Actors Stand By: Storytelling Dual-Character Motion Synthesis

Yingjie Xi
Bournemouth University
Bournemouth, UK
yxi@bournemouth.ac.uk

Boyuan Cheng
Bournemouth University
Bournemouth, UK
bcheng@bournemouth.ac.uk

Rui He
The Hong Kong Polytechnic University
Hong Kong, China
khenryhe@gmail.com

Jinhe Na
Dalian Minzu University
Dalian, China
202311054006@stu.dlnu.edu.cn

Pengjie Wang
Dalian Minzu University
Dalian, China
pengjiewang@gmail.com

Xiaosong Yang
Bournemouth University
Bournemouth, UK
xyang@bournemouth.ac.uk

Jian Jun Zhang
Bournemouth University
Bournemouth, UK
jzhang@bournemouth.ac.uk

Ying Cao*
ShanghaiTech University
Shanghai, China
caoying59@gmail.com



Figure 1: Given an abstract and detailed story description that encompasses fundamental storytelling elements, including event context (green), character intention (red), character relationship (orange), and character emotion (blue), our method is capable of generating the motions of two interacting characters that faithfully tell the story.

Abstract

Text-guided human motion generation has gained significant interest due to its extensive applications in various fields, such as filming, animation and robotics. In this paper, we are interested in a specific form of text-to-motion generation, referred to as storytelling dual-character motion synthesis, which aims to generate the motions of two characters in an interactive scenario to convey the semantics of a given story described by a text prompt. The story descriptions are predominantly composed of abstract, vague concepts, which distinguishes the storytelling motion synthesis from conventional text-to-motion generation that largely relies on concrete action instructions. To solve this task, we present a diffusion-based model that is able to comprehend abstract story text and map it to compelling two-character animations aligned with the input stories. Our model utilizes a novel architectural design that captures intricate interactions between two characters across multiple temporal scales, enabling the generation of highly expressive, storytelling character motions. For training our model, we construct a large-scale two-character motion dataset, consisting of expressive motion sequences of two interacting characters, alongside story-oriented textual descriptions that cover the essential factors of storytelling. Our experiments on the new dataset show that our method can generate the realistic dual-character motions that faithfully convey the given textual stories, significantly outperforming alternative baselines across various metrics. Moreover, we demonstrate the uniqueness and necessity of our dataset for the area of storytelling motion generation of multiple characters, as compared to exiting human motion datasets.

*Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.

MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835729

CCS Concepts

• **Computing methodologies** → **Animation**.

Keywords

Animation, Storytelling, Human Motion, Generative Model

ACM Reference Format:

Yingjie Xi, Boyuan Cheng, Rui He, Jinhe Na, Pengjie Wang, Xiaosong Yang, Jian Jun Zhang, and Ying Cao. 2026. Actors Stand By: Storytelling Dual-Character Motion Synthesis. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835729

1 Introduction

In film, the motions of actors, serving as a form of language for communicating with audiences, are often carefully planned in order to faithfully convey an intended story. The body poses of the actors along with their spatial interactions are critical for storytelling, and variation in these factors can deliver distinct messages to audiences. Given a scene script with abstract descriptions in terms of past and future context, characters' emotional states and intentions, as well as their relationships, professional human actors can determine how to perform based on their understanding of the script, past experience, and even artistic creativity, so that their motions can accurately and efficiently express the intended meaning of the script. This motivates us to ask an intriguing yet challenging question: can computers be equipped with the similar abilities of human actors to generate storytelling 3D motions of *multiple* characters just from a textual story description, with *few or no* concrete action instructions?

Generating realistic 3D human motions has been a longstanding research problem in computer graphics. This field has seen tremendous success with the recent advent of powerful deep generative models. In particular, there is a recent explosion of works on text-to-motion generation, which are able to synthesize highly realistic and diverse human motions from text inputs. While some prior approaches focus on single-character scenarios [8–10, 12, 15, 24, 26, 27, 39, 43, 44, 49, 51], some methods model more complex scenes involving two characters [4, 19, 33]. However, all of these methods rely on relatively short input sentences focusing on specific actions (such as sit, turn left, run, etc.), which differ significantly from story descriptions that are often long and complex, covering much richer concepts.

In this paper, we take a step towards answering the aforementioned question, by solving a task of *storytelling dual-character motion synthesis*: given an input textual story, our aim is to generate the motions of two interacting characters that are able to express the story vividly and faithfully. The core challenges of this task come from several aspects. First, story descriptions (as shown in Figure 1) usually encompass many abstract concepts such as characters' intention relationship, and emotion, and thus the model needs to have a deep understanding of these high-level concepts and how to map them into concrete actions. Second, in addition to the realism of the motions of individual characters, the complex interactions between two characters are necessitated to be plausible, which is vital for storytelling. Third, unlike common actions (e.g., sit and walk) that are consistent across human instances with only small variations, storytelling human motions need to be much more expressive in order to faithfully convey the story. Thus, the model needs to learn a much more complex distribution of human motions.

To address these challenges, we propose a novel approach to story-driven dual-character motion generation. Our approach models the conditional distribution of two-character motion sequences given text prompts, by using a novel diffusion model. The unique part of our model lies in a novel denoising network architecture, which allows for bidirectional information exchange between two characters in motion in both local and global timescales, thereby effectively modeling intricate inter-character relationships across space and time. This architecture design helps our model generate diverse, expressive character motions that well conform to rich semantics in the input stories.

To train our model, we build a large-scale human motion dataset, coined as *StoryDualMotion*, which comprises more than 20k story-motion pairs. Each sample pair in our dataset consists of a story description and a corresponding the motion sequence of two characters in 3D space. Our dataset differs from previous text-to-motion datasets, such as [19, 45], in two primary aspects. First, a large fraction of the motions in our datasets come from professional actors who perform to convey given stories, rather normal people who are instructed to take specific actions. As a result, the motions of our dataset are more expressive, with more nuanced inter-character interaction. Second, as opposed to short captions that describe concrete actions, the textual descriptions of our dataset are more story-oriented, containing various concepts that are critical for storytelling. To the best of our knowledge, our dataset is the first of its kind in the 3D human motion synthesis space.

Through quantitative and qualitative evaluation, we demonstrate our method is capable of generating compelling motion sequences that comply with given story descriptions (as shown in Figure 1), outperforming existing baselines. We envision that our dataset, together with the proposed method, would lay a foundation for future research on the generation of multi-character motions for storytelling. Our code and data are available at: https://github.com/2022yingjie/actors_stand_by

In this paper, our main contributions are as follows:

- We present the first attempt to tackle the problem of storytelling dual-character motion synthesis, which aims to generate the motions of two interacting characters conditioned on input story-oriented text descriptions, paving the way towards the goal of generating animations of multiple characters from only story-like, abstract text prompts, without concrete action instructions.
- We contribute a new two-character motion dataset, which consists of high-quality paired data in the form of stories and their corresponding human motions. The dataset can be used to train and test models for text-guided storytelling human motion synthesis.
- We present a diffusion-based model for synthesizing storytelling dual-character motions from text, with a novel architecture to learn complex inter-character interactions for effective storytelling.

2 Related Work

2.1 Text-based Single-Character Motion Generation

Text-to-motion aims to generate single-character motions from natural language descriptions[8–10, 15, 26]. Over the past few years, several innovative approaches have been introduced for generating single-character motion, showing promising text alignment capabilities. For instance, JL2P [2] introduces an autoencoder with curriculum learning to jointly embed text and pose. AvatarCLIP [11] utilizes a pre-trained VPoser model to generate initial poses, which are then refined through a motion VAE[16]. Text2Action [1] and Language2Pose [2] leverage an encoder-decoder RNN and a shared embedding space, respectively, to correlate language with motion. The works like text2motion [8], TEACH [3], and TEMOS [27], map text to a Gaussian distribution in the latent space. This approach allows for the generation of varied motion outputs by sampling different latent vectors. MotionCLIP [36] trains an autoencoder to align motion with the latent space of CLIP [30], enhancing the fidelity and variety of generated motions. There is a work on addressing the task of story-to-motion [29]. It synthesizes a single-character animation given a story by generating a motion schedule with a large language model, using it to perform motion retrieval and blending the retrieved motion clips.

Recent advancements in diffusion models[34] have led to their widespread adoption for human motion generation. The motion diffusion model (MDM) [37] is trained on a motion space, while MLD [5] employs a latent diffusion model [31] to generate motions based on diverse conditional inputs. T2M-GPT [49] explores a generative framework that combines VQ-VAE[38] with a Generative Pre-trained Transformer (GPT) for motion generation. MotionDiffuse [51] employs a diffusion process combined with a transformer to enhance the diversity of motion samples. Recently, many works have further improved the quality and controllability of text-to-motion diffusion models [7, 13, 14, 18, 20–22, 32, 40, 41, 43, 53, 54].

Unlike all of the aforementioned methods that focus on generating single-character motions, our method aims to generate the animations where two characters are interacting, which necessitates the motions of each character to be realistic and their interactions to be plausible.

2.2 Text-based Two-Character Motion Generation

Compared to single-character motion generation, multi-character motion generation is more challenging, since it not only takes into account each character's motion but also the coordination and interaction between the characters. Xu et al. [47] propose a joint-relation transformer that predicts the future motions of multiple characters in a scene from their historical movement trajectories. MultiAct [17] and ActFormer [46] take semantic motion tags as one of the inputs to guide the model to generate the motion sequences. However, short semantic tags may restrict the generated motions, leading to a lack of variety and creativity in the generated sequences. ComMDM[33] is the first to use textual descriptions for two-character motion generation. InterGen[19] later extends ComMDM with collaborative denoisers to generate more stable text-guided motions. Shan et al.[28] address open-domain multi-task motion generation with a diffusion-based two-stage framework, while FreeMotion[52] adopts a recursive approach with decoupled diffusion modules, enabling control over motions of up to 6 characters. Besides, TIMotion[42] explored the effects of different architectures on two-person motion generation tasks. All of these methods are assumed to process relatively short text inputs that describe what specific actions characters are performing. In contrast, the input story descriptions to our model are much more complex, which are significantly longer and contain numerous storytelling-related abstract concepts. This renders storytelling motion synthesis more difficult than traditional text-to-motion generation.

3 Dataset

We construct a two-character motion dataset, *StoryDualMotion*, tailored for the storytelling dual-character motion synthesis task. Our dataset comprises over 20k two-character motion sequences, each spanning 6 seconds (120 frames); each sequence is accompanied by a textual description. As compared to existing human motion datasets (e.g., [8, 9, 19, 23, 45]), our dataset exhibits two unique features: 1) *expressive motions*: a majority of the motions in our dataset are captured from professional actors (in movies, TV series, stage plays), whose motions have been planned to convey the corresponding stories, and thus are richer and more complex than those obtained from casual videos or captured under controlled settings; 2) *story-oriented text descriptions*: the text descriptions in our dataset cover essential narrative elements, including event context, character intention, character relationship, and character emotion, instead of just specific actions. Figure 2 shows some story-motion pairs from our dataset. See the detailed dataset construction pipeline and a detailed comparison between our dataset and other existing relevant datasets in the supplementary material.

Figure 3 (left) visualizes the distributions of motion sequences of different datasets, including InterHuman, Inter-X and our dataset, in a motion embedding space (obtained by the motion encoder of the MotionBERT [55]). Our dataset has a wider distribution than the other datasets, suggesting its *greater motion diversity*. Figure 3 (middle) shows the motion sequence distributions of the three datasets in an interaction embedding space (will be introduced in Section 5.1). The distribution of our dataset spans a wider region on this space than those of the other datasets, indicating our dataset encompasses *richer interactions* between two characters compared to InterHuman and Inter-X. Figure 3 (right) shows the distributions of text descriptions for InterHuman [19], Inter-X [45] and our dataset in a text embedding space (obtained by LongCLIP[48]). There exists little overlap between our dataset's distribution and those of the other datasets. This highlights the *distinction between story-oriented descriptions and traditional motion captions*.

4 Methodology

Our goal is to learn a generative model of two-character motion sequences conditioned on text conditioning information. Modeling two-character motion is inherently more challenging than single-character motion generation because it needs to account for not only the plausibility of individual character's motions but also



Figure 2: Sample story-motion pairs in our dataset. Each story description covers four aspects: event context (green), character intention (red), character relationship (orange), character emotion (blue).

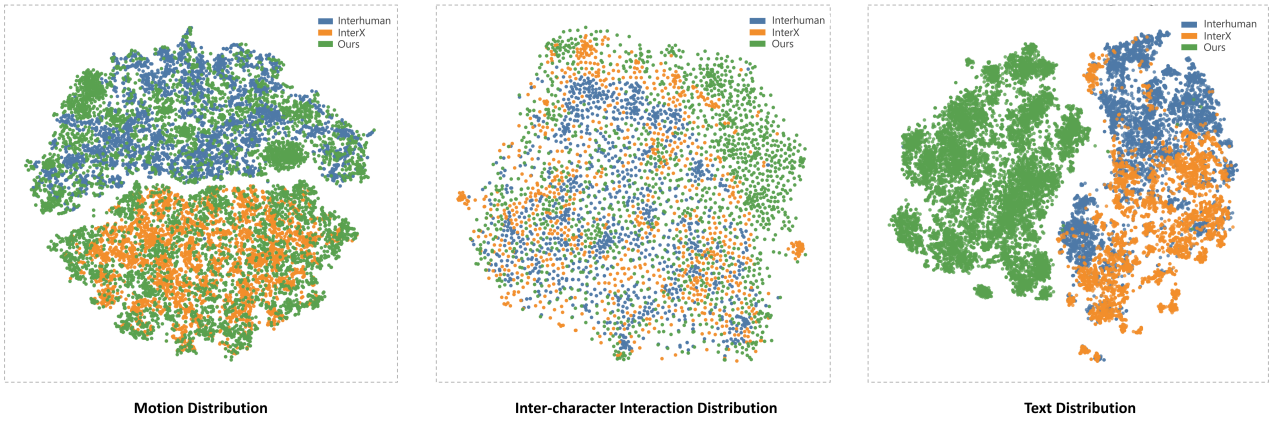


Figure 3: Comparison of our StoryDuaMotion dataset with existing datasets including InterHuman and InterX, in terms of motion distribution (left), inter-character interaction distribution (middle) and text distribution (right).

complex interactions between two characters across spatial and temporal dimensions. Inspired by the recent success of diffusion models in human motion generation [6, 7, 19, 33, 50], we opt to approach our learning problem using a text-conditional diffusion model with a novel network architecture. In the remainder of this section, we first introduce our character motion representation, followed by describing our diffusion model operating in the motion representation space. Then, we elaborate on the architectural design of the denoising network.

4.1 Motion Representation

We represent the motion sequence of N frames for *a single character* c as $X^c = \{x^{c,i}\}_{i=1}^{N}$, where $x^{c,i}$ represents the pose representation of frame i. $x^{c,i}$ comprises the rotations of 22 out of 24 joints in SMPL (with two hand joints excluded) plus the 3D global translation of the root, which can be used to drive a rigged and skinned 3D character (i.e., the SMPL model). The rotation of each joint is encoded as a 3×3 rotation matrix. In our implementation, we only use the first two orthogonal vectors of the rotation matrix, resulting in a 6-D rotation vector. The joint rotations are obtained from our estimated 3D joint positions through inverse kinematics. We further convert the 3D root translation into a 6-D vector with three padding zeros and concatenating it with the rotation vectors. This results in $x^{c,i}$ being a data matrix of size 23×6, which is further flattened into a 138-D vector. Thus, X^c can be represented as a tensor of shape $N \times 138$. We assume that our motion representation for a single character is defined in its local frame, with the root being placed at the origin at the first frame. To put two characters into the same 3D scene, we introduce a vector $D \in \mathbb{R}^{12}$ to represent the orientation and position of each character in the global coordinate system at the first frame. D is composed of a 6-D rotation vector (denoting the character's orientation), a 3-D root position and a 3-D vector of padding zeros. We assume the 3D line connecting the two shoulder points of a character is parallel to the xz-plane and use the angle between the 3D line and the x-axis as the character's orientation. Note that D is computed from the first frame of a two-character motion sequence, but is applied to the remaining frames. We finally extend X_c into shape $N \times 150$ by appending D to the pose representation of each

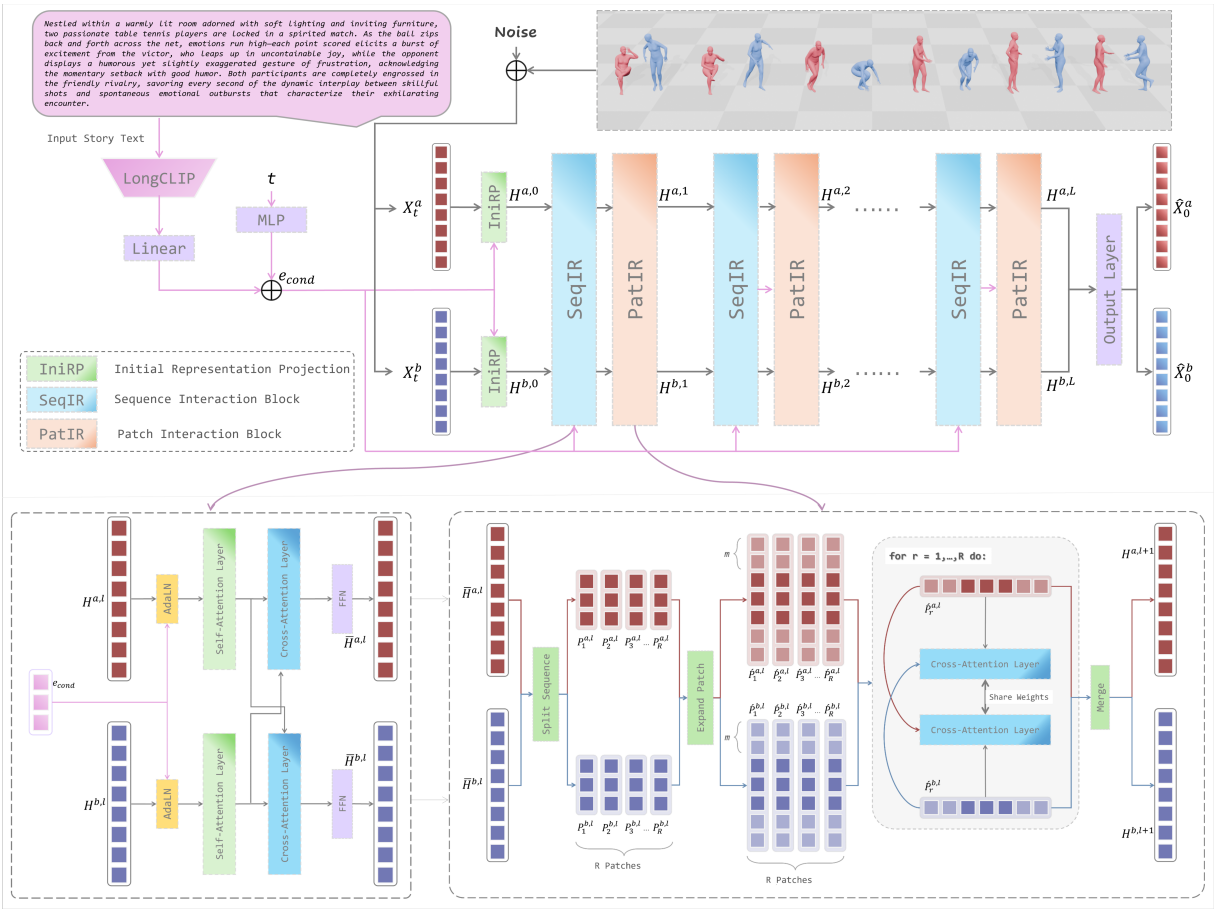


Figure 4: Architecture of the denoising network in our model.

frame. The two-character motion representation X for characters a and b can then be obtained by combining their individual motion representations X^a, X^b as $X = (X^a, X^b)$.

4.2 Our Diffusion Model

Our diffusion model learns a condition distribution $p_\theta(X|s)$ of two-character motions X given textual stories s, by learning the reverse of a forward diffusion process that gradually corrupts a motion sample X_0 into a pure noise $X_T \sim \mathcal{N}(0, I)$ along a fixed T-step Markov chain defined by $q(X_t|X_{t-1})$. The reverse process starts with a random noise and iteratively refines it to recover the original motion sample conditioned on the corresponding story s, based on a learned conditional distribution $p_\theta(X_{t-1}|X_t, s)$. To learn the model, a neural denoising network $\epsilon_\theta(X_t, t, s)$ is trained, which estimates the noiseless sample X_0 from a noisy sample X_t, the text condition c and the diffusion timestep t. X_t is obtained by diffusing X_0 to timestep t along the forward process: $X_t = \Phi(X_0, t, \epsilon_t)$, $\epsilon_t \sim \mathcal{N}(0, I)$. Φ is a sampling operation defined by the forward process marginal distribution $q(X_t|X_0)$. The denoising model is optimized by minimizing a simplified objective:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{X_0, s, t\sim[1,T], \epsilon_t}\left[||X_0 - \epsilon_\theta(X_t, t, s)||_2^2\right]. \quad (1)$$

Note that, following prior work [37], our model predicts X_0 instead of the noise added to X_0.

4.3 Denoising Network Design

The architecture of our denoising network is illustrated in Figure 4. Specifically, the input story is projected into a text embedding using the LongCLIP text encoder [48] followed by a linear layer. The text embedding is added with the timestep embedding to produce a conditional embedding e_{cond} that is then injected into the branches of the two characters. The input representations of the two characters are first processed separately by an *initial representation projection* (IniRP). The resulting representations are then passed through a sequence of interleaved *sequence interaction* (SeqIR) blocks and *patch interaction* (PatIR) blocks, which progressively update the representation of each character and model their interaction. We now describe the key components in the network.

4.3.1 Initial Representation Projection. The input to the initial representation projection module is the noisy motion representation of a character ($X_t^a \in \mathbb{R}^{N\times D}$ or $X_t^b \in \mathbb{R}^{N\times D}$), where N is the number of frames in a motion sequence). This module is composed of a linear layer, followed by a cross-attention layer that attends to the

conditional embedding e_{cond} to integrate the conditioning information. Applying two separate initial representation projection modules to X_t^a and X_t^b, respectively, produces two sequences of embeddings, $H^{a,0} \in \mathbb{R}^{N\times d}$, $H^{b,0} \in \mathbb{R}^{N\times d}$, where d is the internal embedding dimensionality. These embedding sequences are subsequently processed by a series of interleaved sequence interaction blocks and patch interaction blocks.

4.3.2 Sequence Interaction Block. The sequence interaction (SeqIR) block consists of two independent, Transformer-based processing paths, one for each character. Each path comprises an adaptive layer norm (AdaLN) layer (for injecting the conditional embedding e_{cond}), a self-attention layer, a cross-attention layer and a feedforward network (FFN). All the layers except for the AdaLN have a skip connection. For a character, the self-attention captures temporal dependencies between different timesteps within the embedding sequence of the character, while the cross-attention attends to the *full* embedding sequence of the other character, thereby modeling inter-character inte raction at a *global* temporal scale. This design allows the two paths to independently process the input embedding sequences of the two characters: H^a and H^b, while enabling bidirectional information propagation between them. The output of this block is the self-updated and interaction-refined embedding sequences for the two characters: $\bar{H}^a$ and $\bar{H}^b$.

4.3.3 Patch Interaction Block. While the SeqIR block is capable of capturing long-range temporal dependencies between the two characters (through the cross-attention)—how the current pose of one character is impacted by the full sequence of poses of the other character, we argue that modeling their interaction over a small timescale is also crucial for generating detailed, complex and story-aligned character motions. In real-world interactive scenes, a person's pose at the current moment is much more strongly influenced by the poses of another person within a small temporal neighborhood than the poses at relatively distant moments. To account for this intuition, the path interaction (PatIR) block is introduced to better capture inter-character interaction at a *local* temporal scale. The PartIR block, in combination with the SeqIR block, allows our model to capture rich and complex interactions between two characters at multiple temporal scales.

The PatIR block receives as input the output embedding sequences from the SeqIR block, i.e., $\hat{H}^a$, $\hat{H}^b$. Each embedding sequence is first split along the sequence axis into a list of R non-overlapping patches of embeddings: $\hat{H}^a = [P_1^a, P_2^a, \ldots, P_R^a]$, $\hat{H}^b = [P_1^b, P_2^b, \ldots, P_R^b]$, where the patch size is 20. Then, for each path, we extend it bidirectionally to include extra small past and future contexts of size 5, giving rise to two lists of context-enhanced patches: $[\acute{P}_1^a, \acute{P}_2^a, \ldots, \acute{P}_R^a]$, $[\acute{P}_1^b, \acute{P}_2^b, \ldots, \acute{P}_R^b]$. Finally, for each pair of patches $(\acute{P}_r^a, \acute{P}_r^b)$, $r \in 1, \ldots, R$, we model patch-level interactions with cross-attention, updating them to E_r^a and E_r^b, respectively:

$$\begin{aligned} E_r^a &= \text{CrossAttn}(\acute{P}_r^a, \acute{P}_r^b), \\ E_r^b &= \text{CrossAttn}(\acute{P}_r^b, \acute{P}_r^a), \end{aligned} \tag{2}$$

where $\text{CrossAttn}(\cdot,\cdot)$ denotes a cross-attention function, where the query are derived from the first argument, and the key and value are derived from the second argument.

For each updated embedding patch, the extra added contexts are finally removed, and the resulting patches of each character are re-combined along the sequence axis according to their original positions, resulting in an output embedding sequence for the character.

4.3.4 Loss Function. To train our mode, besides the denoising loss $\mathcal{L}_{\text{denoise}}$ (Equation 1), we additionally adopt two auxiliary losses, including a global position loss $\mathcal{L}_{\text{pos}}$ and a global orientation loss $\mathcal{L}_{\text{orien}}$, which we empirically found beneficial to the model performance. By explicitly supervising the global positions and orientations of both characters, the model is encouraged to more accurately predict their spatial configurations, which can benefit the generation of compelling interaction between them. The final loss is written as:

$$\mathcal{L}_{\text{final}} = \lambda_1 \mathcal{L}_{\text{denoise}} + \lambda_2 \mathcal{L}_{\text{pos}} + \lambda_3 \mathcal{L}_{\text{orien}}, \tag{3}$$

where λ_i are weights for balancing different terms. In our implementation, we use $\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.5$.

5 Experiments

Our model uses 1000 diffusion steps, and is trained for 100K iterations with a batch size of 32, and a learning rate of 1e-4. We use 80% of samples in our dataset for training and the rest for testing. Since each sample of our dataset consists of a motion sequence and its corresponding 8 text descriptions. For each motion sequence in the training set, we randomly sample 4 out of 8 text prompts for training and use the remaining ones for evaluation. That is, our testing prompt set is a union of: 1) all the text descriptions of motion sequences in the test set, and 2) the remaining text descriptions of motion sequences that the model has not seen during training.

5.1 Evaluation Metrics

We adopt common evaluation metrics that are used in recent text-to-motion generation works [8], [50], [37], [15]:

- Fréchet Inception Distance (*FID*) quantifies the discrepancy between the distributions of generated and real motions, with a lower value indicating better generation quality.
- *R-Precision* and Multimodal Distance (*MM-Dist*) are metrics for assessing the relevance of generated motions to the input text.
- Multimodality (*MModality*) evaluates the variability in samples resulting from the same text input. *Diversity* measures dissimilarity among generated samples.
- *Density* and *Coverage* have been widely used in recent literature to evaluate the performance of image generation models [25]. We apply them to measure the fidelity and diversity of generated motions.

To calculate these metrics, we follow MotionCLIP [36] to jointly retrain a motion encoder and a text encoder that map two-character motions and text into the same embedding space using the LongCLIP loss that encourages matched motion and text to lie close and unmatched ones to fall apart.

We additionally propose a novel evaluation metric, *InterFID*, to measure the inter-character interaction aspect of generated motions. To compute InterFID, we learn an interaction embedding space and use its embeddings to compute FID between generated and real

Table 1: Quantitative comparison of our model against other methods. All the models are trained on the training split of our StoryDualMotion and evaluated on our test set. For each model, we report the 95% confidence intervals for different metrics by running the model inference 20 times. The best results are in bold.

Method	FID ↓	R-Precision ↑			MM-Dist ↓	MModality ↑	Diversity →	InterFID ↓
		Top 1	Top 2	Top 3				
MDM	16.491 ±.745	0.145 ±.006	0.258 ±.005	0.359 ±.006	5.839 ±.008	3.227 ±.004	12.165±.053	12.096 ±.094
PriorMDM	15.345 ±.084	0.151 ±.006	0.263 ±.007	0.371 ±.007	5.795 ±.008	3.539 ±.004	12.863 ±.064	11.539 ±.068
InterGen	15.049 ±.092	0.154 ±.005	0.268 ±.008	0.378 ±.008	5.737 ±.009	3.676 ±.006	13.150 ±.047	10.508 ±.118
RIG	16.676 ±.073	0.146 ±.005	0.260 ±.008	0.364 ±.007	5.977 ±.006	3.347 ±.008	12.016 ±.058	11.715 ±.079
Ours	**14.977** ±.124	**0.157** ±.005	**0.271** ±.006	**0.383** ±.006	**5.715** ±.011	**3.909** ±.007	**13.292** ±.064	**10.358** ±.167

Table 2: Quantitative comparison of our model against other methods in Density and Coverage. All the models are trained on the training split of our StoryDualMotion and evaluated on our test set. For each model, we report the 95% confidence intervals for different metrics by running the model inference 20 times. The best results are in bold.

Method	Density ↑	Coverage ↑
MDM	0.077 ±.002	0.759 ±.030
PriorMDM	0.083 ±.005	0.893 ±.004
InterGen	0.091 ±.0918	0.906 ±.004
RIG	0.076 ±.003	0.867 ±.029
Ours	**0.096** ±.004	**0.913** ±.016

motions. To quantitatively represent the spatiotemporal interaction between characters a and b in a motion sequence, we define an interaction vector $V = (R_a^r, R_b^r, R_a^j, R_b^j, P_a^r, P_b^r, P_a^j, P_b^j)$, where R_a^r and R_b^r denote their root rotation angles over time, R_a^j and R_b^j denote the rotation angles of their other joints over time, P_a^r and P_b^r denote their root positions over time, and P_a^j and P_b^j denote the positions of their other joints over time. Then, we train a Transformer-based autoencoder on the interaction vectors of our dataset and the InterHuman dataset, to learn the interaction embedding space.

5.2 Comparison to Baselines

We compare our model with four prior text-conditioned two-character motion generation methods: MDM [37], PriorMDM [33], InterGen [19] and RIG [35]. All the models are trained on the training split of our StoryDualMotion dataset and then are tested on the constructed test set as introduced in Section 5.

5.2.1 Quantitative Results. Table 1 presents the quantitative comparison of different methods. Our method outperforms the baselines across all the metrics. These results suggest that, compared with the other methods, our method can generate more *realistic* and *diverse* motions, while exhibiting stronger *storytelling* capabilities—better story adherence. Notably, our method achieves the best InterFID among all the methods, highlighting its superior abilities to replicate complex and rich *interactions* between characters, which is imperative for faithfully conveying input stories. We additionally report the density and coverage scores of different methods in Table 2, which confirms the improved sample fidelity and diversity of our model compared to the baselines.

Table 3: User preference results for different methods in terms of motion realism and story alignment. Each number represents that fraction of time one method is selected as the best among all the methods in terms of one aspect.

Method	Motion Realism	Story Alignment
MDM	4.50 %	4.50%
PriorMDM	13.60 %	9.10 %
InterGen	36.40 %	31.80%
RIG	4.50 %	4.50 %
Ours	**40.90 %**	**50.00%**

5.2.2 Qualitative Results. Figure 5 show a visual comparison of different methods. MDM and RIG struggles with understanding the complex semantics in the input stories, producing the motion sequences that are less relevant to the input text. Furthermore, their results often lack varied character motions and meaningful interaction between two characters. PriorMDM and InterGen are capable of synthesizing richer motions and inter-character interactions, but fall short of accurately conveying the key information in the input text. In contrast, our model shows noticeably improved generation quality compared to the previous methods, generating complex and diverse motions that well align with the input stories while encompassing a variety of semantically meaningful inter-character interactions. In particular, the input text depicts a "comba" between two "opponent" with a "tense" atmosphere. The results of all the existing methods suffer from significant misalignment with the textual input. MDM generates few motions for the two characters, which does not properly express the "tension" of the scene. The outputs of PriorMDM, InterGen and RIG all have two major issues: first, the two character do not face each other; second, only one character is performing some actions, with the other one almost standing still. These issues render their generated inter-character interactions irrelevant to the depicted "confrontation" scenario. In contrast, our method exhibits considerably better text alignment, creating a motion sequence that show the adversarial process of the two characters.

5.3 User Study

To more comprehensively gauge the generation capability of our model, we run a user study comparing our model with the four baselines in terms of motion realism and story alignment. Our study involves 22 subjects recruited from a university. We randomly select

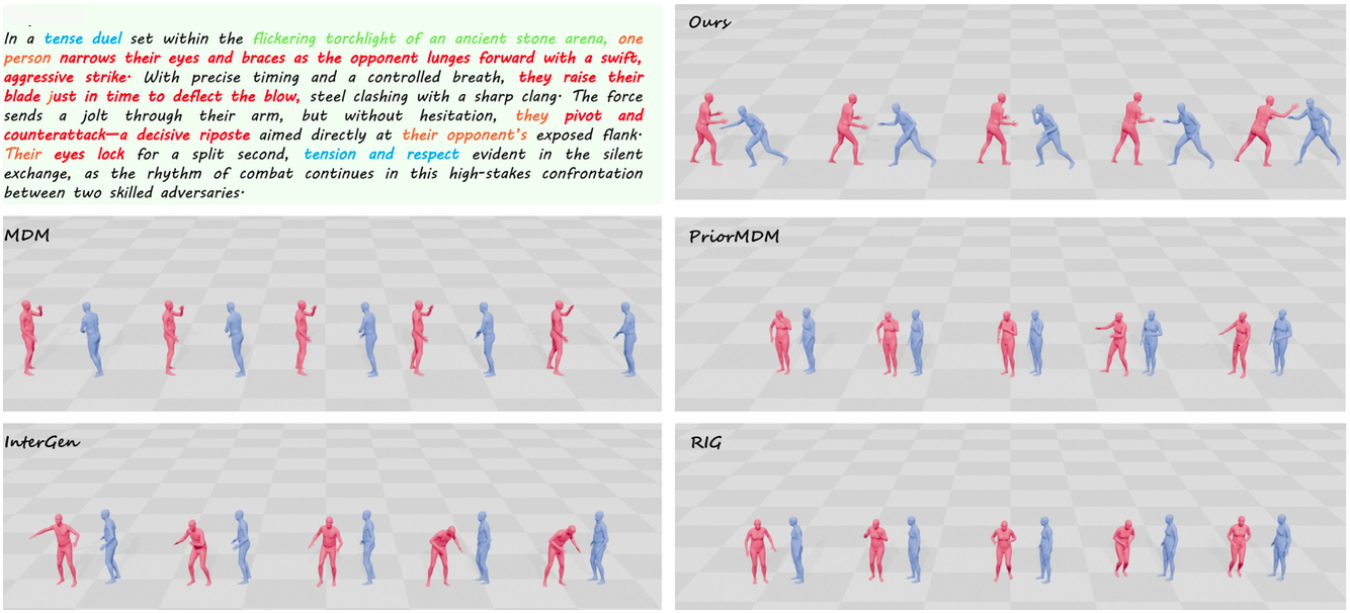


Figure 5: Visual comparison of different methods. The input story description includes event context (green), character intention (red), character relationship (orange), and character emotion (blue).

Table 4: Ablation study results. All the models are trained on our StoryDualMotion. For the ablation experiments, we report the 95% confidence intervals for different metrics by running the model inference 20 times.

Method	FID ↓	R-Precision ↑			MM-Dist ↓	MModality ↑	Diversity →	InterFID ↓
		Top 1	Top 2	Top 3				
SeqIR-only	15.552 ±.118	0.152 ±.006	0.265 ±.007	0.374 ±.008	6.004 ±.009	3.602 ±.007	13.108 ±.041	11.053 ±.129
ParIR-only	15.948 ±.132	0.147 ±.007	0.257 ±.008	0.364 ±.009	5.852 ±.010	3.754 ±.008	13.204 ±.046	11.452 ±.141
w/o $\mathcal{L}_{\text{pos}}$+$\mathcal{L}_{\text{orien}}$	15.457 ±.133	0.150 ±.006	0.262 ±.007	0.375 ±.004	5.792 ±.013	3.860 ±.009	13.145 ±.047	10.802 ±.129
Ours	**14.977** ±.124	**0.157** ±.005	**0.271** ±.006	**0.383** ±.006	**5.715** ±.011	**3.909** ±.007	**13.292** ±.064	**10.358** ±.167

10 input stories from our test set. The subjects are presented with a story text, along with five animations generated by different methods; the order of the animations are randomized. The subjects are asked to choose: 1) which one has the most realistic motion and 2) which best matches the given story. We present the results in Table 3. For motion realism, our method is preferred by the subjects 40.9% of the time, surpassing all the other methods. For story alignment, our method is chosen by the subjects 50.0% of the time, outperforming the second best method (31.8%) by a large margin. This highlight the excellent storytelling abilities of our method again.

5.4 Ablation Study

We first test the importance of the two core components in our model architecture, i.e., *SeqIR* and *PatIR* blocks, by training and evaluating two variants of our full model. The first variant (*SeqIR-only*) solely uses the SeqIR blocks (with the ParIR blocks removed) in the denoising network, while in the second first variant (*ParIR-only*), only the ParIR blocks are included (with the SeqIR blocks removed) in the network, and the conditional embedding $\mathbf{e}_{\text{cond}}$ is incorporated into each ParIR block through AdaLN as in the SeqIR blocks. The results are shown in Table 4. Removing either SeqIR or PatIR leads to performance degradation across all the metrics compared to the full model. This suggests that both components are essential for the generation of high-fidelity, diverse and storytelling motions. Moreover, when the auxiliary losses in Equation 3 are removed, our model's performance drops, implying their effectiveness.

6 Conclusion

In this paper, we make the first attempt to study the task of storytelling dual-character motion generation, which aims to synthesize 3D motion sequences of two interacting characters to tell a given story. To solve the task, we contribute a new two-character motion dataset dedicated to this task, which consists of expressive motion sequences and story-oriented text descriptions, and propose a diffusion model with a novel architecture that captures inter-character interactions at multiple temporal scales. We show the effectiveness of our model in synthesizing animations that faithfully express given stories. We believe that our work can inspire future research in developing more sophisticated methods for storytelling human motion generation.

References

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

[2] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 719–728.

[3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 414–423.

[4] Jianqi Chen, Panwen Hu, Xiaojun Chang, Zhenwei Shi, Michael Kampffmeyer, and Xiaodan Liang. 2024. Sitcom-crafter: A plot-driven human motion generation system in 3d scenes. *arXiv preprint arXiv:2410.10790* (2024).

[5] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[6] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9760–9770.

[7] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[8] Chuan Guo, Chunyuan Zhang, Jie Wu, and Leonid Sigal. 2022. Generating Diverse and Natural 3D Human Motions from Texts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5152–5161.

[9] Chuan Guo, Yuexin Zou, Weipeng Huang, Qing Chen, Chen Change Loy, and Zhiwei Zhang. 2020. Action2Motion: Conditioned Generation of 3D Human Motions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2021–2029.

[10] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*. Springer, 580–597.

[11] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).

[12] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. 2025. SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing. *arXiv preprint arXiv:2503.13836* (2025).

[13] Minjae Jeong, Yechan Hwang, Jaejin Lee, Sungyoon Jung, and Won Hwa Kim. [n. d.]. HGM^3: Hierarchical Generative Masked Motion Modeling with Hard Token Mining. In *The Thirteenth International Conference on Learning Representations*.

[14] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, et al. 2025. PersonaBooth: Personalized Text-to-Motion Generation. *arXiv preprint arXiv:2503.07390* (2025).

[15] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8255–8263.

[16] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. 2023. Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1231–1239.

[18] Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. 2025. SimMotionEdit: Text-Based Human Motion Editing with Motion Similarity Prediction. *arXiv preprint arXiv:2503.18211* (2025).

[19] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision* (2024), 1–21.

[20] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. 2024. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1399–1408.

[21] Jinpeng Liu, Wenxun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, and Xin Tong. 2024. Plan, Posture and Go: Towards Open-Vocabulary Text-to-Motion Generation. In *European Conference on Computer Vision*. Springer, 445–463.

[22] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. 2024. Bridging the gap between human motion and action semantics via kinematic phrases. In *European Conference on Computer Vision*. Springer, 223–240.

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5442–5451.

[24] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 2024. Rethinking Diffusion for Text-Driven Human Motion Generation. *arXiv preprint arXiv:2411.16575* (2024).

[25] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. (2020).

[26] Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10985–10995.

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

[28] Yijun Qian, Jack Urbanek, Alexander Hauptmann, and Jungdam Won. 2023. Text Motion Translator: A Bi - Directional Model for Enhanced 3D Human Motion Generation from Open - Vocabulary Descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2 – 17. doi:10.1109/ICCV.2023.0

[29] Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. 2023. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 Technical Communications*. 1–4.

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

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[32] Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. 2025. MixerMDM: Learnable Composition of Human Motion Diffusion Models. *arXiv preprint arXiv:2504.01019* (2025).

[33] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. 2023. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418* (2023).

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[35] Mikihiro Tanaka and Kent Fujiwara. 2023. Role-aware Interaction Generation from Textual Description . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 15953–15963. doi:10.1109/ICCV51070.2023.01466

[36] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*. Springer, 358–374.

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[38] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[39] Tao Wang, Zhihua Wu, Qiaozhi He, Jiaming Chu, Ling Qian, Yu Cheng, Junliang Xing, Jian Zhao, and Lei Jin. 2025. StickMotion: Generating 3D Human Motions by Drawing a Stickman. *arXiv preprint arXiv:2503.04829* (2025).

[40] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. 2024. MotionGPT-2: A General-Purpose Motion-Language Model for Motion Generation and Understanding. *arXiv preprint arXiv:2410.21747* (2024).

[41] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick WB Li, Ziyao Zhang, and Xiaohui Liang. 2025. Fg-T2M++: LLMs-Augmented Fine-Grained Text Driven Human Motion Generation. *International Journal of Computer Vision* (2025), 1–17.

[42] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. 2025. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7169–7178.

[43] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2024. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. *arXiv preprint arXiv:2405.17013* (2024).

[44] Yingjie Xi, Jian Jun Zhang, and Xiaosong Yang. 2025. PMG: Progressive Motion Generation via Sparse Anchor Postures Curriculum Learning. *arXiv preprint arXiv:2504.16722* (2025).

[45] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. 2024. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22260–22271.

[46] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. 2023. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2228–2238.

[47] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. 2023. Joint-Relation Transformer for Multi-Person Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9816–9826.

[48] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*. Springer, 310–325.

[49] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052* (2023).

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).

[51] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[52] Zhikai Zhang, Yitang Li, Haofeng Huang, Mingxian Lin, and Li Yi. 2025. FreeMotion: MoCap-Free Human Motion Synthesis with Multimodal Large Language Models. In *European Conference on Computer Vision*. Springer, 403–421.

[53] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. 2024. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487* (2024).

[54] Kaifeng Zhao, Gen Li, and Siyu Tang. 2024. Dart: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260* (2024).

[55] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.