FreeMotion: A Unified Framework for Number-free Text-to-Motion Synthesis

Ke Fan[1], Junshu Tang[1], Weijian Cao[2], Ran Yi[1] ✉, Moran Li[2], Jingyu Gong[1], Jiangning Zhang[2], Yabiao Wang[3,2], Chengjie Wang[1,2], and Lizhuang Ma[1,4,5] ✉

[1] Shanghai Jiao Tong University
{slipperyfrank, tangjs, ranyi, gongjingyu, lzma}@sjtu.edu.cn
[2] Tencent Youtu Lab
weijiancao@tencent.com, moranli.aca@gmail.com, {caseywang, jasoncjwang}@tencent.com
[3] Zhejiang University
[4] East China Normal University
[5] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
https://VankouF.github.io/FreeMotion

Abstract. Text-to-motion synthesis is a crucial task in computer vision. Existing methods are limited in their universality, as they are tailored for single-person or two-person scenarios and can not be applied to generate motions for more individuals. To achieve the number-free motion synthesis, this paper reconsiders motion generation and proposes to unify the single and multi-person motion by the conditional motion distribution. Furthermore, a generation module and an interaction module are designed for our *FreeMotion* framework to decouple the process of conditional motion generation and finally support the number-free motion synthesis. Besides, based on our framework, the current single-person motion spatial control method could be seamlessly integrated, achieving precise control of multi-person motion. Extensive experiments demonstrate the superior performance of our method and our capability to infer single and multi-human motions simultaneously.

Keywords: Text-to-motion synthesis · Diffusion models

1 Introduction

Human motion synthesis [7,8,19,28,35] aims at generating human motions driven by the input text, audio, scene, etc., and has extensive applications in areas like robotic control, movies, and animation production. Among these input signals, text, *i.e.*, natural language, can provide rich semantic details and is the most user-friendly and convenient signal for motion synthesis.

✉ Corresponding authors.

Generating single-person motion, as a fundamental task in text-to-motion (T2M), has received significant attention in the research community. Some methods [8, 22] aligned the text and motion in the same space by utilizing the VAE structure, while others [34, 38, 39] utilized the diffusion model conditioned on text to make the generated motion more diverse and realistic. Recently, some works have begun to explore two-person motion synthesis. ComMDM [24] fine-tuned a pre-trained single-human motion generation model on a dataset with only 20 two-person text-annotated motions by few-shot learning. InterGen [17] further introduced a large annotated two-person motion dataset, InterHuman, and employed a shared-weight diffusion network and cross-attention to model the interaction of two-person motion.

Although the aforementioned methods have demonstrated remarkable achievements in both single-person and two-person motion generation, their applicability remains limited in terms of universality. Specifically, these limitations include: (1) their models are tailored either for fitting the marginal motion distribution of single-person or the joint motion distribution of two-person scenarios and cannot support motion inference for both cases simultaneously. (2) Moreover, due to the unavailability of text-annotated multi-person motion datasets, as well as the non-scalability of network design, none of these methods can be applied to generate motion for more than two individuals during the inference phase. (3) While some studies have accomplished spatial control for single-person motion generation, incorporating the spatial signal into the current multi-person motion generation model remains a non-trivial challenge.

To address these problems, we rethink the process of multi-person motion synthesis and propose FreeMotion, a novel unified framework that enables motion generation for any number of individuals and achieves fine-grained spatial control of multiple human motions.

Our key insight is to build a bridge between the marginal and joint distribution with the conditional distribution. Without loss of generality, for the n-person motion generation task, we need to model the joint distribution probability $p(\mathbf{x}^1, ..., \mathbf{x}^n)$. According to the formula of conditional probability, we can decompose the joint distribution as $p(\mathbf{x}^1, ..., \mathbf{x}^n) = p(\mathbf{x}^1) \prod_{i=1}^{n-1} p(\mathbf{x}^{i+1} | \mathbf{x}^i, ..., \mathbf{x}^1)$. Therefore, if we can model the conditional motion distribution, we can generate motions for n individuals in a recursive process, as shown in Fig. 1. Through the conditional probability formula, we reduce the task of multi-person motion synthesis to the single-person motion synthesis under the guidance of other people's motions, *i.e.*, conditional single-person motion synthesis.

To this end, we design a conditional motion diffusion network, which can generate the target motion conditioned on the motions of any number of other individuals. We decouple the modeling of conditional motion distribution into single-person motion generation and multi-person motion interaction, which correspond to the two modules of our network, namely the generation module and the interaction module. The generation module aims to generate diverse and vivid single-person motion according to the text prompt. While the interaction module aims to inject condition motions into the human motion generation by

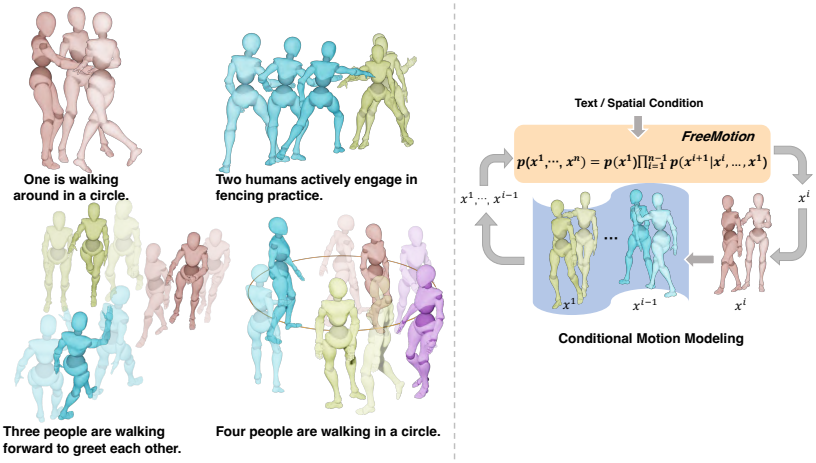

Fig. 1: The **left** shows our model can generate controllable motions for any number (1–4 from the figure) of individuals. Different colors represent the different person's motion. The **right** is an illustration of our new paradigm of motion generation, recursive generation, where every single motion is predicted under the condition of the motions generated before. Best viewed in color.

extracting the interactive information between condition motions and the current motion to be generated. To accommodate the variability in the number of motion conditions in the interaction module, we leverage the length-independent characteristics of self-attention and deliberately design the interactive block to be entirely based on global self-attention. Besides, to describe the motion of each person, we design prompts and employ a large language model [21] to transform multi-person motion descriptions into corresponding descriptions of each person.

Furthermore, we enhance the fine-grained spatial control of multi-person motion generation. Previous works [32, 33] add accurate control on the joint of a single-person motion. Nevertheless, these control signals are not easy to be applied to multi-person motion generation. In our work, we incorporate flexible spatial signals into the interaction module to control the global location of human motions while maintaining the realism of the motion results. As demonstrated in Fig. 1, although our model is only trained on two-person motions, it has the ability to predict the movements of more than two individuals in the inference phase.

In summary, our contributions are as follows: (1) We rethink the process of motion synthesis and propose a new paradigm to unify the synthesis of motions for both single and multiple people. (2) We propose a decoupled generation and interaction module for conditional motion generation, which is the first attempt to achieve high-fidelity number-free motion generation under text conditions. (3) We further achieve precise control of multi-person motion based on flexible spatial control signals utilizing explicit and implicit guidance. Extensive experi-

ments demonstrate that our method outperforms prior works, and achieves vivid multiple motions results.

2 Related Work

2.1 Single-Person Motion Synthesis

At present, the mainstream text2motion methods are mainly divided into two categories: align-based model and the condition-based model.

The align-based model mainly aligns text and motion into a shared latent space. In the inference stage, features are extracted based on the given text, and the Features are regarded as corresponding motion features for action generation. TEMOS [22] and Guo *et al.* [8] leverage the VAE architecture to learn a joint latent space of motion and text constrained on a normal distribution. However, natural language and human motions are quite different with misaligned structure and distribution, which makes the alignment process quite difficult.

Condition-based models are usually based on the diffusion model architecture. It uses pre-trained text encoders, such as CLIP, to extract text features, and inject text features as conditions into the diffusion reconstruction network to guide the network to generate corresponding motions. MDM [31] and Motion Diffuse [37] are the first works to introduce the diffusion model into the motion generation field. MDM additionally introduces a geometric loss to improve the model performance. MLD [1] further leverages the latent diffusion model to significantly drop the training and inference cost. However, Xie *et al.* [34] points out that the latent dimension affects the performance of the model. It cascades two diffusion models with different latent dimensions, promoting the details and the modal consistency. ReMoDiffuse [38] proposes an enhancement mechanism based on data set retrieval to refine the denoising process of Diffusion. MotionLCM [4] proposes to leveraging the consistency model to accelerate the sampling speed. FineMoGen [39] and Motion-X [18] further propose a new large-scale data set to introduce more detailed descriptions, such as hand joints, expression, etc. GMD [13] and OminiControl [33] explore precise trajectory control by the impainting technique or both implicit and explicit spatial guidance. However, these methods are all networks designed for single-person models, and it is difficult to achieve motion synthesis for two or even multiple individuals.

2.2 Multi-Person Motion Synthesis

Multi-person motion synthesis is more difficult than single-human motion generation, which involves an interactive process between multiple individuals. Early works tend to use motion graphs [25] and momentum-based inverse kinematics [14]. Guo *et al.* [9] proposes the Extreme Pose Interaction dataset as well as a two-stream network with cross-interaction attention for interaction modeling. InterFormer [3] uses an attention-based interaction transformer to generate sparse-level reactive motions on the K3HI [12] and the DuetDance [15]

datasets. SocialDiffusion [30] proposes the first diffusion-based stochastic multiperson motion anticipation model. BiGraphDiff [2] further introduces bipartite graph diffusion for geometric constraints between skeleton nodes. Tanaka *et al.* [27] introduces a PIT module the diffusion network, which enables the model to automatically distinguish actors and receivers, thereby better learning the interaction process. To broaden the applications, ReMos [6] and Le *et al.* [16] further propose new datasets that contain hand movements or music. However, these methods depend on either historical motion, action label, or music to give the motion prediction, and can not support the task of T2M.

ComMDM [24] annotates 3DPW manually to obtain a small scale of samples. Furthermore, it fine-tunes a pre-trained single-person T2M model in a few-shot manner and attempted multi-person motion generation for the first time. However, since it only contains 27 two-person motion sequences, The model's ability to generate two-person interactions is greatly restricted. Therefore, InterGen [17] first contributes a large-scale text-annotated two-person motion dataset called InterHuman, and based on this, it proposed a diffusion model with shared weights and multiple regularization losses, outperforming the ComMDM. InterControl [32] attempts to complete the interaction process by controlling the joints of two individuals to a certain position. It uses LLM to generate planning for two individuals during their movements by designing prompts. However, since it was not trained on two individuals, the model can not model the interaction process explicitly. In short, these T2M models almost all directly predict the actions of a single person or a pair of individuals, and cannot support the inference of single and double persons at the same time, let alone support the inference of motion for multiple individuals at the same time.

3 Preliminaries

3.1 Diffusion Model for Motion Synthesis

The diffusion Model [5,10,11,20,26,29,36,40] is a probabilistic model that gradually denoises a Gaussian noise to generate a target output. The key point is to generate a target output by gradually denoising Gaussian noise. It is formulated as a diffusion process and a reverse process and is utilized to approximate the posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$, where T is the total time steps, and $\mathbf{x}_1, ..., \mathbf{x}_T$ are the real data x_0 with t steps of noise added. The diffusion process follows a Markov chain to gradually add Gaussian noise to the data until its distribution is close to the latent distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, according to variance schedules given by β_t:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \end{aligned} \tag{1}$$

During training, the mean-squared error loss is used to optimize the parameters. During inference, we can generate the target output by sampling the initial

noise ϵ and denoising it by predicting the added noise, mean, or x_0 recursively. To fit the arbitrary length of time, the transformer-based denoiser block is commonly preferred in motion generation.

3.2 Motion Interaction Representation

To maintain generality, we model multi-person motion generation as a motion sequence $\mathbf{X} = (\mathbf{x}^p)$, $p \in \{1, 2, ..., N\}$, where N represents the total number of individuals. For simplicity, we use $\mathbf{x}$ to represent the motion of a specific individual.

In the context of generating motion for a single person, the canonical representation proposed by HumanML3D [8] has proven to be well-suited for neural network processing. However, its normalization operation loses the relative position relationship in multi-person scenarios, making it unsuitable for such use cases. To address this limitation, InterGen [17] introduced a non-canonical representation that maintains the relative positions between different individuals. The formulation of this representation is as follows:

$$x^p(i) = [\mathbf{j}_{pg}, \mathbf{j}_{gv}, \mathbf{j}_r, \mathbf{c}_f], \tag{2}$$

where the i-th pose of x^p is defined as a collection of global joint positions $\mathbf{j}_{gp} \in \mathbb{R}^{3J}$, velocities $\mathbf{j}_{gv} \in \mathbb{R}^{3J}$ in the world frame, local rotations $\mathbf{j}_r \in \mathbb{R}^{6J}$ in the root frame, and binary foot-ground contact features $\mathbf{c}_f \in \mathbb{R}^4$, where J denotes the joint number. Since this representation preserves the global coordinates of joints, it well represents the spatial relationship of interactions. In this paper, we follow the non-canonical representation proposed by InterGen.

4 Method

4.1 Overview

In this paper, we propose a unified framework for number-free human motion generation. By reconsidering the generation process via decomposing the joint motion into a conditional distribution through the conditional probability formula, our framework can generate arbitrary number of motions in a recursive manner, as shown in Fig. 1. Specifically, we model the multi-person joint motion distribution as $p(\mathbf{x}^1, ..., \mathbf{x}^n) = p(\mathbf{x}^1) \prod_{i=1}^{n-1} p(\mathbf{x}^{i+1} | \mathbf{x}^i, ..., \mathbf{x}^1)$. Given a text $\mathbf{d}$, we first predict the motion of the first individual. Subsequently, the motion of the second individual is inferred, conditioned upon the motion of the first individual, and the process iterates until the motion of the $\mathbf{n}$-th individual is generated. Therefore, this method essentially transforms a multi-person motion generation process into a sequence of conditional single-person motion generation. Based on this, we can further achieve multi-human spatial motion control by separately controlling the spatial motion of each individual.

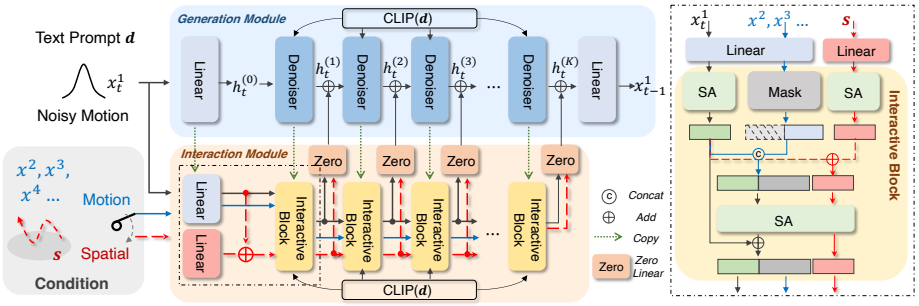

Fig. 2: Overall architecture of FreeMotion, which contains a generation module and an interaction module. Given a text $\mathbf{d}$, our framework can infer a motion x^1 by the generation module independently, or under the condition of multiple motions $x^2, x^3 ...$ or some spatial guidance $\mathbf{s}$. Red line represents the implicit guidance of the spatial control signal.

4.2 Number-free Motion Generation

As introduced before, we consider the multiple motions generation as a conditional generation process. From the perspective of conditional motion distribution, we decouple the modeling process into a motion generation process and a motion interaction process. To start with, we first design a generation module which has the capability of single motion generation. Then we design an interaction module to inject condition signals (motions of $N-1$ individuals) into the human motion generation process by modeling the interaction between condition signals and the current motion to be generated. The detailed architecture of the generation and the interaction module are shown in Fig. 2.

Generation Module. The generation module is designed to synthesize diverse single motion according to the text prompt. Similar to previous text-to-motion methods [31, 37], the module is a Transformer-based diffusion network, containing several denoiser blocks. During training, at timestep t, we add noise to a motion $\mathbf{x} \in \mathbf{R}^{F\times 262}$, where F represents the frame numbers, and get the noised motion $\mathbf{x}_t$, and then use the generation module to denoise $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$. During the inference, the generation module is able to generate a clean motion x_0 starting from the pure Gaussian noise $\mathbf{x}_T$.

Interaction Module. Inspired by ControlNet [36], we design a neural network named Interaction Module to inject the condition signals into the human motion generation. Specifically, we denote the $N-1$ motions in the condition signal as $\mathbf{x}^2, \mathbf{x}^3, .., \mathbf{x}^N$, and the noised motion obtained by adding noise to $\mathbf{x}^1$ as $\mathbf{x_t}^1$. Firstly, we feed $\mathbf{x}_t^1$ and $N-1$ motion conditions $\mathbf{x}^i, i \in \{2, ..., N\}$ into a shared linear layer to encode each motion into a hidden state, denoted as $\mathbf{h}_t^{1,0} \in \mathbf{R}^{F\times H}$, where H represent the latent dimension, and $\mathbf{h}^{i,0}, i \in \{2, ...N\}$. Secondly, to model the interaction information between variable number of motion conditions and the motion to be generated, we design a novel **Interactive Block**, which is sequentially stacked for K times. The interactive block at the

k-th stage ($k \in \{1, ..., K\}$) takes the $\mathbf{h}_t^{1,k-1}$ and $\mathbf{h}^{i,k-1}$ as inputs and outputs the $\mathbf{h}_t^{1,k}$ and $\mathbf{h}^{i,k}$. Each interactive block contains two sequential self-attention modules and a mask module, and the whole calculation process in the k-th interactive block is formulated as:

$$\mathbf{h}_t^{1,k}, \mathbf{h}^{2,k}, ..., \mathbf{h}^{N,k} = SA(SA(\mathbf{h}_t^{1,k-1}), Mask(\mathbf{h}^{2,k-1}, ..., \mathbf{h}^{N,k-1})), \tag{3}$$

where $k \in \{1, ..., K\}$ and SA represents the self-attention operation.

Specifically, in the k-th interactive block, we firstly randomly mask some hidden states $\mathbf{h}^{i,k-1}$, to enable our model to infer the motion under any number of motion conditions. Then, the noised hidden state $\mathbf{h}_t^{1,k-1}$ is input into the first SA module. We concatenate the masked $N-1$ hidden states, and the output of the first SA module along the time dimension, and feed the concatenated result to the second SA module, so that the model learns the interactive information between the noised motion and motion conditions. And the outputs of the second SA module, *i.e.,* $\mathbf{h}_t^{1,k}$ and $\mathbf{h}^{i,k}$s, will be further input to the next interactive block. By leveraging the global self-attention, $\mathbf{h}_t^{1,k}$ successfully fuses the information from the motion conditions.

The output from the interactive block will then be added to the output of each denoiser block in the generation module after going through a linear layer. The linear layers are initialized with zero, thereby ensuring that the interactive module does not impact the generation module at the onset of training. Compared with cross-attention, self-attention is more flexible for modeling the interaction between motions regardless of the number of motions in the condition.

4.3 Training Process

We decompose our training process into two stages. In the first stage, we train the generation module for single-person motion generation. To enable the generation module to synthesize the single-person motion, it is necessary to provide a single-person motion description as its input during training. Therefore, for a given interactive description of multi-person motion, we utilize the ability of the Large Language Model (LLM) [21] to generate N motion descriptions, one for each individual. We use the separated single-person description for the training of the generation module. The text feature is extracted by a pre-trained CLIP [23] model and is injected into the adaptive layer normalization in all attention layers.

In the second stage, we train the interaction module for conditional motion modeling and multi-human motion generation. Inspired by the success of ControlNet, we first freeze the parameters of the generation module and utilize the parameters to initialize the interaction module. Then we use multi-human motions data to train the interaction module. Denote the motions of N individuals as $[\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N]$. We initially randomize the order of these N motions and the corresponding descriptions, and then select the first one as the motion to be noised and reconstructed, leaving the other $N-1$ motions as the clean motion condition. 1) We add t steps of noise to $\mathbf{x}^1$ to obtain $\mathbf{x}_t^1$ and input it to the generation module. 2) For the interaction module, $\mathbf{x}_t^1$ and the remaining $N-1$

motion conditions $\mathbf{x}^p$, $p \in \{2, ..., N\}$ are input into a shared linear layer and K interactive blocks to extract the interactive information. 3) The output of each interactive block is added to the output from the corresponding denoiser block in the generation module after going through a linear layer. 4) Finally, the denoised motion x_{t-1}^1 can be obtained from the generation module.

4.4 Spatial Control

Our proposed framework based on conditional motion generation enables effortless spatial control over multi-human motions. It allows for the integration of the current single-person motion control method without the need for carefully designing the spatial representation of multiple individuals. Inspired by OmniControl [33], we concurrently utilize explicit and implicit guidance to realize our spatial control.

Explicit guidance. Given a desired spatial location $\mathbf{s} \in R^{F \times 3J}$, where F represents the valid motion length and J denotes the joint number. We utilize the L_2 distance $\mathbf{d} = \|\mathbf{s}_{nj} - \mathbf{x}_{nj}\|_2$ to measure the bias between target position s and the predicted motion $\mathbf{x}$ for joint j at frame n. Then we leverage the mechanism of classifier guidance to perturb the predicted motion at each denoising step t to correct the bias, which is formulated as $\mathbf{x}_t = \mathbf{x}_t - \eta \nabla_{\mathbf{x}_t} \mathbf{d}$, where η controls the step of the guidance.

Implicit guidance. In addition to explicit guidance, as the red lines shown in Fig. 2, we further employ implicit spatial guidance. The whole training process is almost the same as the number-free motion generation. A minor difference is that we first input the spatial signal $\mathbf{s}$ into an independent linear layer, and add the output together with the noised motion hidden state. Some frames and joints are randomly selected and the others are subsequently masked out for the whole training process.

After injecting the spatial control signal under explicit and implicit guidance, our proposed unified framework has the capability to control the spatial locations of multi-human motions independently, which further enhances the controllability of our method.

4.5 Loss Function

In addition to reconstruction loss $\mathcal{L}_{rec}$, we also incorporate some regularization losses, which include the contact loss $\mathcal{L}_{foot}$ and joint velocity loss $\mathcal{L}_{vel}$ mentioned in MDM [31] and the bone length loss $\mathcal{L}_{bl}$ as well as the masked joint distance map (DM) loss $\mathcal{L}_{dm}$ mentioned in InterGen [17]. For the first stage of single-motion generation, the total loss is formulated as:

$$\mathcal{L}_1 = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{foot} + \lambda_2 \mathcal{L}_{vel} + \lambda_3 \mathcal{L}_{bl}, \tag{4}$$

where the DM loss is not included. For the second stage of conditional single-motion generation, we further introduce DM loss to model the interactions, where

Table 1: Quantitative comparisons on the InterHuman test set. We run all the evaluations 20 times except MModality runs 5 times. ± indicates the 95% confidence interval. Bold indicates the best result.

Methods	R Precision↑			FID ↓	MM Dist↓	Diversity→	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.748^{\pm.064}$	-
TEMOS [22]	$0.224^{\pm.010}$	$0.316^{\pm.013}$	$0.450^{\pm.018}$	$17.375^{\pm.043}$	$5.342^{\pm.015}$	$6.939^{\pm.071}$	$0.535^{\pm.014}$
T2M [8]	$0.238^{\pm.012}$	$0.325^{\pm.010}$	$0.464^{\pm.014}$	$13.769^{\pm.072}$	$4.731^{\pm.013}$	$7.046^{\pm.022}$	$1.387^{\pm.076}$
MDM [31]	$0.153^{\pm.012}$	$0.260^{\pm.009}$	$0.339^{\pm.012}$	$9.167^{\pm.056}$	$6.125^{\pm.018}$	$7.602^{\pm.045}$	$\mathbf{2.355}^{\pm.080}$
ComMDM* [24]	$0.067^{\pm.013}$	$0.125^{\pm.018}$	$0.184^{\pm.015}$	$38.643^{\pm.098}$	$13.211^{\pm.013}$	$3.520^{\pm.058}$	$0.217^{\pm.018}$
ComMDM [24]	$0.223^{\pm.009}$	$0.334^{\pm.008}$	$0.466^{\pm.010}$	$7.069^{\pm.054}$	$5.212^{\pm.021}$	$7.244^{\pm.038}$	$1.822^{\pm.052}$
InterGen* [17]	$0.264^{\pm.006}$	$0.392^{\pm.005}$	$0.472^{\pm.005}$	$13.404^{\pm.200}$	$3.882^{\pm.001}$	$7.77^{\pm.030}$	$1.451^{\pm.034}$
FreeMotion	$\mathbf{0.326}^{\pm.003}$	$\mathbf{0.462}^{\pm.006}$	$\mathbf{0.544}^{\pm.006}$	$\mathbf{6.740}^{\pm.130}$	$\mathbf{3.848}^{\pm.002}$	$\mathbf{7.828}^{\pm.130}$	$1.226^{\pm.046}$

the total loss is formulated as:

$$\mathcal{L}_2 = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{foot} + \lambda_2 \mathcal{L}_{vel} + \lambda_3 \mathcal{L}_{bl} + \lambda_4 \mathcal{L}_{dm}, \tag{5}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are used to balance the effect of corresponding loss.

5 Experiments

5.1 Datasets and Metrics

Datasets. We evaluate our proposed framework on the **InterHuman** dataset, which is the first text-annotated two-person motion dataset, consisting of 6,022 motions derived from various categories of human actions, labeled with 16,756 unique descriptions composed of 5,656 distinct words.
Metrics. We utilize the same evaluation metrics as InterGen, which are FID, R Precision, Diversity, Multimodality (MModality) and Multi-modal distance (MM Dist). We introduce the details in the supplementary.

5.2 Quantitative Results

Baseline. We mainly compare our method against the current state-of-the-art approach *InterGen* that models the interaction relationship of two people by a cross-attention mechanism. However, it by default can only support the two-person motion inference and cannot generate single-person motion. Thus, we make minor modifications to its model structure and retrain it to make it adapt to the inference of both single and double human motion, ensuring a fair comparison. The corresponding output of cross-attention is set to zero with a 10% probability, thereby gaining some capacity for single-person generation. Our re-annotated single-person texts are utilized during the training phase. Since the

Table 2: Quantitative comparisons with InterGen* on our re-annotated text for single motion generation. The manner of evaluation is the same as Tab. 1. Bold indicates the best result.

Methods	R Precision↑			FID ↓	MM Dist↓	Diversity→	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.452^{\pm.008}$	$0.610^{\pm.009}$	$0.701^{\pm.008}$	$0.273^{\pm.007}$	$3.755^{\pm.008}$	$7.748^{\pm.064}$	-
InterGen* [17]	$0.206^{\pm.004}$	$0.313^{\pm.004}$	$0.389^{\pm.005}$	$23.415^{\pm.222}$	$3.925^{\pm.001}$	$7.514^{\pm.029}$	$\mathbf{1.526}^{\pm.026}$
FreeMotion	$\mathbf{0.264}^{\pm.005}$	$\mathbf{0.394}^{\pm.006}$	$\mathbf{0.473}^{\pm.006}$	$\mathbf{12.975}^{\pm.171}$	$\mathbf{3.885}^{\pm.035}$	$\mathbf{7.702}^{\pm.027}$	$1.300^{\pm.063}$

performance of other methods is inferior to the InterGen on two-person motion synthesis, we directly take the results reported by InterGen for the experiments of two-person motion synthesis.

Comparison Results. For evaluation on two-person motion synthesis, we follow the settings given by InterGen. As for single-person motion synthesis, we take the corresponding re-annotated two single-person descriptions given by LLM to synthesize two separate single-person motions and concatenate them together for performance evaluation. It can be seen from Tab. 1 and Tab. 2 that although the InterGen* can generate both single and two-person motions, the performance is inferior to ours in almost all terms of metrics. Besides, the remaining models in Tab. 1 are only trained on the interactive descriptions and have no ability for single-motion inference, their performance on two-person motion synthesis is still inferior to our method. This fully demonstrates the powerful generality of our proposed framework.

5.3 Ablation Studies

We investigate the influence of several designs on two-person motion generation. Initially, we remove the interaction module, solely employing the generation module (GM). During the second training stage, we finetune the parameters of GM. To endow the network with both single-person and two-person motion capabilities, we drop the motion condition with a 10% probability in the second SA module within each denoiser block. In addition, we exclusively use our re-annotated single-person motion texts, without the interactive description (InterDes) provided in the InterHuman dataset. Subsequently, we incorporate high-quality two-person interactive motion descriptions (2nd row) during the second training stage. Furthermore, we introduce the interaction module as well as InterDes to get our FreeMotion* (3rd row) and FreeMotion (4th row).

Comparing FreeMotion* with GM*, we found it almost gets the same results in all metrics. However, after introducing the high-quality interactive description (InterDes), our FineMotion achieved a significant improvement while the performance of GM dropped a great deal. We analyzed this and found that it is because when using only GM, the GM network parameters need to be updated

Table 3: Ablation study of our proposed framework. All results are reported on the InterHuman test set under the setting of two-person motion generation. **InterDes** and the **GM** represent the interactive description and generation module. We use * to separate whether to leverage InterDes or not.

Methods	InterDes	R-Precision1↑	FID ↓	MM Dist↓	Diversity→
GM*		$0.300^{\pm.005}$	$8.842^{\pm.130}$	$3.863^{\pm.001}$	$7.761^{\pm.036}$
GM	✓	$0.259^{\pm.005}$	$10.749^{\pm.145}$	$3.883^{\pm.001}$	$7.645^{\pm.031}$
FreeMotion*		$0.300^{\pm.004}$	$8.792^{\pm.135}$	$3.865^{\pm.001}$	$7.750^{\pm.022}$
FreeMotion	✓	$\mathbf{0.326}^{\pm.003}$	$\mathbf{6.740}^{\pm.130}$	$\mathbf{3.848}^{\pm.002}$	$\mathbf{7.828}^{\pm.130}$

in the second training stage. When using only single-person text, the text used in the second stage of training was consistent with the first stage and did not produce a large gap, so that the update process was able to ensure that GM* learned as well as FineMotion*. However, when InterDes was introduced, the differences between single and double text prevented the GM from effectively adapting its parameters after the first stage of training. However, as can be seen from the FineMotion results, the introduction of high-quality two-person texts is of great significance in improving the model performance. Overall, the decoupled generation module and interaction module we designed can utilize all the information for training more efficiently.

5.4 Qualitative Results

Single Person Motion Synthesis. To illustrate the effectiveness, we provide a qualitative comparison between the InterGen* and our FreeMotion on single-motion synthesis. As shown in Fig. 3, the synthesized single-person motion given by the proposed method are more consistent with the description. For example, in the second column, our model can produce an obvious squatting pose, which is more consistent with the textual command of grabbing others' waists. This fully demonstrates that the decomposition of the generation and interaction process we proposed can more adequately fit the mapping from text to single motion than randomly mask the interaction process in the training phase.

Two Person Motion Synthesis. We further exhibit the comparison of two-person motion synthesis in the 3rd and 4th columns in Fig. 3. Our approach performs better in two-person interaction coordination and comprehensibility in complex text. It can be seen that InterGen* does not achieve simultaneity when generating the rock-paper-scissors motion at the third column, and even switches the mainly used hands. In the fourth column, InterGen* only focuses on the kicking motion, which results in both people making kicks. In contrast, our method can perform better. We believe that this is also due to the decoupling of the generation and interaction processes. The generation module makes the

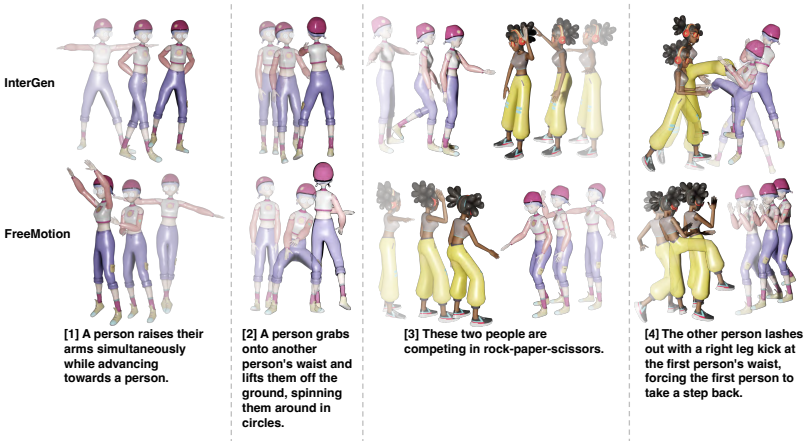

Fig. 3: Comparison with Intergen* on single and two-person motion generation. For single-person motion, we generate it with our re-annotated single description. For two-person motion, we further leverage the original interactive descriptions. For better visualization, some pose frames are shifted to prevent complete overlap.

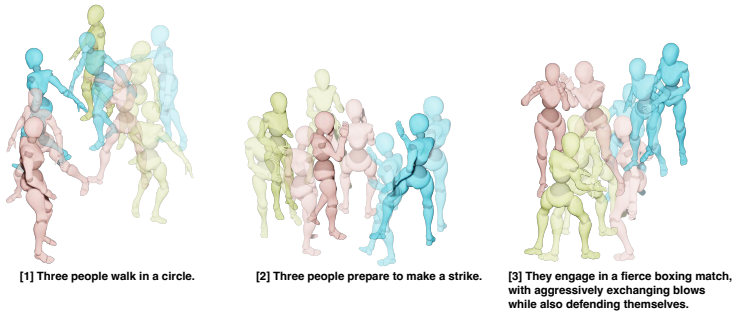

Fig. 4: Qualitative results for generating three-person motions. We manually design some text prompts and feed them to our network for motion generation. For better visualization, some pose frames are slightly shifted to prevent completed overlap.

comprehension of complex text better, and the interaction module enables better two-person synchronization.

Three Person Motion Synthesis. We manually designed the text with three different levels of interaction (low, middle, and high), corresponding to the results from left to right in Fig. 4. Our interactive block is designed with global self-attention, which allows our network to have the ability to support more people's motions as conditions, thus generating three people's motions. More importantly, although our approach is trained using only two people's motions due to data scarcity, the way we decouple the generation and interactive processes enables our network to have the ability to directly reason about three people's motions. The generation module enables our network to generate semantically consistent

Fig. 5: Results of multi-person spatial control. We manually design some text prompts as well as the trajectories and leverage the integrated spatial control module to generate the results.

one-person motions under weak interaction text. The interaction module enables the extraction of motion interaction features under strong interaction text. The above results further illustrate that the interactive block we designed can extract interaction information effectively.

Controllable Motion Synthesis. Thanks to our proposed paradigm of conditional motion modeling, we can seamlessly integrate existing single motion control models into spatial control of multi-person motions. As shown in Fig 5, even though the spatial control module is trained for single motions, mounting the interaction module when generating multi-person motions has no obvious damage to the spatial control and performs well from two to four-person.

6 Conclusion and Discussion

In this paper, we present FreeMotion, a novel motion generation framework for number-free motion synthesis. We rethink the way of multi-person motion generations and propose to recursively generate multi-person motions based on conditional motion modeling. We further propose the decoupled generation module and interaction module, and conduct a large number of quantitative as well as qualitative experiments to prove that our framework can support motion synthesis for any number of motions.

Discussion. The primary aim of our number-free claim is to pave a step for multi-human motion synthesis with limited data. However, we find our method still has some limitations. First, we directly utilize the capability of the large language model to separate single-motion text from the interactive description. Inevitably, there might be some instances where text and movements do not match well. Secondly, although our model can generate multi-person motions, due to the limited understanding of interactions from training only on two-person motions, there might be some interpenetration among different individuals when the number gets large or the text prompt is too complex. Therefore, we do believe high-quality multi-human motion data is still necessary. We hope our work can be the baseline for multiple-human interaction and inspire researchers to propose suitable multiple-human interaction datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 72192821, 62302297), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), the Fundamental Research Funds for the Central Universities (project number: YG2023QNA35). Thanks for Yating Wang providing some technical suggestions.

References

1. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
2. Chopin, B., Tang, H., Daoudi, M.: Bipartite graph diffusion model for human interaction generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5333–5342 (2024)
3. Chopin, B., Tang, H., Otberdout, N., Daoudi, M., Sebe, N.: Interaction transformer for human reaction generation. IEEE Transactions on Multimedia (2023)
4. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. arXiv preprint arXiv:2404.19759 (2024)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Remos: Reactive 3d motion synthesis for two-person interactions. arXiv preprint arXiv:2311.17057 (2023)
7. Gong, J., Wang, M., Liu, W., Qian, C., Zhang, Z., Xie, Y., Ma, L.: Demos: Dynamic environment motion synthesis in 3d scenes via local spherical-bev perception. arXiv preprint arXiv:2403.01740 (2024)
8. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
9. Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13053–13064 (2022)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Hu, T., Zhu, X., Guo, W., Su, K., et al.: Efficient interaction recognition through positive action representation. Mathematical Problems in Engineering **2013** (2013)
13. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Gmd: Controllable human motion synthesis via guided diffusion models. arXiv preprint arXiv:2305.12577 (2023)
14. Komura, T., Ho, E.S., Lau, R.W.: Animating reactive motion using momentum-based inverse kinematics. Computer Animation and Virtual Worlds **16**(3-4), 213–223 (2005)

15. Kundu, J.N., Buckchash, H., Mandikal, P., Jamkhandi, A., Radhakrishnan, V.B., et al.: Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2724–2733 (2020)
16. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8673–8682 (2023)
17. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023)
18. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. arXiv preprint arXiv:2307.00818 (2023)
19. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Zhe, X., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1144–1154 (2024)
20. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
21. OpenAI: Https://openai.com/blog/chatgpt
22. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
24. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
25. Shum, H.P., Komura, T., Yamazaki, S.: Simulating competitive interactions using singly captured motions. In: Proceedings of the 2007 ACM symposium on Virtual reality software and technology. pp. 65–72 (2007)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
27. Tanaka, M., Fujiwara, K.: Role-aware interaction generation from textual description. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15999–16009 (2023)
28. Tang, J., Wang, J., Hu, J.F.: Predicting human poses via recurrent attention network. Visual Intelligence **1**(1), 18 (2023)
29. Tang, J., Zeng, Y., Fan, K., Wang, X., Dai, B., Chen, K., Ma, L.: Make-it-vivid: Dressing your animatable biped cartoon characters from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6243–6253 (2024)
30. Tanke, J., Zhang, L., Zhao, A., Tang, C., Cai, Y., Wang, L., Wu, P.C., Gall, J., Keskin, C.: Social diffusion: Long-term multiple human motion anticipation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9601–9611 (2023)
31. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)

32. Wang, Z., Wang, J., Lin, D., Dai, B.: Intercontrol: Generate human motion interactions by controlling every joint. arXiv preprint arXiv:2311.15864 (2023)
33. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
34. Xie, Z., Wu, Y., Gao, X., Sun, Z., Yang, W., Liang, X.: Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. arXiv preprint arXiv:2312.10960 (2023)
35. Xu, J., Wang, M., Gong, J., Liu, W., Qian, C., Xie, Y., Ma, L.: Exploring versatile prior for human motion via motion frequency guidance. In: 2021 International Conference on 3D Vision (3DV). pp. 606–616. IEEE (2021)
36. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
37. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
38. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023)
39. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. arXiv preprint arXiv:2312.15004 (2023)
40. Zheng, T., Jiang, P.T., Wan, B., Zhang, H., Chen, J., Wang, J., Li, B.: Beta-tuned timestep diffusion model. In: European Conference on Computer Vision (2024)