is aligned according to the initial relative horizontal translation and yaw rotation:

$$\widetilde{\mathbf{X}}^1 = \overline{\mathbf{X}}^1, \qquad \widetilde{\mathbf{X}}^2 = \mathcal{T}_{2\to 1}(\overline{\mathbf{X}}^2), \tag{16}$$

where $\mathcal{T}_{2\to 1}$ denotes the sequence-level rigid transformation from Person 2 to the reference coordinate frame. The same transformation is applied to the corresponding keyframes.

This alignment preserves the relative spatial relationship between the two characters while removing irrelevant global placement differences. Therefore, the subsequent interaction modeling focuses on meaningful relative motion changes, such as distance variation, orientation consistency, and contact coordination, rather than absolute scene locations.

A.2 REFINEMENT BLOCK ARCHITECTURE

The refinement module consists of multiple stacked refinement blocks, which progressively improve the Initialized Individual Motions under the guidance of the Global Interaction Condition (GIC). Each block takes person-wise motion tokens as input and outputs refined hidden representations for subsequent blocks.

Before refinement, the initialized motions $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}^1, \widetilde{\mathbf{X}}^2]$ are converted into frame-level tokens. For each person p and frame t, the initial token is constructed by combining the motion feature with temporal position, person identity, and keyframe information:

$$\mathbf{h}_{p,t}^{(0)} = f_{\text{tok}}(\widetilde{\mathbf{x}}_t^p, \xi_t, \pi_p, m_t), \tag{17}$$

where ξ_t denotes the normalized temporal position, π_p denotes the person identity embedding, and m_t indicates whether the frame is observed.

Each refinement block first models individual temporal dynamics and then extracts interaction-related evidence from the other person. Specifically, temporal self-attention is independently applied to each person stream:

$$\mathbf{H}_p^{(\ell)} = \text{SA}_{\text{temp}}(\mathbf{H}_p^{(\ell-1)}), \quad p \in \{1, 2\}. \tag{18}$$

Then, cross-person attention extracts complementary information from the partner motion:

$$\mathbf{E}_p^{(\ell)} = \text{CA}(\mathbf{H}_p^{(\ell)}, \mathbf{H}_{3-p}^{(\ell)}). \tag{19}$$

The obtained Cross-Person Evidence is only used to estimate local correction requirements and does not directly replace the current motion representations. This design prevents interaction information from overwriting reliable individual motion details.

Algorithm 1 Forward Pass of Refinement Block

Input: Hidden states $\mathbf{H}^{(\ell-1)}$, Global Interaction Condition $\mathbf{S}_{\mathrm{GIC}}$
Output: Updated hidden states $\mathbf{H}^{(\ell)}$

Initialize $\mathbf{H}_p^{(\ell)} \leftarrow \mathbf{H}_p^{(\ell-1)},\ p \in \{1,2\}$
for $p = 1$ to 2 **do**
  $\mathbf{H}_p^{(\ell)} \leftarrow \mathrm{SA}_{\mathrm{temp}}(\mathbf{H}_p^{(\ell-1)})$
end for
for $p = 1$ to 2 **do**
  $\mathbf{E}_p^{(\ell)} \leftarrow \mathrm{CA}(\mathbf{H}_p^{(\ell)}, \mathbf{H}_{3-p}^{(\ell)})$
end for
$\mathbf{u}^{(\ell)} \leftarrow \mathrm{MLP}^{(\ell)}(\mathbf{S}_{\mathrm{GIC}})$
for $p = 1$ to 2 **do**
  for $t = 1$ to T **do**
    $g_{p,t}^{(\ell)} \leftarrow \mathrm{LRG}(\mathbf{h}_{p,t}^{(\ell)}, \mathbf{e}_{p,t}^{(\ell)})$
    $\mathbf{y}_{p,t}^{(\ell)} \leftarrow \mathbf{h}_{p,t}^{(\ell)} + g_{p,t}^{(\ell)}\mathbf{u}^{(\ell)}$
  end for
end for
$\mathbf{H}^{(\ell)} \leftarrow \mathrm{FFN}(\mathbf{Y}^{(\ell)})$
return $\mathbf{H}^{(\ell)}$

Algorithm 2 Local Residual Gate Computation

Input: Motion state $\mathbf{h}_{p,t}^{(\ell)}$, Cross-Person Evidence $\mathbf{e}_{p,t}^{(\ell)}$, maximum gate value $g_{\max}$
Output: Correction strength $g_{p,t}^{(\ell)}$

$\mathbf{r}_{p,t}^{(\ell)} \leftarrow [\mathbf{h}_{p,t}^{(\ell)}; \mathbf{e}_{p,t}^{(\ell)}; \mathbf{h}_{p,t}^{(\ell)} \odot \mathbf{e}_{p,t}^{(\ell)}; \mathbf{h}_{p,t}^{(\ell)} - \mathbf{e}_{p,t}^{(\ell)}]$
$s_{p,t}^{(\ell)} \leftarrow f_g(\mathbf{r}_{p,t}^{(\ell)})$
$g_{p,t}^{(\ell)} \leftarrow g_{\max}\sigma(s_{p,t}^{(\ell)})$
return $g_{p,t}^{(\ell)}$

The Local Residual Gate (LRG) predicts a scalar correction strength for each person and frame. Given the current hidden state and Cross-Person Evidence, the gate is computed as:

$$g_{p,t}^{(\ell)} = g_{\max}\sigma(f_g(\mathbf{h}_{p,t}^{(\ell)}, \mathbf{e}_{p,t}^{(\ell)})), \tag{20}$$

where $g_{p,t}^{(\ell)} \in [0, g_{\max}]$ controls the amount of correction applied at each location. The GIC is projected into a block-specific correction signal:

$$\mathbf{u}^{(\ell)} = \mathrm{MLP}^{(\ell)}(\mathbf{S}_{\mathrm{GIC}}). \tag{21}$$

The hidden states are updated through gated residual refinement:

$$\mathbf{y}_{p,t}^{(\ell)} = \mathbf{h}_{p,t}^{(\ell)} + g_{p,t}^{(\ell)}\mathbf{u}^{(\ell)}. \tag{22}$$

The updated representations are further processed by the feed-forward network (FFN) layers before being passed to the next refinement block.

After the final refinement block, the hidden representations $\mathbf{Z}^{(L)}$ are decoded into a motion-space residual using the Motion Residual Head:

$$\Delta\mathbf{X} = f_{\Delta}(\mathbf{Z}^{(L)}), \qquad \widehat{\mathbf{X}} = \widetilde{\mathbf{X}} + \Delta\mathbf{X}. \tag{23}$$

The observed keyframes are restored exactly after residual addition.

A.3 Pairwise Motion Context Feature Design

The Pairwise Motion Context provides the current interaction-related information from the initialized individual motions. Instead of directly concatenating the two motion streams, we construct explicit pairwise features to capture different aspects of the relationship between two characters.

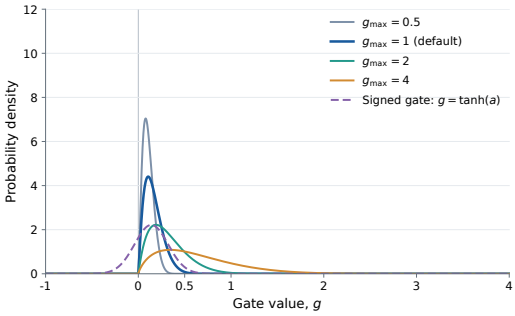


Figure 5: Gate-value distributions with $g_{\max} \in \{0.5, 1, 2, 4\}$ and the signed parameterization $g = \tanh(a)$. The horizontal axis denotes the gate coefficient, and the vertical axis denotes probability density. Each distribution integrates to one.

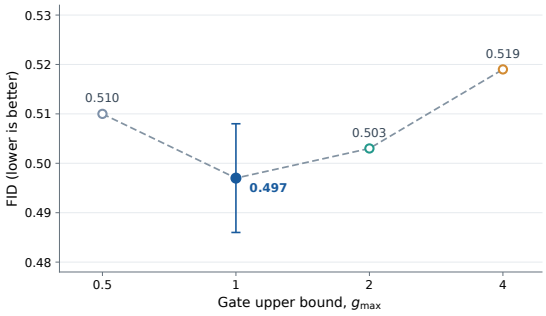


Figure 6: FID sensitivity to the gate upper bound on InterHuman. Lower FID indicates better motion distribution quality.

Given the Initialized Individual Motions $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}^1, \widetilde{\mathbf{X}}^2]$, both motion streams are first projected into a shared latent space:

$$\mathbf{m}_t^p = f_{\text{mot}}(\widetilde{\mathbf{x}}_t^p), \qquad p \in \{1, 2\}. \tag{24}$$

Since the same motion encoder is applied to both persons, the resulting features can be directly compared. Following paired feature comparison designs in metric learning (Wada et al., 2024), we construct the frame-level pairwise descriptor:

$$\boldsymbol{\rho}_t = [\mathbf{m}_t^1 + \mathbf{m}_t^2; \mathbf{m}_t^2 - \mathbf{m}_t^1; |\mathbf{m}_t^2 - \mathbf{m}_t^1|; \mathbf{m}_t^1 \odot \mathbf{m}_t^2; \phi(\xi_t)]. \tag{25}$$

The different components of $\boldsymbol{\rho}_t$ capture complementary pairwise information. Specifically, the summation term describes the overall pair motion state, while the directional difference preserves role-dependent motion discrepancies between the two characters. The absolute difference measures the magnitude of motion inconsistency regardless of character order, and the Hadamard product captures feature-level compatibility between the two motions. The temporal embedding $\phi(\xi_t)$ provides the temporal context of the interaction state.

The frame-level pairwise descriptors are further processed by a temporal relation encoder to model the evolution of interaction-related motion information:

$$\mathbf{R} = [f_{\text{rel}}(\boldsymbol{\rho}_1), \ldots, f_{\text{rel}}(\boldsymbol{\rho}_T)]. \tag{26}$$

An attention pooling operation aggregates the sequence-level relation information into a compact representation:

$$\mathbf{c}_{\text{pair}} = \text{Pool}(\mathbf{R}). \tag{27}$$

The resulting Pairwise Motion Context summarizes the current pairwise motion state and is used by GIC to retrieve target interaction constraints from the Text and Keyframe Features.

To further analyze the contribution of each pairwise feature component, we conduct an additional feature ablation study by removing individual terms from $\boldsymbol{\rho}_t$. The results demonstrate that each component contributes complementary information for constructing the global interaction guidance.

A.4 Analysis of LRG Parameterization

Gate-Value Distributions. Figure 5 shows the distributions of gate coefficients under different parameterizations. For nonnegative gates, the values concentrate towards the lower part of the admissible range, indicating that the model primarily uses attenuated global-guidance injection. Increasing $g_{\max}$ broadens the distribution towards larger coefficients, showing that the additional range is utilized rather than simply remaining available. The signed parameterization additionally assigns probability mass to negative values, allowing individual persons and frames to reverse the shared correction direction. These distributions characterize the strength and sign of the gating coefficients; the resulting latent update also depends on the magnitude of the global correction vector $u^{(\ell)}$.

Sensitivity to the Gate Upper Bound. Figure 6 compares nonnegative gates with $g_{\max} \in \{0.5, 1, 2, 4\}$. The default setting, $g_{\max} = 1$, achieves the lowest mean FID among the evaluated bounds. Reducing the bound to 0.5 increases FID, while expanding it to 2 produces a relatively small change. A further increase to 4 leads to a larger degradation, indicating that a wider coefficient range does not necessarily improve motion distribution quality. Together with the gate-value distributions, these results show that admitting and using larger gate coefficients does not consistently benefit refinement. They support $g_{\max} = 1$ as a practical default within the evaluated setting.

B Baseline Adaptation Details

We evaluate GLoRe against both single-person motion in-betweening and two-person motion generation approaches. Since these baselines are originally designed for different settings, we adapt them to the text-driven two-person motion in-betweening task under a unified keyframe protocol.

For all evaluated methods, training and evaluation follow the same keyframe sampling strategy as GLoRe. During training, 10% of valid frames are uniformly sampled without replacement as observed keyframes. The two persons share identical keyframe timestamps, while their observed motion values correspond to their own 262-dimensional motion features. For 10% of training samples, all keyframe conditions are removed to improve robustness under sparse observations. During evaluation, all methods receive the same fixed number of keyframes. Each observed frame provides the complete motion feature rather than partial joint observations.

All methods are retrained from scratch under this setting. For two-person text-to-motion methods, we preserve their defining interaction modeling modules while augmenting the conditioning and diffusion procedures for sparse motion observations. Specifically, the adapted models incorporate observation-value and observation-mask representations and enforce observed frames during training and sampling.

B.1 Initializer Selection

Our task requires motion completion conditioned jointly on an interaction description and sparse poses observed at randomly sampled timesteps. We select CondMDI (Cohan et al., 2024) because its conditioning interface directly supports text prompts and arbitrary sparse keyframe placement. Representative single-person in-betweening methods, such as Harvey et al. (2020) and Qin et al. (2022), formulate motion completion using motion context and target poses, without text conditioning in their original formulations. Although some methods additionally support pose constraints within a transition, this capability alone does not provide the joint text-and-keyframe conditioning required by our setting. Using these methods as alternative initializers would therefore require a text-conditioning extension and adaptation to our observation protocol. We use CondMDI consistently to preserve the same conditioning interface and assess interaction refinement from a fixed initialization model.

B.2 CondMDI-Concat

CondMDI (Cohan et al., 2024) is originally designed for text-conditioned single-person motion in-betweening. We use it as an individual motion initialization baseline by retraining CondMDI on the individual motion streams extracted from paired two-person motions.

Given a pair-level interaction caption s and person-specific observed motion constraints, the same CondMDI model independently generates two motion streams:

$$\overline{\mathbf{X}}^p = \mathcal{P}_\phi(s, \mathcal{K}^p, T), \qquad p \in \{1, 2\}. \tag{28}$$

The two streams share the same interaction caption and keyframe timestamps, while using their corresponding observed motion values. They are generated independently and stacked in the canonical pair coordinate system without additional interaction refinement:

$$\overline{\mathbf{X}}^{\text{concat}} = \text{Stack}(\overline{\mathbf{X}}^1, \overline{\mathbf{X}}^2). \tag{29}$$

This baseline evaluates whether independent individual motion priors alone can produce coordinated two-person interactions.

B.3 InterGen Adaptation

InterGen (Liang et al., 2024) is originally designed for text-conditioned two-person motion generation through cooperative diffusion models and mutual attention between two character streams. To adapt InterGen to motion in-betweening, we add a frame-level observation interface while preserving its original cooperative denoisers and interaction modeling blocks.

For each person, the hidden input representation is augmented with observed motion features and binary observation masks:

$$\mathbf{h}_t^p = E_x(\mathbf{x}_t^p) + E_o(m_t \mathbf{c}_t^p) + E_m(m_t), \tag{30}$$

where E_o and E_m denote the observation-value and observation-mask embeddings, respectively. During diffusion training and sampling, observed frames are enforced through hard replacement:

$$\mathbf{x}_t \leftarrow (1 - \mathbf{m}) \odot \mathbf{x}_t + \mathbf{m} \odot \mathbf{c}. \tag{31}$$

The original cooperative denoisers and mutual attention mechanism are kept unchanged, while the additional observation interface enables sparse keyframe completion.

B.4 TIMotion Adaptation

TIMotion (Wang et al., 2025) introduces temporal and interaction modeling through Causal Interactive Injection, Role-Evolving Scanning, and Localized Pattern Amplification. We retain these core interaction modules and adapt the model for motion in-betweening by introducing sparse observation conditioning.

Specifically, observed motion values and binary masks are projected into the latent space through additional embeddings. Similar to the InterGen adaptation, the observed frames are enforced during diffusion training and sampling through hard replacement. This allows TIMotion to complete missing motion segments while preserving its original temporal and interaction modeling design.

B.5 Interact2Ar Adaptation

Interact2Ar (Ruiz-Ponce et al., 2026) is an autoregressive diffusion framework that generates human-human interactions through a mixed-memory autoregressive rollout. To adapt it to motion in-betweening, we preserve its defining autoregressive diffusion and pair-interaction mechanism while modifying the representation-specific output branches.

The original Interact2Ar is designed for SMPL-X full-body and hand representations. Since our experiments use the InterHuman 22-joint 262-dimensional representation, we replace the SMPL-X-specific output branches with compatible motion prediction heads, including position and velocity features, rotation features, and foot-contact features.

During autoregressive rollout, the model predicts missing intervals between observed keyframes. The observed keyframes are hard-replaced at each diffusion step within every autoregressive chunk, and the training loss is computed only on valid unobserved target frames. This adaptation preserves the original mixed-memory autoregressive interaction modeling while enabling evaluation under the same sparse motion in-betweening setting.

All evaluated baselines therefore operate under the same text, keyframe availability, and motion representation constraints, allowing a fair comparison with GLoRe.

B.6 Cross-Space In-Betweening.

Motion In-Betweening for Densely Interacting Characters (Zhang et al., 2025) introduces Cross-Space In-Betweening for two-character motion completion. It generates transitions autoregressively from observed motion history towards target keyposes. Our setting instead specifies a text description and sparse paired poses at randomly sampled timesteps. The authors' public repository contained the project description and teaser but no training or inference code or pretrained checkpoints. We therefore do not include this method in the quantitative comparison. Its published results also follow a different conditioning protocol and cannot be directly incorporated into our evaluation tables.

ID	Action category	Hits	Clips
6	Posing	35	285
9	Sit on leg	6	240
14	Knock over	30	283
15	Step on foot	19	318
16	High-five	36	294
17	Chase	25	286
18	Whisper in ear	2	288
20	Finger-guessing	21	289
22	Link arms	33	286
23	Shoulder to shoulder	39	286
25	Carry on back	6	282
26	Massage shoulder	2	281
27	Massage leg	0	289
28	Hand wrestling	2	281
30	Pat on cheek	2	296
31	Thumb up	16	268
33	Imitate	45	278
34	Kiss on cheek	4	171
36	Cover mouth	7	282
37	Look back	24	249
39	Fly kiss	0	261
Total: 21 categories			5,793

Table 5: Selected Inter-X categories.

C Inter-X Subset Construction

We construct an Inter-X subset to complement InterHuman with interaction types that are absent or sparsely represented in its textual annotations. Inter-X provides 40 predefined action categories (Xu et al., 2024). We estimate their coverage by searching the InterHuman captions using category-specific action keywords. Matches are counted at the clip level: each clip contributes at most one hit, even if multiple captions match. We retain categories with at most 50 matching clips, including categories with zero matches, yielding the 21 categories listed in Table 5.

For the selected categories, we retained available two-person motion clips with corresponding textual descriptions and preserved the original Inter-X training, validation, and test assignments. The resulting subset contains 5,793 interaction clips, with 4,634 for training, 295 for validation, and 864 for testing.

D Training Objective

GLoRe learns interaction refinement by combining motion fidelity, pair-relation dynamics, and residual regularization. During refiner training, the single-person initialization model is frozen. Let $\overline{\mathbf{X}}$, $\widehat{\mathbf{X}}$, and $\mathbf{X}$ denote the initialized, refined, and ground-truth motion features, respectively. For this section, these tensors use the normalized representation $\mathbb{R}^{B\times 2\times T\times D}$, where B is the minibatch size and $D = 262$. The GIC and LRG modules are optimized jointly through the refined motion.

Let v_{bt} indicate a valid frame and M_{bptd} an observed feature, where $p \in \{1, 2\}$ indexes the two people. We define the supervision sets

$$\Omega = \{(b, p, t, d) : v_{bt} = 1,\ M_{bptd} = 0\}, \tag{32}$$

$$\mathcal{U} = \{(b, p, t) : v_{bt} = 1,\ \max_{d} M_{bptd} = 0\}, \tag{33}$$

$$\mathcal{E} = \{(b, t) : v_{bt} = v_{b,t+1} = 1\}. \tag{34}$$

Here, Ω contains unobserved feature entries, $\mathcal{U}$ contains unobserved person-frame pairs, and $\mathcal{E}$ contains adjacent valid frame pairs. Observed features are preserved by directly replacing the corresponding output entries with the supplied observations. Each loss is averaged over its eligible entries within the minibatch; an empty supervision set contributes zero.

Motion fidelity. We supervise the completed motion in both feature and joint-position spaces. Feature reconstruction is measured on valid, unobserved entries:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{(b,p,t,d)\in\Omega} \left(\widehat{X}_{bptd} - X_{bptd}\right)^2. \tag{35}$$

Let $\mathbf{P}$ and $\widehat{\mathbf{P}}$ denote the $J = 22$ joint positions extracted from the first 66 channels after reversing the feature normalization. Both people are represented in the same pair-centric coordinate system. The joint-position loss is

$$\mathcal{L}_{\text{pos}} = \frac{1}{3J|\mathcal{U}|} \sum_{(b,p,t)\in\mathcal{U}} \sum_{j=0}^{J-1} \left\| \widehat{\mathbf{P}}_{bptj} - \mathbf{P}_{bptj} \right\|_2^2. \tag{36}$$

While feature reconstruction supervises the full motion representation, this term directly constrains spatial pose accuracy in the shared coordinate system.

We additionally anchor supporting feet to reduce horizontal drift. For each of the four foot joints, a valid, unobserved frame is classified as a support frame when its ground-truth contact indicator exceeds 0.5 or its ground-truth height is below 0.05 in the denormalized coordinate units. These frames are partitioned into maximal contiguous segments separately for each sample, person, and foot. Let $\mathcal{S}$ contain the resulting tuples (b, p, f, I), where I is a segment and f is a foot joint. With $t_I = \min I$, its horizontal anchor is $\mathbf{a}_{bpfI} = \mathbf{P}^{xz}_{bpt_I f}$. We minimize

$$\mathcal{L}_{\text{foot}} = \frac{\displaystyle\sum_{(b,p,f,I)\in\mathcal{S}} \sum_{t\in I} \left\| \widehat{\mathbf{P}}^{xz}_{bptf} - \mathbf{a}_{bpfI} \right\|_2^2}{\displaystyle 2 \sum_{(b,p,f,I)\in\mathcal{S}} |I|}. \tag{37}$$

Observed frames and invalid frames break these segments, so each supervised segment uses its own initial ground-truth foot location. Together, the motion-fidelity objective is

$$\mathcal{L}_{\text{mot}} = \mathcal{L}_{\text{rec}} + \alpha_{\text{pos}}\mathcal{L}_{\text{pos}} + \alpha_{\text{foot}}\mathcal{L}_{\text{foot}}. \tag{38}$$

Pair-relation dynamics. To supervise how the interaction evolves, we construct a geometric relation state from the denormalized joint positions at each frame. Suppressing the sample index, let $\mathbf{d}_t = \mathbf{P}_{2,t,0} - \mathbf{P}_{1,t,0}$ be the relative root displacement. Its bearing in the horizontal plane is $\theta_t = \operatorname{atan2}(d_t^x, d_t^z)$. Let ψ_t be the signed angle between the two horizontal forward directions, each derived from the hip joints and the vertical axis. We also compute

$$d_t^{\text{hand}} = \min_{i,j\in\mathcal{H}} \|\mathbf{P}_{1,t,i} - \mathbf{P}_{2,t,j}\|_2, \tag{39}$$

$$d_t^{\text{foot}} = \min_{i,j\in\mathcal{F}} \|\mathbf{P}_{1,t,i} - \mathbf{P}_{2,t,j}\|_2, \tag{40}$$

where $\mathcal{H} = \{20, 21\}$ and $\mathcal{F} = \{7, 8, 10, 11\}$ are the hand and foot joint sets under zero-based indexing. The resulting state is

$$\begin{aligned} \mathbf{r}_t = \big[& d_t^x,\ d_t^z,\ \sin\theta_t,\ \cos\theta_t,\ \|\mathbf{d}_t\|_2, \\ & d_t^{\text{hand}},\ d_t^{\text{foot}},\ \sin\psi_t,\ \cos\psi_t \big]^\top \in \mathbb{R}^9. \end{aligned} \tag{41}$$

The root distance and inter-person joint distances use three-dimensional positions, while root bearing and forward-direction angles are computed in the horizontal plane. Angular quantities are represented by their sine and cosine.

Applying the same construction to the refined motion gives $\widehat{\mathbf{r}}_t$. We match the first-order temporal differences of these states:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{9|\mathcal{E}|} \sum_{(b,t)\in\mathcal{E}} \|(\widehat{\mathbf{r}}_{b,t+1} - \widehat{\mathbf{r}}_{bt}) - (\mathbf{r}_{b,t+1} - \mathbf{r}_{bt})\|_2^2. \tag{42}$$

This term couples the two motion streams by supervising changes in relative placement, orientation, and hand/foot proximity. All adjacent valid frame pairs participate, including transitions into and out of observed keyframes. The differences are computed per frame in the geometric state above.

Residual regularization. The initialized motions provide a reference for retaining individual motion content. We penalize the magnitude of the applied refinement on valid, unobserved feature entries:

$$\mathcal{L}_{\text{res}} = \frac{1}{|\Omega|} \sum_{(b,p,t,d)\in\Omega} \left(\widehat{X}_{bptd} - \overline{X}_{bptd}\right)^2. \tag{43}$$

This regularizer favors smaller departures from the initialized motion, balancing the reconstruction and interaction-dynamics objectives.

Overall objective. The refiner is trained with

$$\mathcal{L}_{\text{GLoRe}} = \mathcal{L}_{\text{mot}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{res}}\mathcal{L}_{\text{res}}. \tag{44}$$

We use $\alpha_{\text{pos}} = 0.5$, $\alpha_{\text{foot}} = 0.1$, $\lambda_{\text{dyn}} = 0.05$, and $\lambda_{\text{res}} = 0.005$ for both InterHuman and the Inter-X.

E Evaluation Metrics

We evaluate GLoRe from two perspectives: motion in-betweening quality and text-conditioned generation quality. Following the official evaluation protocols of LaFAN1 (Harvey et al., 2020) and InterGen (Liang et al., 2024), we report L2P, L2Q, NPSS, FID, Diversity, R-Precision, MM-Dist, and MModality. For each test condition, we generate $R = 10$ motions with different random seeds and average the results over all repetitions.

Given the test set

$$\mathcal{D}_{\text{test}} = \{(c_i, M_i, O_i)\}_{i=1}^{N}, \tag{45}$$

where c_i denotes the text condition, M_i denotes the ground-truth two-person motion, and O_i denotes the observed keyframe set. Metrics for motion reconstruction are computed only on the unobserved frames:

$$\Omega_i = \{t : t \notin O_i,\ t \text{ valid}\}. \tag{46}$$

E.1 Motion In-Betweening Quality

L2P. L2P measures the average global joint position error between generated motions and ground-truth motions. Specifically, it evaluates the Euclidean distance between the 22 global joint positions of the two characters on unobserved frames:

$$\text{L2P} = \frac{\sum_{i=1}^{N}\sum_{t\in\Omega_i} d_{i,t}^{P}}{\sum_{i=1}^{N}|\Omega_i|}, \tag{47}$$

where $d_{i,t}^{P}$ denotes the normalized global position distance at frame t. Lower L2P indicates more accurate spatial reconstruction.

L2Q. L2Q measures the average rotation error between generated and ground-truth motions. Following the official protocol, local rotations are converted into global joint rotations through forward kinematics, and the quaternion distance is computed on unobserved frames:

$$\text{L2Q} = \frac{\sum_{i=1}^{N}\sum_{t\in\Omega_i} d_{i,t}^{Q}}{\sum_{i=1}^{N}|\Omega_i|}, \tag{48}$$

where $d_{i,t}^{Q}$ denotes the quaternion-space rotation distance. Lower L2Q indicates more accurate body orientations and joint poses.

NPSS. NPSS evaluates the temporal dynamics of generated motions by measuring the discrepancy between generated and ground-truth frequency distributions. Different from frame-wise reconstruction errors, NPSS captures motion rhythm and temporal variation through the cumulative power spectrum distance:

$$\text{NPSS} = \frac{\sum_{i,l,d} E_{i,l,d} D_{i,l,d}}{\sum_{i,l,d} E_{i,l,d}}, \tag{49}$$

Table 6: Robustness evaluation under different keyframe conditions on Inter-X in terms of motion in-betweening quality. K denotes the number of keyframes provided during motion in-betweening.

Condition	Method	L2P↓	L2Q↓	NPSS↓	FID↓
$K = 10$	CondMDI-Concat(Cohan et al., 2024)	$1.514^{\pm.003}$	$1.057^{\pm.002}$	$0.808^{\pm.004}$	$1.203^{\pm.013}$
	InterGen(Liang et al., 2024)	$0.825^{\pm.005}$	$1.113^{\pm.002}$	$0.654^{\pm.006}$	$0.370^{\pm.009}$
	TIMotion(Wang et al., 2025)	$3.251^{\pm.005}$	$1.858^{\pm.003}$	$1.892^{\pm.005}$	$0.292^{\pm.005}$
	Interact2Ar(Ruiz-Ponce et al., 2026)	$1.195^{\pm.006}$	$0.921^{\pm.001}$	$0.721^{\pm.003}$	$0.494^{\pm.016}$
	Ours	$\mathbf{0.802}^{\pm.006}$	$\mathbf{0.768}^{\pm.002}$	$\mathbf{0.432}^{\pm.006}$	$\mathbf{0.204}^{\pm.010}$
$K = 20$	CondMDI-Concat(Cohan et al., 2024)	$1.349^{\pm.003}$	$0.943^{\pm.002}$	$0.709^{\pm.003}$	$0.960^{\pm.018}$
	InterGen(Liang et al., 2024)	$0.703^{\pm.005}$	$0.818^{\pm.002}$	$0.709^{\pm.003}$	$0.299^{\pm.007}$
	TIMotion(Wang et al., 2025)	$3.083^{\pm.008}$	$1.731^{\pm.003}$	$1.807^{\pm.005}$	$0.179^{\pm.003}$
	Interact2Ar(Ruiz-Ponce et al., 2026)	$0.933^{\pm.003}$	$0.803^{\pm.001}$	$0.617^{\pm.002}$	$0.351^{\pm.010}$
	Ours	$\mathbf{0.508}^{\pm.005}$	$\mathbf{0.513}^{\pm.003}$	$\mathbf{0.350}^{\pm.004}$	$\mathbf{0.083}^{\pm.002}$

where $D_{i,l,d}$ denotes the Earth Mover Distance between generated and ground-truth cumulative spectra, and $E_{i,l,d}$ denotes the spectral energy of the ground-truth motion. Lower NPSS indicates more consistent temporal dynamics.

FID. FID measures the distribution discrepancy between generated motions and real motions in the evaluator latent space. Following InterGen, motion features are extracted using a pretrained motion evaluator, and Gaussian distributions are fitted to generated and ground-truth motion features. The FID score is computed as:

$$\mathrm{FID} = \|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{1/2}), \tag{50}$$

where μ_g, Σ_g and μ_r, Σ_r denote the means and covariances of generated and real motion distributions, respectively. Lower FID indicates that generated motions have a distribution closer to real motions.

E.2 Text-Conditioned Generation Quality

All text-conditioned metrics are computed using the pretrained InterGen text-motion evaluator, which contains a text encoder E_t and a motion encoder E_m:

$$e_i = E_t(c_i), \qquad z_i = E_m(M_i). \tag{51}$$

R-Precision. R-Precision measures text-motion semantic alignment by evaluating whether the motion corresponding to a text query can be retrieved from a candidate motion set. Given the ranking based on text-motion embedding distances, R-Precision is computed as:

$$R@k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[i \in \mathrm{TopK}(D_i)], \quad k \in \{1, 2, 3\}, \tag{52}$$

where D_i denotes the ranking results for the i-th text condition. Higher R-Precision indicates stronger text-motion consistency.

MM-Dist. MM-Dist measures the average embedding distance between a text condition and its corresponding generated motion in the shared text-motion embedding space:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N} \|e_i - z_i\|_2. \tag{53}$$

Lower MM-Dist indicates better semantic correspondence between text and motion.

Diversity. Diversity measures the variation among generated motions across different test conditions. Following InterGen, randomly sampled motion pairs are compared in the evaluator embedding space:

$$\text{Diversity} = \frac{1}{K}\sum_{k=1}^{K} \left\| \frac{z_{a_k} - z_{b_k}}{2} \right\|_2. \tag{54}$$

Table 7: Robustness evaluation under different keyframe conditions on Inter-X in terms of text-conditioned generation quality. K denotes the number of keyframes provided during motion in-betweening.

Condition	Method	Diversity→	R-Precision ↑			MM-Dist↓	MModality↑
			Top-1	Top-2	Top-3		
$K = 10$	CondMDI-Concat(Cohan et al., 2024)	$5.660^{\pm.009}$	$0.006^{\pm.001}$	$\mathbf{0.013}^{\pm.001}$	$0.017^{\pm.001}$	$4.144^{\pm.001}$	$0.162^{\pm.007}$
	InterGen(Liang et al., 2024)	$5.917^{\pm.009}$	$0.004^{\pm.001}$	$0.009^{\pm.001}$	$0.015^{\pm.002}$	$4.145^{\pm.001}$	$\mathbf{0.195}^{\pm.008}$
	TIMotion(Wang et al., 2025)	$5.828^{\pm.011}$	$0.004^{\pm.001}$	$0.008^{\pm.001}$	$0.015^{\pm.002}$	$4.151^{\pm.001}$	$0.175^{\pm.008}$
	Interact2Ar(Ruiz-Ponce et al., 2026)	$\mathbf{5.923}^{\pm.009}$	$\mathbf{0.007}^{\pm.001}$	$0.011^{\pm.002}$	$0.018^{\pm.001}$	$4.144^{\pm.001}$	$0.165^{\pm.007}$
	Ours	$6.032^{\pm.010}$	$0.006^{\pm.001}$	$0.011^{\pm.001}$	$\mathbf{0.019}^{\pm.001}$	$\mathbf{4.142}^{\pm.001}$	$0.104^{\pm.011}$
$K = 20$	CondMDI-Concat(Cohan et al., 2024)	$5.666^{\pm.009}$	$0.005^{\pm.001}$	$0.010^{\pm.002}$	$0.017^{\pm.001}$	$4.147^{\pm.001}$	$0.134^{\pm.006}$
	InterGen(Liang et al., 2024)	$5.935^{\pm.009}$	$0.004^{\pm.001}$	$0.009^{\pm.001}$	$0.015^{\pm.002}$	$4.143^{\pm.001}$	$\mathbf{0.152}^{\pm.006}$
	TIMotion(Wang et al., 2025)	$5.850^{\pm.004}$	$0.003^{\pm.001}$	$0.008^{\pm.001}$	$0.019^{\pm.001}$	$4.150^{\pm.001}$	$0.092^{\pm.004}$
	Interact2Ar(Ruiz-Ponce et al., 2026)	$\mathbf{5.943}^{\pm.006}$	$0.006^{\pm.001}$	$0.011^{\pm.001}$	$0.018^{\pm.001}$	$4.146^{\pm.001}$	$0.134^{\pm.006}$
	Ours	$5.995^{\pm.001}$	$\mathbf{0.006}^{\pm.001}$	$\mathbf{0.013}^{\pm.001}$	$\mathbf{0.020}^{\pm.002}$	$\mathbf{4.141}^{\pm.001}$	$0.065^{\pm.003}$

Unlike MModality, Diversity evaluates global variation across different conditions rather than multiple generations under the same condition. A value closer to real-motion diversity indicates better coverage of the motion distribution.

MModality. MModality measures the diversity of multiple generated motions under the same text and keyframe condition. Given multiple generations obtained with different random seeds, it computes the average pairwise distance between their motion embeddings:

$$\text{MModality} = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{|\mathcal{P}_i|}\sum_{(a,b)\in\mathcal{P}_i}\|z_i^{(a)} - z_i^{(b)}\|_2, \tag{55}$$

where $\mathcal{P}_i$ denotes the sampled generation pairs for condition i. Higher MModality indicates stronger conditional diversity while maintaining the same interaction requirement.

E.3 Interaction Coordination Quality

Contact Distance Accuracy (CDA). CDA measures the accuracy of contact-related inter-person distances between generated and ground-truth motions. For each sequence i, we construct a set $\mathcal{C}_i$ of joint-pair tuples (t, j, k) on unobserved frames, where the ground-truth distance between joint j of the first character and joint k of the second character is below 20 cm. These pairs serve as geometric proxies for contact. CDA is computed as

$$\text{CDA@}\epsilon = \frac{1}{|\mathcal{I}_C|}\sum_{i\in\mathcal{I}_C}\frac{\sum_{(t,j,k)\in\mathcal{C}_i}\mathbb{I}\left[\left|\hat{d}_{i,t,j,k} - d^{\text{GT}}_{i,t,j,k}\right| < \epsilon\right]}{|\mathcal{C}_i|}, \tag{56}$$

where $\hat{d}_{i,t,j,k}$ and $d^{\text{GT}}_{i,t,j,k}$ denote the generated and ground-truth Euclidean distances for the same cross-person joint pair, respectively. $\mathcal{I}_C$ contains sequences with at least one eligible joint pair, and $\mathbb{I}[\cdot]$ is the indicator function. We report CDA@5cm with $\epsilon = 5$ cm. Higher CDA indicates more accurate reconstruction of contact-related inter-person distances. We additionally report CDA@10cm and CDA@15cm in the supplementary experiments to evaluate contact-distance accuracy under more relaxed error tolerances.

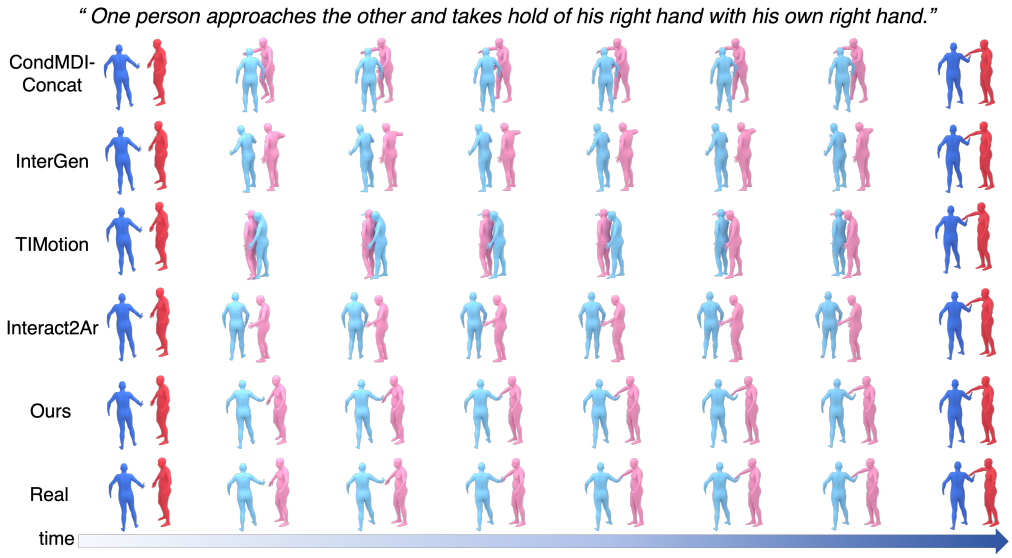


Figure 7: Qualitative comparison with prior state-of-the-art methods on a Inter-X sample. Darker color indicates keyframes. Our method produces smooth and interaction-coordinated in-between keyposes.

Table 8: Ablation study of pairwise motion context construction on InterHuman.

Feature Design	L2P↓	L2Q↓	NPSS↓	FID↓	Diversity→	R@1/2/3↑	MM-Dist↓	MModality↑
w/o $(h_1 + h_2)$	0.710	1.041	0.650	0.520	7.890	0.087/0.148/0.204	3.796	0.309
w/o $(h_2 - h_1)$	0.719	1.041	0.650	0.512	7.814	0.093/0.157/0.203	3.796	0.314
w/o $\lvert h_2 - h_1 \rvert$	0.709	1.037	0.644	0.504	7.797	0.092/0.156/0.203	3.797	0.318
w/o $(h_1 \odot h_2)$	0.708	1.037	0.645	0.508	7.800	**0.098**/0.155/0.205	3.795	0.316
Direct concat $[h_1, h_2, t]$	0.755	1.060	0.662	0.525	7.795	0.090/0.155/0.200	3.796	0.301
Ours Full	**0.700**	**1.024**	**0.634**	**0.497**	**7.784**	0.092/**0.159/0.206**	**3.785**	**0.326**

Relative Orientation Accuracy (ROA). ROA measures the consistency of the two characters' relative pelvis orientations between generated and ground-truth motions. We estimate each character's horizontal heading from its hip axis and compute the signed heading difference between the two characters. ROA is defined as

$$\mathrm{ROA@}\alpha = \frac{1}{|\mathcal{I}_R|} \sum_{i \in \mathcal{I}_R} \frac{\sum_{t \in \Omega_i^R} \mathbb{I}\left[\left|\mathrm{wrap}_{180}\left(\hat{\phi}_{i,t} - \phi_{i,t}^{\mathrm{GT}}\right)\right| < \alpha\right]}{|\Omega_i^R|}, \tag{57}$$

where $\hat{\phi}_{i,t}$ and $\phi_{i,t}^{\mathrm{GT}}$ denote the generated and ground-truth relative heading angles, respectively, and $\mathrm{wrap}_{180}(\cdot)$ maps angular differences to $[-180°, 180°)$. Ω_i^R contains unobserved frames with valid ground-truth headings, and $\mathcal{I}_R$ contains sequences with at least one such frame. Undefined predicted headings are counted as failures. We report ROA@15° with $\alpha = 15°$. Higher ROA indicates better preservation of relative facing directions between the two characters. We additionally report ROA@5° and ROA@30° in the supplementary experiments to evaluate relative orientation accuracy under stricter and more relaxed angular tolerances.

Table 9: Quantitative comparison on InterHuman in terms of interaction coordination quality.

Method	CDA ↑			ROA ↑		
	@5 cm	@10 cm	@15 cm	@5°	@15°	@30°
Real	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$	$1.000^{\pm 0.000}$
CondMDI-Concat(Cohan et al., 2024)	$0.379^{\pm 0.103}$	$0.765^{\pm 0.111}$	$0.851^{\pm 0.108}$	$0.207^{\pm 0.062}$	$0.546^{\pm 0.102}$	$0.755^{\pm 0.064}$
InterGen(Liang et al., 2024)	$0.391^{\pm 0.164}$	$0.550^{\pm 0.204}$	$0.650^{\pm 0.191}$	$0.307^{\pm 0.052}$	$0.640^{\pm 0.099}$	$0.793^{\pm 0.103}$
TIMotion(Wang et al., 2025)	$0.214^{\pm 0.091}$	$0.381^{\pm 0.142}$	$0.499^{\pm 0.160}$	$0.039^{\pm 0.024}$	$0.125^{\pm 0.063}$	$0.274^{\pm 0.102}$
Interact2Ar(Ruiz-Ponce et al., 2026)	$0.315^{\pm 0.147}$	$0.498^{\pm 0.194}$	$0.602^{\pm 0.189}$	$0.145^{\pm 0.058}$	$0.444^{\pm 0.121}$	$0.682^{\pm 0.114}$
Ours	$\mathbf{0.491}^{\pm 0.110}$	$\mathbf{0.816}^{\pm 0.082}$	$\mathbf{0.926}^{\pm 0.047}$	$\mathbf{0.317}^{\pm 0.077}$	$\mathbf{0.706}^{\pm 0.088}$	$\mathbf{0.930}^{\pm 0.042}$

F Additional Experimental Results

F.1 Qualitative Results on Inter-X

To further evaluate the generalization ability of our method, we provide additional qualitative comparisons on the Inter-X dataset. Figure 7 shows a representative example under the interaction prompt "One person approaches the other and takes hold of his right hand with his own right hand."

Compared with existing approaches, our method generates a more coherent interaction transition while preserving the individual motion patterns of both characters. In particular, our method maintains a natural approaching process and gradually reduces the interpersonal distance before hand contact, resulting in a more consistent spatial relationship between the two characters. In contrast, CondMDI-Concat independently generates two individual motions and therefore struggles to establish precise interaction coordination. Other two-person generation methods produce plausible interaction responses but may exhibit less accurate transition dynamics or weaker preservation of individual motion details.

This example demonstrates that our method effectively combines reliable individual motion initialization with interaction-aware refinement, allowing the generated motions to satisfy both interpersonal coordination and fine-grained motion quality requirements.

F.2 Interaction Coordination Quality.

Table 9 evaluates interaction coordination on InterHuman using CDA and ROA at multiple thresholds. Our method achieves the highest mean scores among all evaluated generation methods across all six metrics. For CDA, it outperforms the strongest baseline, CondMDI-Concat, at each distance threshold, with the largest absolute improvement at 15 cm (0.926 versus 0.851). For ROA, our method consistently exceeds InterGen, the strongest baseline across the evaluated angular thresholds, reaching 0.930 at 30° compared with 0.793. These correspond to gains of 7.5 and 13.7 percentage points in CDA@15 cm and ROA@30°, respectively. Although the gains at the strictest thresholds are smaller, the consistent improvements in mean scores across both metrics suggest better interaction coordination over a range of evaluation tolerances.

F.3 InterHuman Keyframe Robustness

We further compare our work with existing approaches under different numbers of provided keyframes. As shown in Table 10, our method consistently achieves competitive performance across different keyframe settings. These results demonstrate that our method generalizes well to different levels of keyframe availability and can effectively perform interaction refinement under diverse constraint conditions.

F.4 Inter-X Keyframe Robustness

To further evaluate the robustness of our method under different constraint densities, we conduct additional experiments on the Inter-X subset with different numbers of provided keyframes. Following the InterHuman setting, we randomly sample $K = 10$ and $K = 20$ keyframes from each motion

Table 10: Robustness evaluation under different keyframe conditions on InterHuman. K denotes the number of randomly sampled keyframes provided during motion in-betweening.

Condition	Method	L2P↓	L2Q↓	NPSS↓	FID↓
$K = 10$	CondMDI-Concat(Cohan et al., 2024)	0.663	0.986	1.415	0.910
	InterGen(Liang et al., 2024)	0.609	1.109	1.657	0.534
	TIMotion(Wang et al., 2025)	4.584	2.444	3.415	1.422
	Interact2Ar(Ruiz-Ponce et al., 2026)	0.946	1.089	1.771	1.104
	Ours	**0.602**	**0.851**	**0.482**	**0.339**
$K = 20$	CondMDI-Concat(Cohan et al., 2024)	0.591	0.879	1.242	0.726
	InterGen(Liang et al., 2024)	**0.519**	1.084	1.396	0.432
	TIMotion(Wang et al., 2025)	4.518	2.293	3.038	1.010
	Interact2Ar (Ruiz-Ponce et al., 2026)	0.773	0.971	1.580	0.849
	Ours	0.588	**0.744**	**0.420**	**0.254**

sequence and compare our method with all adapted baselines under the same text and keyframe conditions. The results are reported in Tables 6 and 7.

Our method consistently maintains strong performance under both sparse and denser keyframe settings. When more keyframes are provided, our method effectively utilizes the additional motion observations while preserving the quality of the initialized individual motions. This demonstrates that the proposed refinement strategy does not rely on a specific constraint density and can adapt to different levels of motion availability. In contrast, methods that directly generate interacting motions show larger variations when the amount of observed motion changes, indicating higher sensitivity to the constraint conditions.

Moreover, our method maintains stable text-motion consistency and motion variation under different keyframe settings. The results show that interaction refinement guided by global interaction information and locally controlled residual correction can improve coordination without over-constraining the generated motions. Overall, these experiments verify that our method generalizes well to diverse keyframe conditions on the more challenging Inter-X dataset.

F.5 Ablation Study of Pairwise Motion Context Features

To analyze the contribution of different components in the Pairwise Motion Context, we conduct an additional feature ablation study by removing individual feature terms from the pairwise descriptor. As shown in Table 8, each component provides complementary information for characterizing the current pairwise motion state.

Removing the summation term $(\mathbf{h}_1 + \mathbf{h}_2)$ leads to performance degradation, indicating that the overall pair motion state provides useful global information for interaction refinement. Removing the directional difference term $(\mathbf{h}_2 - \mathbf{h}_1)$ also reduces performance, demonstrating the importance of preserving role-dependent motion differences between the two characters. The absolute difference term $|\mathbf{h}_2 - \mathbf{h}_1|$ captures the magnitude of pairwise discrepancy, and removing it decreases the ability to identify coordination deficiencies. Meanwhile, removing the Hadamard product $\mathbf{h}_1 \odot \mathbf{h}_2$ affects the performance less significantly but still results in degradation, suggesting that feature-level compatibility provides additional interaction cues.

Compared with direct concatenation of the two individual motion features, the designed pairwise descriptor achieves better performance, showing that explicitly modeling symmetric states, directional differences, discrepancy magnitude, and feature compatibility is more effective than leaving relation discovery entirely to subsequent networks. The full feature combination achieves the best overall results, demonstrating that these components provide complementary information for constructing the Pairwise Motion Context.

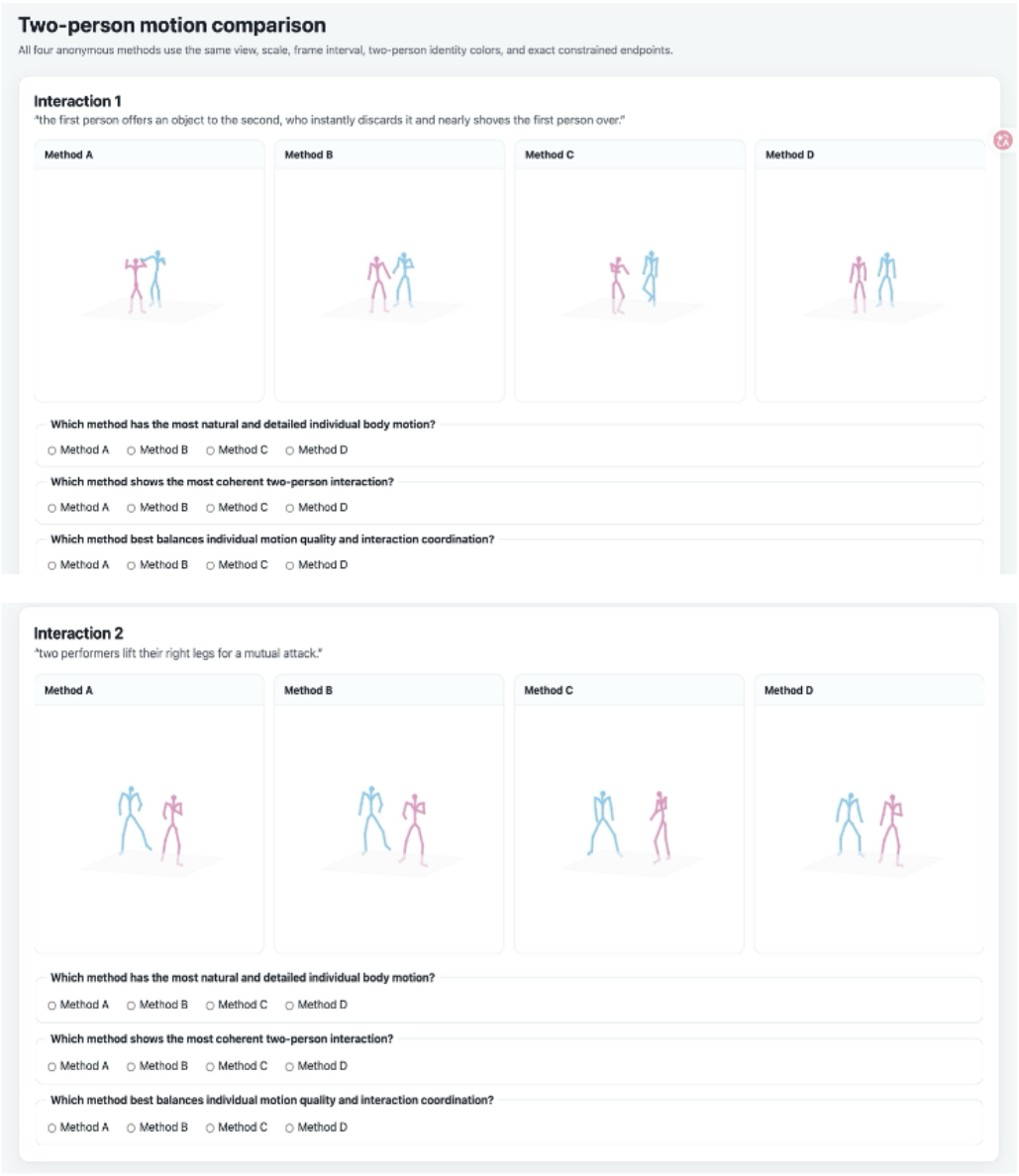


Figure 8: A screenshot of the user study.

F.6 User Study

To further evaluate the perceptual quality of generated two-person motions, we conduct a user study comparing our method with several strong baselines as showed in Figure 8. The evaluation is performed using an anonymous pairwise comparison setting, where participants are asked to select the motion sequence that provides better overall quality.

We randomly select five interaction samples from the InterHuman dataset. For each sample, participants compare the generated motion sequences from different methods under the same text condition and keyframe constraints. The evaluation focuses on two aspects: motion smoothness and interpersonal interaction harmony. To avoid potential bias, the method identities are hidden and presented as anonymous options during the evaluation. We collect responses from 40 participants. Our method achieves the highest user preference and receives nearly 50% preference, demonstrating that the generated motions are perceived as smoother and more naturally coordinated compared with exist-

ing approaches. These results provide additional human-level evidence that our method effectively balances individual motion quality and interaction coordination.