A. Appendix Section

In the following sections, we present detailed insights into the design and performance of our method for multi-person motion generation and coordination. The Design Details of the $U(\cdot)$ Function section elaborates on the key aspects of motion alignment, smooth transitions, and orientation adjustments, which are crucial for generating coherent and realistic motion sequence. More additional experiments results are displayed in the Additional Experiments section. In the Multi-Agent Motion Generation Results and Analysis section, we demonstrate how the dual-person interaction alignment module can be extended to multiple agents, illustrating the process and limitations of generating interactive motion sequences for more complex multi-agent scenarios. Finally, the Additional Visual Results section provides a series of visualizations that highlight the model's effectiveness in synthesizing and coordinating diverse motion sequences across different scenarios, offering further insight into its ability to handle complex interactions and maintain synchronization between multiple agents.

A.1. Design Details of the $U(\cdot)$ Function

The function $U(\cdot)$ manages motion alignment, smooth transitions, and orientation adjustments between segments, resulting in a coherent interleaved sequence.

Motion Alignment. The transformation from sk_2 to sk_1 can be expressed as:

$$j_k^{sk_1} = R_k \cdot j_{k_1}^{sk_2} + T_k j_k^{sk_2} + (1 - T_k)\, j_{k-1}^{sk_2} \tag{1}$$

where R_k is the rotation matrix for the k-th joint relative to the parent joint j_{k-1} in sk_2 and sk_1, and T_k is the translation matrix corresponding to the bone length difference between the k-th joint in sk_2 and sk_1 relative to the parent joint j_{k-1}.

Given the quaternions $q_k^{sk_1}$ and $q_k^{sk_2}$, which represent the rotations of the k-th joint relative to its parent joint j_{k-1} in skeletons sk_1 and sk_2, respectively, the relative rotation quaternion q_i between sk_1 and sk_2 is computed as:

$$q_k = q_k^{sk_1} \cdot (q_k^{sk_2})^{-1} \tag{2}$$

where $(q_i^{sk_2})^{-1}$ is the inverse of $q_i^{sk_2}$, equivalent to the conjugate of $q_i^{sk_2}$. The rotation matrix R_k can then be obtained from the quaternion q_i using the quaternion-to-rotation matrix conversion:

$$R_k = \text{QuatToRot}(q_i) \tag{3}$$

where $q_i = (w_i, v_i)$ with w_i being the scalar part and $v_i = (v_{i1}, v_{i2}, v_{i3})$ being the vector part of the quaternion.

To align the bone lengths between two skeletons at the k-th joint relative to the parent joint, we define T_k as the ratio of the bone length in sk_1 to the bone length in sk_2. This ratio accounts for the difference in bone lengths between the corresponding joints in the two skeletons.

The formula for T_k can be expressed as follows:

$$T_k = \frac{L_k^{sk_1}}{L_k^{sk_2}} \tag{4}$$

where $L_k^{sk_1}$ is the bone length between the k-th joint and its parent joint in sk_1, $L_k^{sk_2}$ is the bone length between the k-th joint and its parent joint in sk_2.

Smooth Transitions. Depending on the values of t_s and t_i (for example, when $t_s = 0$ and t_i is fixed), we align the processed data by the root node, which, although keeps the positions of the human body on the original trajectory, inevitably results in discontinuities between consecutive frames. To mitigate this issue, we employ Spherical Linear Interpolation (SLERP) to smooth the transitions. Specifically, for the frames before and after the t_i-th frame, we apply SLERP over five frames to interpolate between them. SLERP provides a smooth transition between two rotations by interpolating along the shortest path on a sphere, ensuring continuous motion. This interpolation technique allows us to create more fluid and natural motion sequences.

Additionally, when computing L_{rec}, we mask the features corresponding to these frames to avoid noise caused by abrupt transitions, and apply L_{smooth} to handle this smoothing process. This ensures that the model's training is not influenced by noisy data, while also enabling it to learn the ability to transition smoothly between different motion segments.

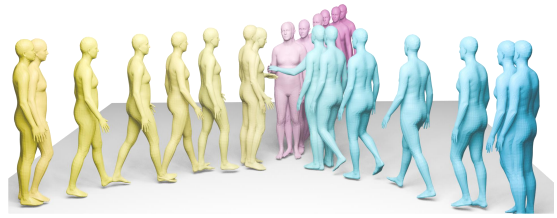
Orientation Adjustments. When inserting different data, although we ensure that the position of the root point remains relatively unchanged, errors in orientation may occur. Specifically, the orientation may either remain the same as the previous frame or be rotated 180 degrees around the z-axis. This issue arises because we use the 263-dimensional recursive representation of the HumanML3D dataset. During the recursion process, there can be singularities in the representation of the rotation angle, meaning that a rotation can represent orientation information that differs by 180 degrees. To address this, we add an extra cross-product check. For the last frame of the previous data and the first frame of the inserted data, we compute the cross product of the root node and the corresponding two nodes on the inner thigh. We then check whether the dot product of the resulting vectors is greater than 0. If the dot product is greater than 0, there is no issue. However, if it is less than 0, we apply a 180-degree rotation transformation around the z-axis to the inserted data to correct the orientation.

A.2. Additional Experiments

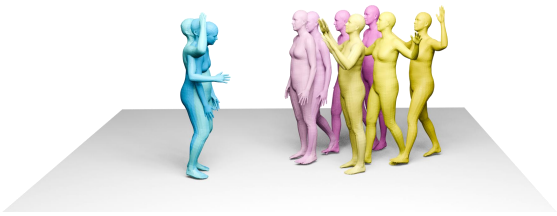
We conduct two sets of experiments to evaluate our model: the human study and the quantitative evaluation on HumanML3D dataset. In the human study, we evaluated mod-

Human Study	Method	Text-Motion ↑	Motion Naturalness ↑	Interactivity ↑
	FreeMotion	2.57	3.41	2.17
	Ours	3.86	4.02	3.64
Random Samples	Method	R Precision Top 1 ↑	R Precision Top 2 ↑	FID ↓
Single-person Dataset	FreeMotion	$0.179^{\pm.022}$	$0.294^{\pm.017}$	$1.785^{\pm.068}$
	Ours	$\mathbf{0.204}^{\pm\mathbf{.002}}$	$\mathbf{0.357}^{\pm\mathbf{.008}}$	$\mathbf{0.740}^{\pm\mathbf{.023}}$
	Method	MM Dist ↓	Diversity →	MModality ↑
	FreeMotion	$5.261^{\pm.009}$	$9.372^{\pm.090}$	$1.028^{\pm.059}$
	Ours	$\mathbf{3.928}^{\pm\mathbf{.010}}$	$9.127^{\pm.045}$	$\mathbf{1.476}^{\pm\mathbf{.021}}$

Table 1. Experimental results for human study and experiments in single-human dataset, respectively.



(a). A Example of Generating Multi-person Motion Sequence with Same texts



(b). A Example of Generating Multi-person Motion Sequence with Different texts

Figure 1. Multi-Person Motion Sequence Generation with Varied Text Inputs.

els on text-motion alignment, motion naturalness, and dual-human interactivity, with 42 participants rating the results (1–5 scale). Our model consistently receives higher ratings across all criteria, demonstrating better semantic alignment and more natural, interactive motion generation. In the quantitative evaluation, our approach achieves higher retrieval accuracy, better generation quality, and improved multimodal consistency, confirming its effectiveness in producing accurate, diverse, and semantically aligned motions.

A.3. Multi-Agent Motion Generation Results and Analysis

In our work, we first introduced the framework for generating single-agent motions and then extended it to handle dual-agent interactions through the concept of interleaved motion synthesis. This approach allows the model to generate motion sequences for two agents, where each agent's motion is interwoven in a way that respects both individual action and interaction between the agents. By leveraging the conditional information, we generate coherent sequences for each agent, ensuring that their motions are synchronized and contextually meaningful in the interaction. This dual-agent interaction framework serves as a foundation for extending the model to more complex scenarios.

Building upon this, our method is capable of generating motion sequences for single agents, dual-agent interactions according to Fig 2, and can even be extended to multi-agent scenarios. This scalability is made possible by the modular design of our system, where the core principles of motion synthesis and coordination are applied iteratively across multiple agents. The model can generate multiple independent motion sequences and coordinate interactions across agents, ensuring that even in more crowded or complex scenes, the generated motions remain natural and coherent.

We can generalize the dual-person interaction alignment module to multiple agents by leveraging the properties of the coordinator in the main text. Specifically, we generate multiple independent motion sequences (including interactive actions) through the INS module. Each motion sequence can be treated as a sequence generated under the constraints of other motion sequences. We iteratively feed these motions into the coordinator, using the example of three-person motions. Figure 1 illustrates an example of three-person motion generation. Although our current network can support the generation of multi-agent motions, the model's sensitivity to interaction constraints is still limited. Additionally, it requires relatively structured and routine actions (e.g., walking over and shaking hands) to generate high-quality results. For more complex actions (e.g., A greets B, then hugs C), the model has limited generalization ability with respect to spatial positions. In future work, we may consider modeling trajectory constraints for multiple agents.

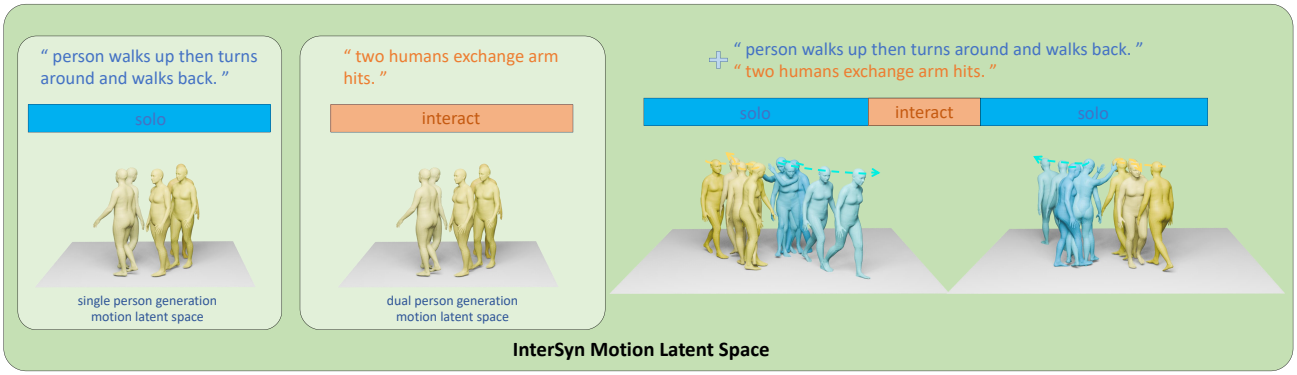


Figure 2. Additional Generation of Interleaved Motion Sequences.

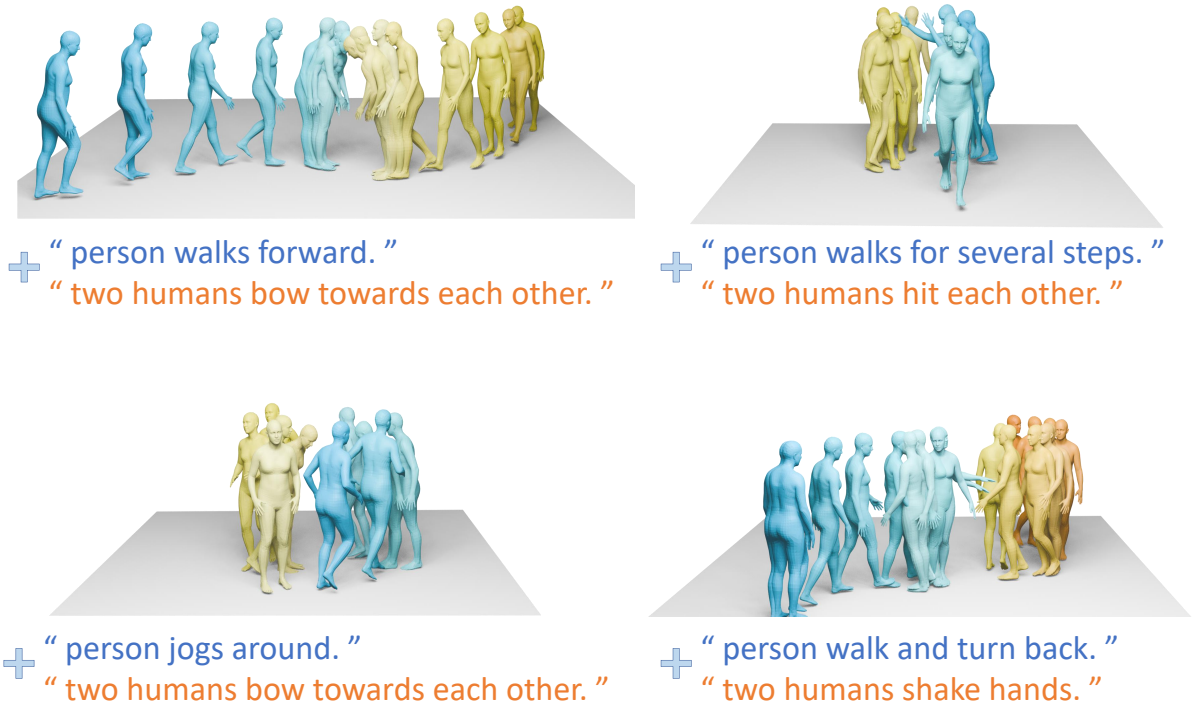


Figure 3. Additional Generation of Interleaved Motion Sequences.

A.4. Additional Visual Results

In this section, we present additional visual results to further demonstrate the effectiveness of our proposed method. The provided visualization highlights the motion synthesis and coordination processes across different scenarios. By visualizing both the individual and coordinated actions, we can observe how the model handles complex interactions between characters, ensuring smooth transitions and realistic behavior. These results offer further insight into the model's ability to synthesize diverse motion sequences and manage coordination tasks, such as maintaining proper alignment and synchronization across multiple agents in a dynamic environment.