HSPI The Open Access Publisher

JAIRI ISSN 3069-762X JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION

Research Article

Trajectory-Aware Motion Generation for Enhanced Naturalness in Interactive Applications

Xuan Liu[1,2#], Shaojun Yuan[1#], Zhiyang Zhang[1], Xiangyu Qu[1], Yidian Liu[3], Chaomurilige[1*], Zheng Liu[1*] and Shan Jiang[1]

[1]School of Information Engineering, Minzu University of China, Zhongguancun South Street No.27, Beijing, 100081, China

[2]Hainan International College of Minzu University of China, Li'an International Education Innovation pilot Zone, Hainan, 572499, China

[3]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Ministry of Education, Minzu University of China, Beijing, 100081, China

[#]These authors contributed equally to this work.

Abstract

Human motion generation is a pivotal task in the field of data generation, with trajectory-guided methods emerging as a prominent approach due to their ability to provide precise control over motion outcomes. However, achieving a balance between motion naturalness and trajectory accuracy remains a significant challenge. In this paper, we present a novel method, Trajectory-Aware Motion Generator (TAMG) that optimally addresses this challenge. TAMG integrates third-order dynamic features, namely position, velocity, and acceleration, to enhance the naturalness of generated motions while maintaining precise trajectory control. We propose a multimodal feature fusion strategy that combines biomechanical features to ensure accurate motion representation, alongside a sparse sampling strategy based on motion importance distribution to focus on key phases of joint motion. The effectiveness of TAMG is validated through extensive experiments, which demonstrate its superior performance in both trajectory accuracy and motion quality compared to existing methods. This approach offers a simple, effective solution for interactive motion generation tasks, advancing the state of the art in trajectory-guided motion generation.

More Information

***Address for correspondence:** Chaomurilige, Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Ministry of Education, Minzu University of China, China,
Email: chaomurilige@muc.edu.cn

Zheng Liu, School of Information Engineering, Minzu University of China, Zhongguancun South Street No.27, Beijing, 100081, China,
Email: liuzheng@muc.edu.cn

Submitted: November 04, 2025
Approved: November 17, 2025
Published: November 18, 2025

How to cite this article: Liu X, Yuan S, Zhang Z, Qu X, Liu Y, Chaomurilige, et al. Trajectory-Aware Motion Generation for Enhanced Naturalness in Interactive Applications. J Artif Intell Res Innov. 2025; 1(1): 085-093. Available from: https://dx.doi.org/10.29328/journal.jairi.1001010

Keywords: Motion generation; Motion naturalness; Motion prediction; Interactive data generation

1. Introduction

Human motion generation is a significant field within data generation, with trajectory-guided motion generation emerging as a prominent research focus due to its precise control over results. This method has broad application in areas such as virtual character animation [1-4] and intelligent human-computer interaction [5-8]. By integrating spatial trajectory constraints and behavioral semantic instructions, this approach generates complex motions that adhere to both trajectory limits and behavioral semantics.

In motion generation models, the masking mechanism is crucial: it promotes data diversity, mitigates overfitting, and improves the model's robustness. As a key component of the training process, the masking strategy significantly influences the quality of generated results. Numerous models incorporate masking modules. For instance, MDM [9] improves the model's generalization by randomly masking all spatio-temporal units. Omnicontrol [10] masks features in frames lacking valid control signals, enabling the model to concentrate on valid signals, ensuring motion generation is consistent with control instructions. However, in practical applications, problems persist, such as motions deviating from preset trajectories or violating physical laws. This is primarily due to the simplistic random masking strategy that excessively focuses on joint trajectories, neglecting the kinematic dynamics of real motion. This limitation results in insufficient naturalness in generated motions. Therefore, ensuring both trajectory accuracy and motion naturalness remains a major challenge in trajectory-guided motion generation.

To address this, we propose the Trajectory-Aware Motion Generator (TAMG), a model designed to balance

trajectory accuracy with motion naturalness by co-optimizing biomechanical multimodal features. Specifically, we introduce a multimodal feature extraction module based on the third-order kinetic properties (position, velocity, and acceleration) of joint motion. This module quantifies kinetic parameters of joint movements and uses a multimodal feature fusion mechanism to combine the extracted features with predefined weights. Additionally, we apply a sparse sampling strategy based on motion importance distribution to accurately mask critical motion phases. This method enhances the naturalness of generated motions while maintaining sub-centimeter trajectory accuracy with minimal computational overhead.

We validate the effectiveness of TAMG through extensive experiments, optimizing the balance between trajectory accuracy and motion naturalness. This approach successfully reduces trajectory-guided motion offsets. Compared to existing methods, TAMG significantly improves generation quality and offers new insights for research in related areas.

In summary, the main contributions of this paper are:

1. We examine the coupling mechanism of multimodal motion features in trajectory-guided motion generation, highlighting its significance in interactive motion models through extensive experimentation.

2. We introduce TAMG, a straightforward yet effective motion generation method, combining multimodal feature fusion with sparse masking of critical motion phases. This method effectively balances trajectory constraints with biomechanical rationality and improves the coordination of generated motions.

3. We demonstrate the effectiveness of our approach through extensive experiments, achieving a 29.6% improvement in FID performance on the HumanML3D dataset.

2.Related work

2.1 Human motion generation

Text-to-motion generation tasks [11-16] aim to create realistic human motion sequences based on natural language descriptions. Early approaches primarily utilized Generative Adversarial Networks (GAN) [17-19] or Variational Auto-Encoders (VAE) [20-22]. In recent years, diffusion models have gained prominence for multimodal motion generation tasks due to their powerful generative capabilities. For instance, MotionDiffuse [23], the first text-driven motion diffusion model, enhances multimodal representation through fine-grained part control. MDM [9] further strengthens language-driven control for generating complex motions, while TLControl [24] combines linguistic descriptions with trajectory signals, utilizing potential spatial decoupling to generate high-fidelity motions. For trajectory control, GMD [25] propagates trajectory signals to neighboring nodes by introducing simple spatial constraints, while OmniControl [10] incorporates multi-joint spatial constraints for better coordination. MotionLCM [26] combines latent space control with ControlNet [27] to achieve efficient trajectory control.

Diffusion-based generation methods [28-33], renowned for their robust generative abilities and diversity, have been widely adopted for multimodal motion generation tasks. For example, MotionDiffuse [23], the first text-based motion diffusion model, enriches multimodal generation by enabling fine-grained control of body parts. MDM [9] enhances language-driven control, allowing users to generate complex actions using natural language. TLControl [24] integrates language descriptions with trajectory signals, generating high-fidelity motions through latent space decoupling. However, these methods often fall short in terms of control when actions involve significant dynamic changes or when multimodal input signals conflict. Moreover, the modeling of dependencies between body keypoints in trajectory-guided generation tasks remains insufficiently detailed. These models generally assume that the trajectory uniformly guides all keypoints, which can lead to conflicts between local actions and global consistency.

2.2 Masking strategies and kinematic modeling

2.2.1 Masking strategies: Masking strategies are widely utilized in self-supervised learning for data representation tasks. MAE [34] introduced a large-scale mask reconstruction method to effectively learn transferable visual representations in the image domain. In the human motion domain, similar masking strategies have been applied to self-supervised representation learning [35] and motion prediction [36]. However, most existing strategies rely on uniformly randomized masks, overlooking the temporal dynamics and interjoint dependencies of human motion. This limitation can lead to a lack of naturalness in the generated motions during motion generation processes.

2.2.2 Kinematic modeling: Human motions are commonly represented as time series of joint rotations or 3D joint positions. Biomechanical analysis by Winter [37] demonstrates that the naturalness of motion is primarily determined by the multimodal coupling of positional accuracy, velocity continuity, and acceleration smoothness. Several approaches have been proposed to incorporate physical laws to enhance naturalness. For example, PhysDiff [38] corrects global motion errors (such as penetration) using a physics simulator but neglects joint-level dynamic properties. Li, et al. [39] proposed a velocity/acceleration feature extraction method based on difference operators, which, while incorporating rich kinematic knowledge, is limited to short-term prediction tasks. Recently, FinePhys [40] introduced a three-bias physics fusion framework based on Euler-Lagrange equations. Despite advancements in rigid-body dynamics modeling, this framework imposes global constraints that

hinder its ability to capture biomechanical coordination among joints and does not optimize synergistically with conditional injection architectures.

While existing trajectory-guided motion generation research has made notable progress in model architecture and trajectory control, current masking strategies [9,10,36,41,42] often fail to adequately model the intrinsic biomechanical properties of human motion, particularly the coupled effects of position, velocity, and acceleration dynamics. Meanwhile, related kinematic modeling studies focus either on global physical corrections, ignoring joint-level dynamics, or on kinetic features that are inadequately integrated into the core mechanisms (e.g., masking strategies) of generative models. These limitations hinder the effective synergy of trajectory accuracy and motion naturalness. To address these issues, we propose TAMG, a model designed to optimally balance trajectory accuracy and motion naturalness, providing a simple and effective solution for motion generation.

3. Method

3.1 Overview

In this paper, we propose the Trajectory-Aware Motion Generator (TAMG), which integrates multimodal feature fusion with sparse masking of key motion phases to address the challenge of balancing trajectory accuracy and motion naturalness in trajectory-guided motion generation. The proposed method achieves high-precision trajectory control and natural motion generation by incorporating the third-order dynamic properties (position, velocity, and acceleration) of joint motion and optimizing the masking strategy through multimodal biomechanics features. The model architecture is illustrated in Figure 1. Our approach consists of the following steps:

1. **Multimodal feature extraction for joint motion:** We design a multimodal feature extraction module based on three biomechanical features (position, velocity, acceleration) to quantify the kinematic parameters of joint motion.
2. **Multimodal feature fusion and importance calculation:** We compute a multimodal importance measure for each moment by weighting and fusing the features from different modalities with predefined weights.
3. **Sparse masking strategy based on importance distribution:** Using the computed multimodal importance, we apply a sparse sampling masking strategy to accurately control the key phases of joint motions, enhancing both the naturalness of the generated motions and trajectory accuracy.

3.2 Multimodal masking strategy

The core of our multimodal masking strategy lies in modeling the third-order dynamics of the joint motions and generating sparse masks of the key motion phases through weighted fusion. In the specific implementation, the velocity and acceleration of the joints are first computed to extract the motion features of each joint in the time dimension. These features are combined with positional features to generate multimodal joint motion information.

3.2.1 Feature extraction: Given the input joint motion sequence $J \in R^{T\times N\times 3}$ (T denotes the number of time frames, N the number of joints, and 3 the 3D coordinates), we first extract third-order dynamic features:

Position feature: Directly uses the original joint

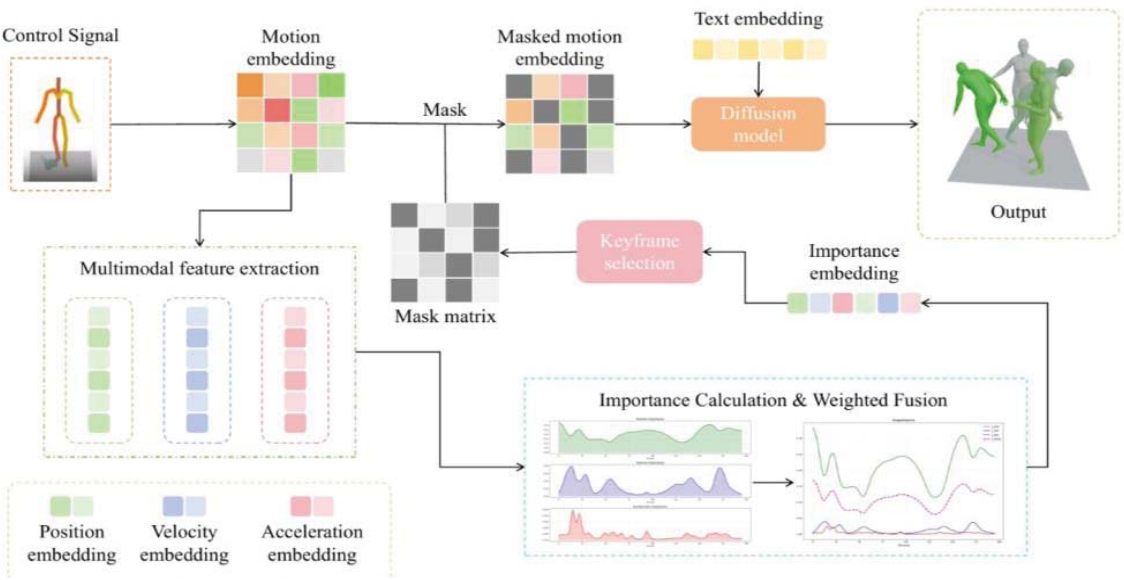


Figure 1: Overview of TAMG.

coordinates $P_t = J_t$, representing static spatial position.

Velocity feature: Computes joint motion velocity through first-order differentiation:

$$vt = \begin{cases} \left\| J_t - J_{t-1} \right\|_2, & t > 1 \\ 0 & , t = 1 \end{cases} \quad (1)$$

Acceleration feature: Computed based on second-order differentiation of velocity:

$$a_t = \begin{cases} \left\| v_t - v_{t-1} \right\|_2, & t > 2 \\ 0 & , t \le 2 \end{cases} \quad (2)$$

This third-order feature extraction is based on biomechanical principles: motion naturalness is essentially determined by the coupled effects of position accuracy, velocity continuity, and acceleration smoothness. The velocity feature quantifies the instantaneous rate of motion change, while the acceleration feature reveals the smooth transition characteristics of motion, together forming the dynamic basis of motion fluency.

3.2.2 Materiality calculation: We calculate the motion importance for each frame based on the third-order dynamic characteristics (position, velocity, acceleration) of joints. Specifically, the motion importance is computed through deviation from the mean:

$$I_t = \left\| P_t - \bar{P} \right\|_2 \quad (3)$$

Where P is the mean of all joint position data, and $\|\cdot\|_2$ denotes the Euclidean norm.

We designed several different sampling strategies to evaluate their impact on generation results (see ablation study in Sec 4.4.1).

1. Gaussian filtering enhancement: Enhances Gaussian smoothing of motion importance to better capture details of motion changes.

$$I_t = \text{Gaussian Filter}\left(\left\| P_t - \bar{P} \right\|_2, \sigma \right) \quad (4)$$

where σ is the standard deviation of the Gaussian filter, typically $\sigma = \min\left(5, \frac{T}{5}\right)$, ensuring a smoothing effect.

2. Peak focusing: Emphasizes motion peaks by amplifying joint movements with significant changes to increase sampling probability at key motion moments:

$$I_t = \left(\left\| P_t - \bar{P} \right\|_2 \right)^2 \quad (5)$$

Squaring the motion deviation emphasizes joint movements with larger changes.

3. Temporal smoothing: Enhances temporal continuity by computing temporal differences to improve motion smoothness and reduce short-term fluctuations.

$$I_t = \left\| P_t - P_{t-1} \right\|_2 \quad (6)$$

Computing motion differences between adjacent frames focuses on the smoothness of continuous motion.

3.2.3 Fusion sampling: We perform weighted fusion of position, velocity, and acceleration motion importance to obtain a comprehensive importance distribution. This distribution reflects the influence of motion features at different periods and joints on generated motions. Let I_t^{pos}, I_t^{vel}, and I_t^{acc} denote position, velocity, and acceleration motion importance, respectively. The total motion importance is:

$$It = \omega_{pos} \cdot I_t^{pos} + \omega_{vel} \cdot I_t^{vel} + \omega_{acc} \cdot I_t^{acc} \quad (7)$$

where ω_{pos}, ω_{vel}, and ω_{acc} are, respectively, weight parameters for position, velocity, and acceleration, satisfying

$$\omega_{pos} + \omega_{vel} + \omega_{acc} = 1 \quad (8)$$

Figure 2 shows the motion importance of a randomly selected dataset. Figure 2a illustrates the calculated importance of position, velocity, and acceleration, and Figure 2b shows the importance of fusion versus them. Sec 4.4.2 and Sec 4.4.3 discuss the impact of different parameters on generation results.

For each time t, we determine whether to sample that frame based on the computed motion importance I_t. The sampling probability p_t is proportional to it:

$$pt = \frac{I_t}{\sum_{t=1}^{T} I_t} \quad (9)$$

where $\sum_{t=1}^{T} I_t$ is the total importance across all time steps t, used to normalize the sampling probability for each frame. We sample a new frame sequence from all frames according to the probability p_t.

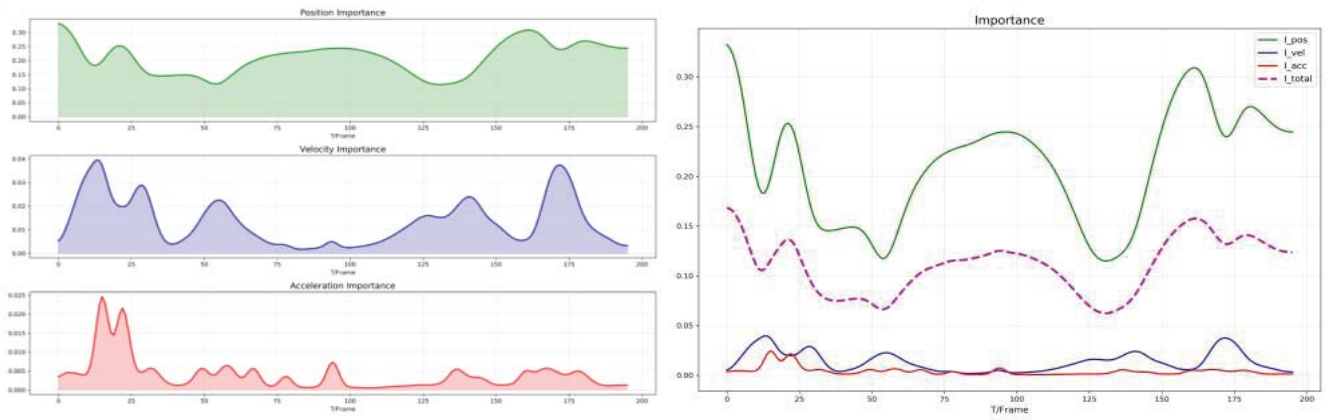
3.3 Training strategy

During the training process, we introduced a random masking strategy to improve the generalization ability of the model. In the application of the masking strategy, the model first normalizes the motion data of each frame to ensure that the motion data of different joints are compared at the same scale. Then, the joints are selected and masks are applied through a multimodal masking strategy to ensure that the model can focus on the most important motion phases during the learning process.

During the training process, we also used a random selection of joints for masking to simulate the uncertainty in the real scenario. By properly masking the joints, the model can avoid over-fitting and generate more natural motion sequences while maintaining trajectory accuracy.

4. Experiment

We performed several experiments with human motion generation on the model to initially assess its performance in terms of balance trajectory accuracy and naturalness of motion. The experiments were conducted on the widely used



(a) The motion importance.

(b) The fusion importance.

Figure 2: Examples of the motion importance. We randomly selected a dataset, computed its motion significance, and visualized it.

HumanML3D [43] motion dataset and included comparative analyses with existing benchmark methods. Additionally, we performed ablation studies to investigate the specific impact of each module on the generation quality.

4.1 Datasets

We selected HumanML3D [43] as the dataset for our experiments to ensure that the model is evaluated across various motion scenarios. HumanML3D contains 14,646 motion sequences extracted from the AMASS [44] and HumanAct12 [45] datasets, each with accompanying textual annotations. It covers a broad range of human actions such as daily activities (e.g., 'walking', 'jumping'), sports (e.g., 'swimming', 'playing golf'), acrobatics (e.g., 'cartwheel'), and artistry (e.g., 'dancing'). HumanML3D data follows the SMPL skeleton structure with 22 joints. In experiments, we consider the pelvis, left foot, right foot, head, left wrist, and right wrist, given their common usage in interactions with objects and the surrounding scene.

4.2 Evaluation metrics

We adopt the evaluation protocol from [43]. We use three main metrics, Frechet Inception Distance (FID), R-Precision, and Diversity, to assess the realism and diversity of the generated motions. FID measures the quality of the generated motions by calculating the distance between the generated and real distributions in a pre-trained feature space. R-Precision evaluates the alignment between the generated motions and the textual descriptions by matching each generated motion sequence with 32 text descriptions (1 correct description and 31 random). The Euclidean distance between motion and text embeddings is computed, and accuracy is reported based on the top-3 matches. Diversity measures the variability in generated motions by randomly pairing sequences in the test set and calculating the average cumulative difference within each pair. To evaluate the controlling performance, following [25], we report the foot skating ratio as a proxy for the incoherence between trajectory and human motion and physical plausibility.

Additionally, to quantify the performance of the model in trajectory control, we adopted control accuracy metrics in calculating Trajectory Error, Location Error, and Average Error to evaluate the 3D joint position control in key frames. Trajectory Error measures the precision of trajectory tracking: if any joint position in a key frame deviates beyond a set threshold, the trajectory is considered unsuccessful. Location Error calculates the percentage of key frame positions that fail to fall within a specified proximity. Average Error quantifies the overall control accuracy by computing the mean Euclidean distance between the joint positions in the generated motion and the given control trajectories at each key frame motion step.

4.3 Comparison with state-of-the-art methods

In Table 1, we compared the performance of our method with that of the Omni control method on multiple joints. The experimental results show that our method outperforms Omnicontrol on most of the evaluation metrics, especially on FID, R-precision, and motion diversity. Specifically, our method effectively improves the realism and accuracy of motions, while reducing the foot-slip ratio and significantly improving the physical plausibility. In addition, our method also performs well in terms of trajectory control and joint positioning accuracy, especially in terms of trajectory error and positioning error of the right foot and head, which are both greatly optimized. Overall, our method demonstrates enhanced capabilities in terms of generation quality, motion recognition, and physical reasonableness, validating the effectiveness of multi-feature fusion and dynamic weighting strategies to generate more natural and smooth motions while ensuring accuracy.

4.4 Ablation study

We conducted several ablation experiments on

HumanML3D [43] to verify the effectiveness of our model design choices.

4.4.1 Effect of different sampling strategies: In order to evaluate the impact of different sampling strategies on the generation results, we designed a series of ablation experiments to observe their performance in motion generation by varying the details of the sampling strategies.

In Table 2, we present a detailed comparison of the different variants of the sampling strategy to verify their effectiveness in terms of generation quality. Specifically, base (first row) employs only bias computation; Gaussian-3 (second row) uses a Gaussian filter with a standard deviation of 3; Gaussian-5 (third row) employs a Gaussian filter with a standard deviation of 5; Peak (fourth row) emphasizes joint motion variations by amplifying motion peaks; and Temporal (fifth row) enhances temporal smoothing by calculating the motion differences of neighboring frames.

In FID and control accuracy (Traj. err., Loc. err.) and other metrics in the experiments, Gaussian filtering (σ = 3) significantly reduces the FID value and the local error, indicating that moderate smoothing can effectively improve the fidelity of details. However, when the standard deviation was increased to 5, excessive smoothing of the Gaussian filter led to a significant deterioration of the FID values and produced large trajectory distortions, suggesting that excessive smoothing may lead to loss of motion details. The peak focusing strategy performs well in R-precision and Foot skating, but it is also accompanied by an increase in trajectory error, reflecting the limitations of this strategy in modelling continuous motion. In contrast, the temporal continuity enhancement strategy shows good overall performance on all evaluation metrics. These experimental results suggest that the choice of sampling strategy must be weighed against the task requirements: Gaussian filtering is suitable for high-fidelity scenes, peak focusing excels at capturing key frames, and temporal continuity.

The enhancement strategy has the most significant advantage in terms of overall generation quality. This further validates the critical impact of multi-scale motion feature modeling on generation quality.

4.4.2 Effect of different dynamic features: To evaluate the importance of position, velocity, and acceleration features in motion generation, we set up multiple distinct models and experimentally compared the impact of each feature on the generated results.

In Table 3, we systematically evaluate the impact of different dynamic features on the quality of generated motion by comparing models utilizing the full set of dynamic features against models employing only a single feature. The experimental results show that when relying only on a single feature, the FID values deteriorate significantly when using pure position features (second row), although the trajectory error performs moderately well, suggesting that the spatial localization features need to be compensated by the velocity and acceleration features during the generation process. Specifically, the pure position weights can achieve

Table 1: Quantitative results of comparison with state-of-the-art methods on.

HumanML3D Test Set Results.								
Method	**Joint**	**FID ↓**	**R-precision ↑(Top-3)**	**Diversity →**	**Foot-skating ratio ↓**	**Traj. err. ↓ (50 cm)**	**Loc. err. ↓ (50 cm)**	**Avg. err. ↓**
real	–	0.002	0.797	9.503	0.000	0.000	0.000	0.000
Omnicontrol	Pelvis	0.323	0.691	9.854	0.0431	0.0781	0.0131	0.0635
Ours (on all)	Pelvis	0.234	0.702	9.715	0.0594	0.0557	0.0096	0.0585
Omnicontrol	Left foot	0.292	0.689	9.855	0.0538	0.1230	0.0166	0.0623
Ours (on all)	Left foot	0.252	0.703	9.576	0.0628	0.0977	0.0124	0.0555
Omnicontrol	Right foot	0.307	0.693	9.901	0.0518	0.1426	0.0191	0.0654
Ours (on all)	Right foot	0.209	0.709	9.654	0.0645	0.1152	0.0153	0.0592
Omnicontrol	Head	0.316	0.687	9.921	0.0414	0.0928	0.0126	0.0757
Ours (on all)	Head	0.205	0.701	9.581	0.0605	0.0713	0.0100	0.0708
Omnicontrol	Left wrist	0.264	0.69	9.661	0.0527	0.2002	0.0284	0.1187
Ours (on all)	Left wrist	0.185	0.696	9.553	0.0586	0.1289	0.0148	0.0947
Omnicontrol	Right wrist	0.277	0.693	9.775	0.0513	0.1953	0.0251	0.1141
Ours (on all)	Right wrist	0.169	0.698	9.795	0.0601	0.1074	0.0146	0.0931
Omnicontrol	Average	0.297	0.691	9.828	0.0490	0.1387	0.0192	0.0833
Ours (on all)	Average	0.209	0.702	9.612	0.0610	0.0960	0.0128	0.0720

Table 2: Ablation study for different sampling strategies.

Method	**Joint**	**FID**	**↓R-precision**	**↑ (Top-3)**	**Diversity**	**→ Foot skating ratio**	**↓ Traj. err. (↓50 cm)**	**↓ Loc. err. (↓50 cm)**	**↓ Avg. err. ↓**
Base	Average	0.251	0.691		9.393	0.0570	0.1078	0.0148	0.0738
Guassian-3	Average	0.209	0.702		9.612	0.0610	0.0960	0.0128	0.0720
Guassian-5	Average	0.312	0.700		9.794	0.0558	0.1099	0.0134	0.0709
Peak	Average	0.253	0.703		9.424	0.0522	0.1210	0.0171	0.0787
Temporal	Average	0.248	0.706		9.891	0.0546	0.1135	0.0132	0.0752

Table 3: Ablation study for different dynamic features.

Method	Joint	FID	↓R-precision	↑ (Top-3)	Diversity	→ Foot skating ratio	↓ Traj. err. (↓50 cm)	↓ Loc. err. (↓50 cm)	Avg. err. ↓
Ours (on all)	Average	0.253	0.703		9.424	0.0522	0.1210	0.0171	0.0787
only-pos	Average	0.313	0.694		9.400	0.0572	0.1133	0.0147	0.0748
only-vel	Average	0.254	0.706		9.478	0.0689	0.1146	0.0156	0.0759
only-acc	Average	0.280	0.698		9.644	0.0516	0.1149	0.0162	0.0765

some performance in terms of trajectory accuracy, but the lack of velocity and acceleration supplementation makes the generated results lack naturalness and detail.

When using pure velocity features (third row), the model obtains the best R-precision; however, the Foot skating ratio also rises significantly, exposing the problem of transient motion instability. This suggests that velocity features are effective in enhancing motion coherence, but over-reliance on velocity features may lead to a lack of physical plausibility and unnatural sliding, especially at the joints.

For the case of using pure acceleration features (fourth row), despite the improvement of Foot skating, which reduces the sliding phenomenon in the movement, the overall performance is still not optimal and is relatively weak in other metrics. The single acceleration feature fails to adequately address the balance between trajectory accuracy and motion details, and thus has limitations in terms of generation quality.

4.4.3 Effect of different combinations of weights: In Section 4.4.2, we demonstrated the limitations of relying solely on a single kinematic feature for motion generation. To further explore how integrating position, velocity, and acceleration features can enhance generation quality, we designed models incorporating multiple feature fusions and experimentally evaluated the impact of different weight combinations on generated results. In our experiments, we independently adjusted the weights for position features (ω_{pos}), velocity features (ω_{vel}), and acceleration features (ω_{acc}), comparing their contributions to the generated results.

Table 4 illustrates the impact of different feature fusion strategies on motion generation quality. Experimental results indicate that the full feature fusion model (i.e., balanced weights for position, velocity, and acceleration) achieves optimal outcomes in both naturalness and detail representation. This demonstrates that the effective integration of all three features compensates for the limitations of individual features, ensuring trajectory accuracy while enhancing action detail and naturalness.

We further analyzed the impact of different weight combinations. Under combinations with larger ω_{pos}, although trajectory accuracy was outstanding, the naturalness of the motion decreased, manifested as higher Foot skating values. This indicates that position features significantly impact trajectory precision, but over-reliance on them leads to unnatural sliding and reduced fluidity. Conversely, combinations with higher ω_{vel} values showed reduced Foot skating values and less sliding, though trajectory accuracy was slightly compromised in some cases.

When adjusting acceleration feature weights, increasing ω_{acc} reduces foot skating and enhances physical plausibility. However, overemphasizing acceleration degrades overall performance, particularly in fine-grained motion details compared to balanced combinations.

In summary, weighted feature fusion plays a crucial role in motion generation. Appropriate weight combinations achieve a favourable balance between trajectory precision and naturalness, thereby enhancing the quality of generated actions.

Further optimization of experimental results indicates that when the weights of position, velocity, and acceleration features are relatively balanced, the generated motion maintains high-quality trajectory accuracy while possessing sufficient physical plausibility and naturalness. Therefore, the comprehensive utilization of these three features significantly enhances the effectiveness of action generation.

5. Conclusion

To address the challenge of balancing trajectory accuracy and motion naturalness, we propose the Trajectory-Aware Motion Generator (TAMG). By leveraging third-order dynamic features and an innovative multimodal masking strategy, our approach ensures both the naturalness and smoothness of the generated motion, while maintaining high-precision trajectory control. We validate the effectiveness of the model through a comprehensive series of experiments, demonstrating that TAMG significantly outperforms existing methods in both trajectory accuracy and overall motion quality. Our results highlight the importance of integrating

Table 4: Ablation study for different combinations of weights.

ω_{pos}	ω_{vel}	ω_{acc}	Joint	FID ↓	R-precision ↑(Top-3)	Diversity →	Foot-skating ratio ↓	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err. ↓
0.5	0.3	0.2	Average	0.253	0.703	9.424	0.0522	0.1210	0.0171	0.0787
0.7	0.3	0		0.266	0.699	9.752	0.0535	0.1138	0.0156	0.0750
0.34	0.33	0.33		0.245	0.696	9.711	0.0429	0.1084	0.0140	0.0729
0.5	0.3	0.2	Average	0.209	0.702	9.612	0.0610	0.0960	0.0128	0.0720
0.3	0.5	0.2		0.251	0.694	9.447	0.0561	0.1078	0.0144	0.0737

multiple biomechanical features and offer new insights for future research in trajectory-guided motion synthesis.

6. Acknowledgements

We would like to acknowledge all students and teachers of the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of the Ministry of Education of Minzu University of China, for their support and contributions to this research.

Author contributions

Xuan Liu was responsible for conceptualization, methodology, software development, validation, original draft preparation, reviewing and editing, formal analysis, and resources; Shaojun Yuan contributed to conceptualization, methodology, software development, validation, original draft preparation, reviewing and editing, and formal analysis; Zhiyang Zhang handled methodology, reviewing and editing, formal analysis, and supervision; Xiangyu Qu managed software development, validation, resources, and supervision; Yidian Liu worked on methodology, software development, validation, and supervision; Chaomurilige contributed to validation, resources, and supervision; Zheng Liu participated in reviewing and editing, resources, and supervision; Shan Jiang was involved in software development, validation, and resources.

Declarations

Data availability: The dataset used in this manuscript is a publicly available dataset. Detailed information about this dataset is provided in Section 4.1 of this manuscript.

Funding: This work was supported in part by the Hainan Li'an Education Technology Innovation Alliance (Natural Science Foundation NO.624LALH004) and the Grant of Open Project of Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Ministry of Education (NO. OPC-202401).

References

1. Wang CY, Zhou Q, Fitzmaurice G, Anderson F. Videoposevr: authoring virtual reality character animations with online videos. Proc ACM Hum Comput Interact. 2022;6(ISS). Available from: https://doi.org/10.1145/3567728
2. Ye H, Kwan KC, Su W, Fu H. Aranimator: in-situ character animation in mobile AR with user-defined motion gestures. ACM Trans Graph. 2020;39(4). Available from: https://doi.org/10.1145/3386569.3392404
3. Hu L, Zhang B, Zhang P, Qi J, Cao J, Gao D, Zhao H, Feng X, Wang Q, Zhuo L, Pan P, Xu Y. A virtual character generation and animation system for e-commerce live streaming. In: Proceedings of the 29th ACM International Conference on Multimedia. MM'21. New York: Association for Computing Machinery; 2021;1202–1211. Available from: https://doi.org/10.1145/3474085.3481547
4. Thomas S, Ferstl Y, McDonnell R, Ennis C. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). 2022;11–20. Available from: https://doi.org/10.1109/VR51125.2022.00018
5. Qi J, Jiang G, Li G, Sun Y, Tao B. Intelligent human-computer interaction based on surface EMG gesture recognition. IEEE Access. 2019;7:61378–61387. Available from: https://doi.org/10.1109/ACCESS.2019.2914728
6. Wang X, Yan K. Immersive human-computer interactive virtual environment using large-scale display system. Future Gener Comput Syst. 2019;96:649–659. Available from: https://doi.org/10.1016/j.future.2017.07.058
7. Zhou H, Wang D, Yu Y, Zhang Z. Research progress of human-computer interaction technology based on gesture recognition. Electronics. 2023;12(13). Available from: https://doi.org/10.3390/electronics12132805
8. Xu P. A real-time hand gesture recognition and human-computer interaction system. arXiv e-prints. 2017;1704.07296. Available from: https://doi.org/10.48550/arXiv.1704.07296
9. Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH. Human motion diffusion model. arXiv e-prints. 2022;2209.14916. Available from: https://doi.org/10.48550/arXiv.2209.14916
10. Xie Y, Jampani V, Zhong L, Sun D, Jiang H. OmniControl: control any joint at any time for human motion generation. arXiv e-prints. 2023;2310.08580. Available from: https://doi.org/10.48550/arXiv.2310.08580
11. Ahn H, Ha T, Choi Y, Yoo H, Oh S. Text2Action: generative adversarial synthesis from language to action. arXiv e-prints. 2017;1710.05298. Available from: https://doi.org/10.48550/arXiv.1710.05298
12. Messina N, Sedmidubsky J, Falchi F, Rebok T. Text-to-motion retrieval: towards joint understanding of human motion data and natural language. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23. New York: Association for Computing Machinery; 2023;2420–2425. Available from: https://doi.org/10.1145/3539618.3592069
13. Ahuja C, Morency LP. Language2Pose: natural language grounded pose forecasting. arXiv e-prints. 2019;1907.01108. Available from: https://doi.org/10.48550/arXiv.1907.01108
14. Zhang J, Zhang Y, Cun X, Zhang Y, Zhao H, Lu H, Shen X, Shan Y. Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;14730–14740. Available from: https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_Generating_Human_Motion_From_Textual_Descriptions_With_Discrete_Representations_CVPR_2023_paper.pdf
15. Athanasiou N, Petrovich M, Black MJ, Varol G. TEACH: temporal action composition for 3D humans. arXiv e-prints. 2022;2209.04066. Available from: https://doi.org/10.48550/arXiv.2209.04066
16. Poole B, Jain A, Barron JT, Mildenhall B. DreamFusion: text-to-3D using 2D diffusion. arXiv e-prints. 2022;2209.14988. Available from: https://doi.org/10.48550/arXiv.2209.14988
17. Xu L, Song Z, Wang D, Su J, Fang Z, Ding C, Gan W, Yan Y, Jin X, Yang X. Actformer: a GAN-based transformer towards general action-conditioned 3D human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;2228–2238. Available from: https://doi.org/10.48550/arXiv.2203.07706
18. Ma F, Xia G, Liu Q. Spatial consistency constrained GAN for human motion transfer. IEEE Trans Circuits Syst Video Technol. 2021;32(2):730–742.
19. Kundu JN, Gor M, Babu RV. Bihmp-gan: bidirectional 3D human motion prediction GAN. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:8553–8560. Available from: https://doi.org/10.1609/aaai.v33i01.33018553
20. Petrovich M, Black MJ, Varol G. Action-conditioned 3D human motion synthesis with transformer VAE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021;10985–10995. Available from: https://openaccess.thecvf.com/content/ICCV2021/papers/Petrovich_Action-Conditioned_3D_Human_Motion_Synthesis_With_Transformer_VAE_ICCV_2021_paper.pdf

21. Bie X, Guo W, Leglaive S, Girin L, Moreno-Noguer F, Alameda-Pineda X. Hit-dvae: human motion generation via hierarchical transformer dynamical VAE. arXiv preprint. 2022;2204.01565. Available from: https://doi.org/10.48550/arXiv.2204.01565

22. Kim H, Kong K, Kim JK, Lee J, Cha G, Jang HD, Wee D, Kang SJ. Enhanced control of human motion generation using action-conditioned transformer VAE with low-rank factorization. IEIE Trans Smart Process Comput. 2024;13(6):609–621. Available from: https://ieiespc.org/ieiespc/XmlViewer/f434509

23. Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L, Liu Z. MotionDiffuse: text-driven human motion generation with diffusion model. arXiv e-prints. 2022;2208.15001. Available from: https://doi.org/10.48550/arXiv.2208.15001

24. Wan W, Dou Z, Komura T, Wang W, Jayaraman D, Liu L. TLControl: trajectory and language control for human motion synthesis. arXiv e-prints. 2023;2311.17135. Available from: https://doi.org/10.48550/arXiv.2311.17135

25. Karunratanakul K, Preechakul K, Suwajanakorn S, Tang S. Guided motion diffusion for controllable human motion synthesis. arXiv e-prints. 2023;2305.12577. Available from: https://doi.org/10.48550/arXiv.2305.12577

26. Dai W, Chen LH, Wang J, Liu J, Dai B, Tang Y. MotionLCM: real-time controllable motion generation via latent consistency model. arXiv e-prints. 2024;2404.19759. Available from: https://doi.org/10.48550/arXiv.2305.12577

27. Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. 2023. Available from: https://arxiv.org/abs/2302.05543

28. Dabral R, Hamza Mughal M, Golyanik V, Theobalt C. MoFusion: a framework for denoising-diffusion-based motion synthesis. arXiv e-prints. 2022;2212.04495. Available from: https://doi.org/10.48550/arXiv.2212.04495

29. Ma J, Bai S, Zhou C. Pretrained diffusion models for unified human motion synthesis. arXiv e-prints. 2022;2212.02837. Available from: https://doi.org/10.48550/arXiv.2212.02837

30. Zhao M, Liu M, Ren B, Dai S, Sebe N. Modiff: action-conditioned 3D motion generation with denoising diffusion probabilistic models. arXiv e-prints. 2023;2301.03949. Available from: https://doi.org/10.48550/arXiv.2301.03949

31. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. arXiv e-prints. 2021;2112.10741. Available from: https://doi.org/10.48550/arXiv.2112.10741

32. Popov V, Vovk I, Gogoryan V, Sadekova T, Kudinov M. Grad-TTS: a diffusion probabilistic model for text-to-speech. arXiv e-prints. 2021;2105.06337. Available from: https://doi.org/10.48550/arXiv.2105.06337

33. Xu J, Wang X, Cheng W, Cao YP, Shan Y, Qie X, Gao S. Dream3D: zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models. arXiv e-prints. 2022;2212.14704. Available from: https://doi.org/10.48550/arXiv.2212.14704

34. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022;15979–15988. Available from: https://doi.org/10.1109/CVPR52688.2022.01553

35. Kundu JN, Seth S, Jampani V, Rakesh M, Babu RV, Chakraborty A. Self-supervised 3D human pose estimation via part guided novel image synthesis. 2020. Available from: https://arxiv.org/abs/2004.04400

36. Pinyoanuntapong E, Usama Saleem M, Karunratanakul K, Wang P, Xue H, Chen C, Guo C, Cao J, Ren J, Tulyakov S. ControlMM: controllable masked motion generation. arXiv e-prints. 2024;2410.10780. Available from: https://doi.org/10.48550/arXiv.2410.10780

37. Three-Dimensional Kinematics and Kinetics. In: John Wiley & Sons, Ltd. 2009;176–199. Chapter 7. Available from: https://doi.org/10.1002/9780470549148.ch7

38. Yuan Y, Song J, Iqbal U, Vahdat A, Kautz J. PhysDiff: physics-guided human motion diffusion model. 2023. Available from: https://arxiv.org/abs/2212.02500

39. Zhou H, Guo C, Zhang H, Wang Y. Learning multiscale correlations for human motion prediction. 2021. Available from: https://arxiv.org/abs/2103.10674

40. Shao D, Shi M, Xu S, Chen H, Huang Y, Wang B. FinePhys: fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. 2025. Available from: https://arxiv.org/abs/2505.13437

41. Rempe D, Luo Z, Bin Peng X, Yuan Y, Kitani K, Kreis K, Fidler S, Litany O. Trace and pace: controllable pedestrian animation via guided trajectory diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023;13756–13766. Available from: https://doi.org/10.48550/arXiv.2304.01893

42. 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 (ICCV). 2023; 3836–3847. Available from: https://doi.org/10.48550/arXiv.2302.05543

43. Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, Cheng L. Generating diverse and natural 3D human motions from text. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022;5142–5151. Available from: https://ieeexplore.ieee.org/document/9880214

44. Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ. AMASS: archive of motion capture as surface shapes. arXiv e-prints. 2019;1904.03278. Available from: https://doi.org/10.48550/arXiv.1904.03278

45. Guo C, Zuo X, Wang S, Zou S, Sun Q, Deng A, Gong M, Cheng L. Action2Motion: conditioned generation of 3D human motions. arXiv e-prints. 2020;2007.15240. Available from: https://doi.org/10.1145/3394171.3413635