HSPI The Open Access Publisher

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

Research Article

Dynamic Masking Strategy: An Effective Approach to Enhancing Accurate Human Motion Generation

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

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

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

[3]Information Security Center, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China

[#]These authors contributed equally.

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

Submitted: November 04, 2025
Approved: November 14, 2025
Published: November 17, 2025

How to cite this article: Liu X, Zhang Z, Qu X, Yuan S, Liu Y, Chaomurilige, et al. Dynamic Masking Strategy: An Effective Approach to Enhancing Accurate Human Motion Generation. J Artif Intell Res Innov. 2025; 1(1): 075-084. Available from: https://dx.doi.org/10.29328/journal.jairi.1001009

Keywords: Human motion generation; Dynamic masking strategy; Motion amplitude; Temporal features

Abstract

Human motion generation has become an important research direction in computer vision and human motion modelling. Current motion generation methods typically rely on static or random masking during training, which fail to adequately capture dynamic variations in joint movement amplitude and temporal characteristics, resulting in suboptimal accuracy in generated motions. To address this, we propose a dynamic masking strategy (DMS) based on motion amplitude, which dynamically adjusts the mask probability distribution by incorporating both motion amplitude and temporal features. By calculating the motion amplitude of each joint and adapting the mask timing, the model is directed to focus on key movements during training, enhancing the quality of motion generation. Experimental results demonstrate that DMS outperforms traditional methods across multiple evaluation metrics, achieving a 15.3% reduction in FID, a 9.0% reduction in trajectory error, and a 6.3% reduction in location error, thereby validating the effectiveness and sophistication of the proposed method.

Introduction

Motion generation, particularly interactive motion generation, has become a hot and highly valuable research direction in the fields of computer vision and human motion modelling. With the continued progress of deep learning, motion generation models have demonstrated substantial potential across various fields [1]. The core challenge of these models is to generate motions that align with the motion trajectories provided by the user, in order to accurately produce behaviors that align with the user's intentions.

During training, masking techniques are commonly employed to enhance data diversity and prevent overfitting. In some motion modelling tasks, masking is used to obscure joint movements, thereby enhancing the model's robustness. However, most existing masking strategies rely on fixed patterns or simple random selection [2], which fail to fully leverage temporal dynamics and the differences in joint motion amplitude. This leads to models that cannot efficiently capture key motions and timing-dependent dynamics, as the masking strategies do not adapt to the unique motion properties of different joints and time periods. Thus, designing a strategy that dynamically adjusts the mask based on motion amplitude and timing characteristics has become crucial for improving motion model training and achieving more precise behavioural representation.

To address these challenges, we propose a dynamic masking strategy based on motion amplitude. This method dynamically adjusts the mask by combining joint motion amplitude and time series motion features. First, it helps the

model focus on critical motions by calculating the motion amplitude of each joint and assigning a masking probability based on this value, with small-amplitude joints being more likely to be masked. Second, the adaptive selection strategy dynamically adjusts the masking timing, aligning it with periods of mild motion in the time dimension. This ensures that the model applies more masking during these periods. Finally, our strategy prevents overmasking during periods of intense motion or highly relevant time intervals, enabling the model to learn key motion information more effectively, thereby improving the efficiency and effectiveness of model training.

We conducted a series of experiments to validate the proposed dynamic masking strategy based on motion amplitude. The results show that this strategy enhances the accuracy of model-generated motions, reduces trajectory errors, and addresses the shortcomings of traditional masking methods in adapting to dynamic motion characteristics. Our approach provides novel insights and a design framework for motion generation tasks. The main contributions of this paper are as follows:

- In-depth analysis of the impact of joint motion amplitude and time series characteristics on motion generation control.
- Proposal of a dynamic masking strategy based on motion amplitude, which overcomes the limitations of traditional methods and enhances the accuracy and diversity of generated motions.
- The effectiveness of this strategy was validated on a key dataset, resulting in significant reductions in FID values and error metrics. This provides new insights for future research on human motion generation.

Related work

Human motion generation

Human motion has become a focal point of academic research, particularly in the area of conditional motion generation, which has seen significant advancements. Motion generation typically employs multimodal inputs, encompassing various control modalities such as text [3-12], music [13-17], objects [18-25], 3D scenes [26-28], images [29], movements [30,31], and trajectories [32-39].

In the domain of human motion generation, researchers have employed diverse methods to enhance the accuracy and generalization of models. Traditional motion generation techniques largely rely on deep neural networks, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). With the continued advancement of deep learning, methods based on Generative Adversarial Networks (GANs) [40] have also gained popularity for generating high-quality motion trajectories. These approaches are effective in generating motion sequences that exhibit a sense of realism, achieved through adversarial training between the generative and discriminative networks.

In recent years, several novel approaches have emerged to further enhance the performance of motion generation models by incorporating interpretability and adaptability. For instance, attention mechanism-based methods [41,42] are widely utilized in sequence modeling tasks, particularly for generating motion trajectories. These methods significantly improve the quality and accuracy of generated motions by automatically focusing on key frames or essential parts of the action [43]. Additionally, reinforcement learning-based approaches [44] have been applied to motion generation, especially in dynamic environments, by adaptively adjusting the generation strategy, thereby enhancing the flexibility and adaptability of the models.

Dynamic masking strategies

Dynamic masking strategies have been successfully applied across various domains, demonstrating their ability to enhance model performance and efficiency, particularly when handling complex, time-sensitive data. In fields such as natural language processing, computer vision, and machine learning, dynamic masking strategies improve model robustness, generalization, and training efficiency by adaptively adjusting mask positions. In human movement prediction, dynamic masking strategies are also of significant value, helping models focus on large movement changes.

In Natural Language Processing (NLP), dynamic masking enhances the model's contextual understanding and generalization ability by randomly masking different parts of the input. RoBERTa [45] optimized BERT by employing dynamic masking, allowing different words to be randomly masked in each training cycle, thereby helping the model learn richer contextual information. Similarly, SpanBERT [46] improves the effectiveness of question-answering and inference tasks by masking consecutive word spans rather than individual words, enabling the model to capture long-distance dependencies.

In computer vision, dynamic masking strategies are widely used in tasks such as image restoration, object detection, and instance segmentation. In image restoration, dynamic masking adaptively selects regions to be masked based on the image content, enabling the model to focus more on restoring important areas. In object detection and instance segmentation, DynaMask [47] introduces a dynamic mask resolution selection module, which allows flexible adjustment of the masking strategy, thus balancing computational efficiency and segmentation accuracy.

For human motion prediction, the introduction of dynamic masking strategies helps models handle complex motion sequences and temporal data more effectively, thereby

improving the quality of generated motions. Traditional motion generation models typically rely on fixed or random masking strategies, which may not adequately address situations with significant movement amplitude variations. HumanMAC [48], proposed by Chen, et al. predicts future motions through a masked motion complementation framework that combines observed motion data and generative models, enhancing continuity and diversity.

However, these methods often overlook the motion amplitude and timing characteristics of joints. To address this, we propose a dynamic masking strategy based on motion amplitude, which adaptively adjusts the masking probability according to the motion amplitude and timing characteristics of each joint. This approach not only improves the accuracy and diversity of generated results but also addresses the limitations of traditional methods that struggle to adapt to large motion variations.

Methods

Human motion generation presents several challenges, with the primary one being the efficient generation of natural and realistic motion trajectories. A key difficulty lies in accurately generating the complex dynamic properties of human joints, where the motion amplitude and its variation across different joints in a time series must be properly handled. Traditional motion generation models often overlook these variations, leading to issues such as unnatural or distorted movements, especially in areas where joint motion amplitudes are either very small or exceptionally large.

To address this challenge, we propose an advanced training methodology that dynamically adjusts the masking probability based on motion amplitude. This approach allows the model to better capture the subtle differences in motion across joints, resulting in more realistic and accurate motion generation. Specifically, the method assigns different masking probabilities to regions with varying motion amplitudes, enabling the model to focus on critical motion differences between joints and improve the quality of generated motions.

The key idea behind this method is the dynamic masking strategy, which adjusts joint information during the training process. By selectively masking certain parts of the motion data based on amplitude, the model is forced to focus more on the important motion dynamics. This process not only mitigates the influence of irrelevant or low-dynamic regions but also enhances the model's ability to generate motion sequences that align with the underlying behavioural patterns.

The generation process using our proposed method is illustrated in Figure 1. The steps involved are as follows: First, human motion data and their corresponding textual descriptions are extracted from a dataset. The motion sequences and the text encodings are then generated. These motion sequences are subsequently input into the masking strategy module, where the motion amplitudes are computed, and masking probabilities are dynamically assigned to different regions. The dynamic masking process is then applied to adjust the motion information, ensuring that critical motion details are preserved. Finally, the adjusted motion sequences, along with the associated text encodings, are passed into a diffusion model to generate the corresponding human motions. This dynamic approach significantly improves the quality and naturalness of the generated motions while capturing the nuanced variations in joint movements. This comprehensive method facilitates the creation of high-quality, realistic motions by dynamically adjusting for motion amplitude differences, thereby offering a more accurate and natural representation of human movement in motion generation tasks.

Joint masking strategy based on motion amplitude

To improve the training efficiency and effectiveness of motion generation models, this study introduces a dynamic joint masking strategy based on motion amplitude. This strategy aims to refine the model's learning process by dynamically adjusting the masking probabilities according to the motion amplitude of each joint. The core objective is to enable the model to focus on the most critical movement patterns, which are often characterised by larger motion amplitudes, ensuring a more efficient learning process.

The masking strategy is designed to be adaptive, with the motion amplitude of each joint playing a crucial role in determining its masking probability. Joints with larger motion amplitudes, which typically correspond to more significant or expressive movements, are assigned lower masking probabilities, whereas joints with smaller motion amplitudes, often indicative of less critical or background motions, are given higher masking probabilities. This dynamic adjustment helps the model concentrate on the key motion details that are most relevant for accurately generating emotionally expressive movements.

The rationale behind this approach is to leverage the natural variations in joint motion amplitudes to prioritise learning the more expressive, dynamic, and important motions. By doing so, the model can focus its attention on these essential movements during training, leading to improved accuracy and realism in the generated motions. This comprehensive method facilitates the creation of high-quality, behaviorally consistent motions by dynamically adjusting for motion amplitude differences.

Motion amplitude calculation: The motion amplitude for each joint is calculated by evaluating the displacement differences between consecutive frames over the course of the time series. Specifically, the model calculates the difference in the position of the nth joint between two consecutive time steps $J_t + 1, n$, using the Euclidean distance. This displacement difference is then averaged across all the frames in the time

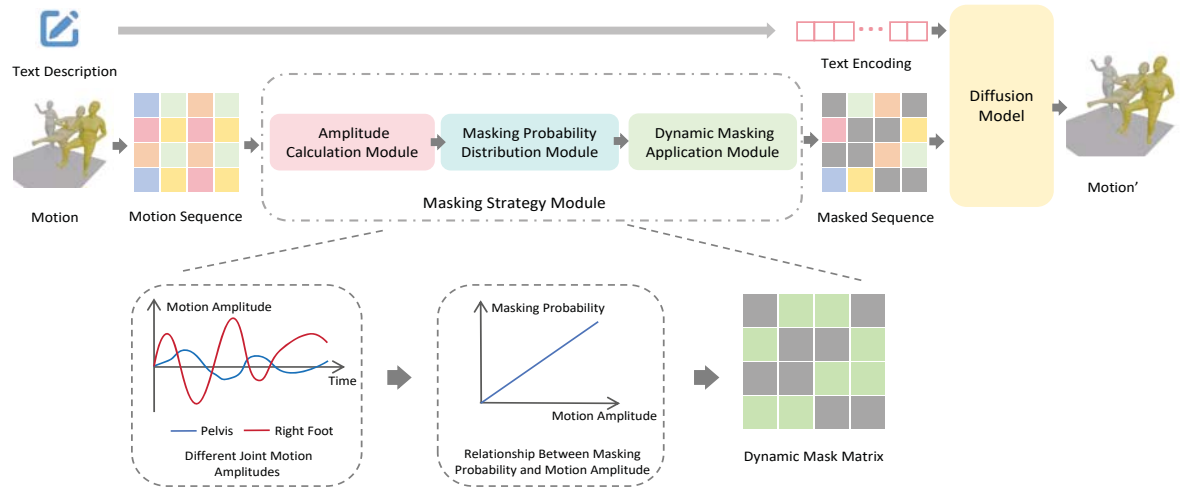

Figure 1: Flowchart of motion generation and dynamic masking strategy.

sequence to yield a single value representing the joint's overall motion amplitude, M_n.

The formula is expressed as follows:

$$M_n = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\| J_{t+1,n} - J_{t,n} \right\|_2 \quad (1)$$

where M_n denotes the motion amplitude of the nth joint, T is the length of the time sequence, and $J\{t,n\}$ denotes the position of the n joint in the t frame.

Here:

- M_n is the motion amplitude of the nth joint.
- T is the total number of frames in the time sequence.
- J_t is the position of the n joint at the t frame.
- The$||\cdot||$ denotes the Euclidean norm, which measures the straight-line distance between the positions of the joint at consecutive time steps.

By calculating this motion amplitude for each joint, the model quantifies how much each joint is moving over time. Joints with larger motion amplitudes, indicating they are more dynamic or play a more significant role in the overall motion, are assigned a higher probability of being retained. This approach ensures that the model prioritises masking those joints with greater influence on the motion, promoting better generalisation during training.

Mask probability assignment and joint selection: In the mask probability assignment and joint selection process, the motion amplitude of each joint is used to determine the probability that the joint will be masked during training. Specifically, the model normalises the motion amplitudes of all controllable joints to ensure that joints with small motion amplitudes are more likely to be selected for masking, and joints with larger motion amplitudes are more likely to be retained. This normalisation process helps to prioritise the more dynamic joints, which are crucial for the motion, while still allowing for diversity in the selection of joints to be masked.

The formula for calculating the mask probability is:

$$P_n^{joint} = \frac{M_n}{\sum_{k\in C} M_k}, \forall n \in C \quad (2)$$

Where C denotes the set of controllable joints and P_n^{joint} denotes the probability that the n joint is retained in the mask sequence. Where:

- P_n^{joint} represents the probability that the n joint is selected to be retained in the mask sequence.
- M_n is the motion amplitude of the n joint.
- The denominator is the sum of the motion amplitudes of all controllable joints in the set C, which normalises the individual probabilities.

At each training iteration, the model dynamically determines how many joints to mask, selecting a random number between 1 and 3. This variability in the number of masked joints during training enhances the model's ability to generalise by preventing overfitting to specific joints. By using this approach, the model ensures a more diverse selection of joints for masking, which ultimately contributes to better performance across different scenarios and conditions.

Dynamic masking strategy based on time dimension

In addition to the masking strategy applied to the joint dimension, this study introduces a dynamic masking approach specifically designed for the time dimension. This temporal masking strategy recognizes that not all frames within a motion sequence contribute equally to the overall motion information. Certain moments within the sequence

may carry more significant motion data, which is crucial for training a robust model. Therefore, we dynamically assign mask weights to each frame based on its motion amplitude, which is a measure of how much movement occurs within a particular time frame.

Frames with larger motion amplitudes, indicating that they contain more significant motion, are assigned lower masking probabilities. This allows the model to place more focus on these critical moments where the motion is more intense, which are likely to contain the most informative features necessary for learning complex motion patterns. Conversely, frames with smaller motion amplitudes, representing less significant motion, are assigned higher probabilities for masking.

By implementing this strategy, the model is encouraged to prioritise frames that are rich in motion data, ensuring that the most important aspects of the motion sequence are preserved during training. This dynamic masking approach not only enhances the model's ability to learn from the most relevant parts of the motion sequence but also helps improve its overall generalisation by focusing on the most temporally significant moments.

This method facilitates more targeted learning, where the model is exposed to the varying dynamics of the motion while being trained in a way that reduces the risk of overfitting to less informative frames. As a result, the model becomes better equipped to handle complex motion patterns and exhibit improved performance during inference.

Time weight calculation: In the time dimension, the motion of the selected joints in each frame is initially calculated to assess how much movement occurs between consecutive frames. This motion magnitude is determined by computing the positional change of each joint between the current frame t and the previous frame $t-1$. The displacement is measured using the Euclidean distance between the joint positions at these two time points. To account for the overall temporal continuity of the motion sequence and ensure consistency, the motion amplitude of each frame is then normalised across the entire sequence.

The normalisation process ensures that the motion amplitudes are scaled in a way that reflects their relative significance across the entire sequence, allowing for a fair comparison between frames. This normalised value is used to calculate a temporal weight for each frame, denoted as wt, which quantifies the importance of each frame in the training process.

Frames with larger motion amplitudes, signifying more significant or dynamic movement, are assigned higher temporal weights. These frames are considered more important for training, as they are likely to contain crucial information about the motion. On the other hand, frames with smaller motion amplitudes, indicating minimal movement, are assigned lower weights. This differential weighting mechanism ensures that the model places more emphasis on the frames that contribute more significantly to the overall motion sequence.

The formula for calculating the time weight is:

$$w_t = \frac{\sum_{n \in S} \left\| J_{t,n} - J_{t-1,n} \right\|_2}{\sum_{T=1}^{T} \sum_{n \in S} \left\| J_{\tau,n} - J_{\tau-1,n} \right\|_2} \tag{3}$$

- S represents the set of selected joints.
- J_t denotes the position of the nth joint at time t.
- The Euclidean norm $||\cdot||$ measures the displacement of each joint between consecutive frames.
- T is the total number of frames in the sequence.

By using this temporal weight calculation, the model is guided to focus more on the frames that carry significant motion information, thus optimising the training process to capture the most critical dynamic features of the motion.

Construction and application of masking matrix: Once the time weights are calculated for each frame, the next step is to construct the masking matrix. The masking matrix is a binary matrix that indicates which frames in the motion sequence should be masked. Using the time weights as a basis, we randomly select a certain percentage of time frames to be masked during each training iteration. The proportion of frames selected for masking is determined according to their respective time weights: frames with higher motion amplitudes (i.e., more critical moments) will have higher probabilities of being selected for retaining in the mask sequence, ensuring that more significant moments are retained.

The construction process of the mask matrix involves several steps:

- First, initialize a matrix with all values set to false, indicating that initially, all joint mask states are set to False.
- **Selection of Frames to be retained:** Based on the time weights, a certain percentage of frames are selected for retaining. These frames correspond to moments with higher motion intensity, as indicated by their time weight.
- **Application of retain:** Apply the retained to the selected frames by setting the retained values of the corresponding joints to True at the designated time points. The retained positions for these frames are concentrated in regions where joint movement is significant and temporally important, ensuring the preservation of critical motion information.

By constructing the mask matrix in this way, the model is

guided to focus its learning efforts on the most important time moments of the motion. This ensures that sufficient semantic information about the motion is preserved, particularly for the high-intensity motion moments, while less important or static moments will receive less attention. As a result, the model learns to generate more realistic and dynamic human motion, with improved accuracy and coherence in its output.

Theoretical connection and cognitive motivation

The proposed Dynamic Masking Strategy (DMS) is not only an effective technique for improving model generalization but also aligns conceptually with human motion perception and attention mechanisms. Studies in cognitive science indicate that humans tend to focus attention on body parts or time segments with large motion amplitudes, significant velocity changes, or event boundaries, as these regions carry higher informational value for motion understanding and prediction. By computing per-joint motion amplitudes and frame-wise dynamic variations, DMS automatically identifies and prioritizes these highly informative spatiotemporal regions during training. This design can be regarded as a computational analogue of human attention allocation, where the model selectively emphasizes salient motion cues while suppressing redundant or less informative ones. Consequently, DMS enables the model to learn motion representations that are more consistent with human perceptual focus, enhancing both semantic coherence and naturalness of the generated motion. In this sense, the proposed strategy not only improves performance empirically but also provides a solid theoretical grounding inspired by human perceptual and predictive coding mechanisms.

Experiments

Dataset and evaluation methods

In this study, we utilise the HumanML3D dataset for the text-based motion generation task. The HumanML3D dataset [10] contains 14,616 unique motion capture data samples sourced from the AMASS [49] and HumanAct12 [10] datasets, along with 44,970 corresponding text descriptions. This dataset provides extensive human motion data, making it highly suitable for evaluating the effectiveness of text-based motion generation models.

We adopt the evaluation framework proposed by Guo, et al. [50] to comprehensively assess the generated motion sequences using multiple metrics. First, the Fréchet Inception Distance (FID) is used to evaluate the naturalness and fidelity of the generated motions by measuring the difference between the generated and true distributions in the feature space of a pre-trained model. A lower FID value indicates higher quality in the generated motions.

Next, R-Precision is employed to assess the relevance of the generated motion sequences to the corresponding text descriptions. Specifically, the relevance of a motion sequence is measured against 32 textual descriptions, one of which is the true description, and the other 31 are random. The accuracy is evaluated based on the first 3 matches by calculating the Euclidean distance between the motion embedding and the textual description.

Additionally, we use Diversity to measure the variation in the generated motions. Diversity is quantified by calculating the average cumulative difference between randomly paired generated sequences in the test set.

To evaluate control accuracy, we introduce several control performance metrics. These include trajectory error, location error, and mean error, which measure the 3D control accuracy of the controlled joint locations at keyframes. Trajectory error indicates the percentage of keyframe joint positions deviating from a set trajectory threshold. Location error measures the percentage of keyframe locations that fall outside a given distance constraint. Mean error calculates the average Euclidean distance between a joint's position and the corresponding control trajectory at each keyframe.

During evaluation, all models were trained to generate 196-frame motion sequences at five keyframe density levels: 1, 2, 5, 49 (25% density), and 196 (100% density). The time step of the keyframes was randomly sampled, and all models were provided with real trajectories as spatial control signals during both training and evaluation. We report the average performance across all density levels. For training, we followed the warm-up procedure outlined in the OmniControl article.

Comparison with other methods

We conducted a comprehensive comparison between our proposed method and the OmniControl method, using data from our replicated model. The results were analysed using multiple evaluation metrics, and the comparison outcomes are presented in the table. The metrics compared include FID (Fréchet Inception Distance), R-precision (Top-3), Diversity, Trajectory Error (Traj. Err.), Location Error (Loc. Err.), and Average Error (Avg. Err.). For each metric, the performance is shown separately based on different control joints.

As shown in Table 1, our method outperforms OmniControl on most key metrics. Specifically, on the FID metric, our method achieves a clear advantage with an average value of 0.180, compared to 0.297 for OmniControl, indicating significantly lower error in the quality of the generated motions. Additionally, our method demonstrates more stable performance in terms of Diversity, with a mean value of 9.529, compared to 9.828 for OmniControl, and shows superior performance across multiple control joints, suggesting that our model maintains a higher level of diversity throughout the generation process.

In terms of error-related metrics (Trajectory Error, Location Error, and Average Error), our method generally

Table 1: Comparison of quantitative results of our method with the current OmniControl method on the HumanML3D test set. Best scores are highlighted in bold.

Method	Control	FID	R-precision	Diversity	Traj. Err.	Loc. Err.	Avg. Err.
Real	-	0.002	0.797	9.503	0.000	0.000	0.000
OmniControl	Pelvis	0.323	0.691	9.854	0.078	0.013	0.064
Ours(On all)	Pelvis	0.160	0.662	9.376	0.075	0.014	0.057
OmniControl	Head	0.316	0.687	9.921	0.093	0.017	0.076
Ours(On all)	Head	0.222	0.662	9.565	0.086	0.015	0.056
OmniControl	Left Hand	0.264	0.690	9.661	0.200	0.028	0.119
Ours(On all)	Left Hand	0.199	0.678	9.432	0.105	0.021	0.060
OmniControl	Right Hand	0.264	0.690	9.661	0.200	0.028	0.119
Ours(On all)	Right Hand	0.204	0.685	9.399	0.067	0.012	0.054
OmniControl	Left Foot	0.292	0.689	9.855	0.123	0.017	0.062
Ours(On all)	Left Foot	0.136	0.652	9.741	0.099	0.015	0.070
OmniControl	Right Foot	0.307	0.693	9.901	0.143	0.019	0.065
Ours(On all)	Right Foot	0.161	0.663	9.658	0.101	0.016	0.070
OmniControl	Average	0.297	0.691	9.828	0.139	0.019	0.083
Ours(On all)	Average	0.180	0.667	9.529	0.089	0.015	0.061

performs better. For instance, the mean trajectory error is 0.089 compared to 0.139 for OmniControl; the mean location error is 0.015, notably lower than 0.019 for OmniControl; and the mean error is 0.061, compared to 0.083 for OmniControl, indicating that our method generates more accurate results.

In summary, our method demonstrates significant advantages in FID, diversity, and error-related metrics, particularly in terms of generation quality and error control. This suggests that our proposed model not only maintains diversity and stability but also produces more accurate, high-quality results, thus more effectively simulating and representing human motion.

Ablation experiments

Removal of motion magnitude-based joint masks: To further validate the effectiveness of our proposed dynamic masking strategy based on motion amplitude, we conducted an ablation experiment by removing the motion amplitude-based joint mask. In this experiment, the dynamic masks based on motion amplitude were removed, and only the standard global masks were used for comparative analysis.

The results of the ablation experiment are shown in Table 2. From the table, it is evident that removing the motion amplitude-based joint mask leads to a decrease in the model's overall performance, as reflected in the trajectory error (Traj. Err.), location error (Loc. Err.), and average error (Avg. Err.).

In terms of trajectory error, our method achieves a mean value of 0.089, while the error after removing the joint mask increases to 0.097, indicating an approximate 9.0% increase in error. For location error, the mean value of our method is 0.015, which is 6.3% lower than 0.016 after removing the mask, demonstrating the positive impact of the motion amplitude mask on accurate localization. Finally, the average error in our method is 0.061, while after removing the mask, it increases to 0.063, reflecting a 3.3% increase.

These initial results highlight the significant role of the motion amplitude-based joint mask in improving the accuracy and stability of the model. The removal of this mask results in degradation of the model's performance across several key metrics, particularly in trajectory error and location accuracy. This suggests that our dynamic masking strategy effectively enhances generation quality and improves the model's accuracy in the motion generation task.

This ablation Study further confirms the crucial role of motion amplitude-based joint masking in the motion generation task, providing strong evidence for the effectiveness of our proposed masking strategy.

Removal of time dimension masks: To further validate the effectiveness of our proposed dynamic masking strategy based on motion magnitude, we conducted an ablation experiment by removing the time dimension mask. This experiment compares the performance changes after removing the time dimension mask, which involves excluding the temporal magnitude information from the motion data.

Table 3 presents the experimental results. Our method outperforms the strategy without the time dimension mask across three key metrics: trajectory error (Traj. Err.), location error (Loc. Err.), and average error (Avg. Err.). With the removal of the time dimension mask, the trajectory error increases from 0.089 to 0.100, the location error rises from 0.015 to 0.016, and the average error increases from 0.061 to 0.063. These results indicate that the inclusion of the time

Table 2: Ablation Study to Remove the Effect of Joint Masking Based on Motion Amplitude.

Method	Joint	Traj. Err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err. ↓ (cm)
Ours	Average	0.089	0.015	0.061
Removing Joint Dimension Mask	Average	0.097	0.016	0.063

Table 3: Ablation study of the effect of removing time dimension masks.

Method	Joint	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err. ↓ (cm)
Ours	Average	0.089	0.015	0.061
Removing Time Dimension Mask	Average	0.100	0.016	0.063

dimension mask is effective in reducing errors during the motion generation process, particularly in enhancing the accuracy of the trajectory and location.

These findings clearly demonstrate the effectiveness of our dynamic masking strategy that adjusts the mask based on both motion amplitude and temporal characteristics. Removing the time dimension mask leads to performance degradation, as the model fails to adequately capture the temporal dynamics of the motion. Therefore, retaining the temporal dimension mask is crucial for enhancing the model's ability to understand and generate movements.

This ablation experiment further underscores the important role of the time dimension mask in improving model robustness and generation accuracy, providing strong evidence for the effectiveness of our proposed masking strategy.

Comparison of masking strategies: In this section, we further validate the effectiveness of our proposed strategy, which assigns higher masking probabilities to regions with smaller motion amplitudes, through two ablation experiments. The two experiments are: 1) higher masking probabilities in regions with large motion amplitudes, and 2) higher masking probabilities in regions with medium motion amplitudes. We compare the results of these experiments with our proposed method to analyse the impact of the masking strategy on model performance.

Table 4 presents the results of these experiments. As shown in the table, the model errors exhibit varying degrees of change. Compared to our method, the strategy of assigning higher masking probabilities to regions with larger motion amplitudes leads to an increase in trajectory error (Traj. Err.) from 0.089 to 0.113, which represents an increase of approximately 26.9%. Similarly, location error (Loc. Err.) increased from 0.015 to 0.018, a 20.0% increase.

For the strategy with higher masking probabilities in regions with moderate motion amplitudes, the trajectory error is 0.114, and the location error is 0.020, representing increases of about 28.1% and 33.3%, respectively, compared to our method. However, the average error (Avg. Err.) shows less variation across the three methods, with values of 0.061, 0.070, and 0.070, respectively.

These results demonstrate that when the masking probability is higher in regions with larger or medium motion amplitudes, the model's error increases. Compared to our proposed strategy, which assigns higher probabilities to regions with smaller motion amplitudes, the use of larger or medium motion amplitude regions for masking negatively impacts the generation accuracy. This further validates the effectiveness of our dynamic masking strategy based on motion amplitude in improving both the quality and accuracy of generation.

Table 4: Comparison of the results of the ablation experiments with higher masking probability in large and moderate motion amplitude regions, respectively.

Method	Joint	Trai. Err. ↓ (50 cm)	Loc. Err. ↓ (50 cm)	Avg. err. ↓ (50 cm)
Ours	Average	0.089	0.015	0.061
Higher masking in lager amplitude	Average	0.113	0.018	0.070
Higer masking in moderate amplitude	Average	0.114	0.020	0.070

Potential applications and future directions

The enhanced accuracy and stability achieved by the Dynamic Masking Strategy suggest its broad applicability across multiple domains:

Virtual Human Animation and Film Production – DMS can generate more natural and precise human motions, reducing manual correction efforts in digital character animation and motion completion tasks.

Immersive Interaction and Virtual/Augmented Reality (VR/AR) – By improving the realism of generated motions, DMS can enhance the responsiveness and immersion of virtual agents in real-time interactive environments.

Intelligent Motion Synthesis and Robotics – The attention-guided masking of key joints enables better coordination and adaptability in robotic motion planning and biomimetic control.

Action Understanding and Behavior Analysis – The masking and attention mechanism introduced in DMS can also benefit recognition and prediction tasks by encouraging models to focus on the most discriminative spatiotemporal features. Overall, DMS not only advances human motion generation but also offers a promising foundation for future research in multimodal interaction, digital humans, and intelligent robotic systems.

Conclusion

This paper proposes a dynamic masking strategy (DMS) that integrates both motion amplitude and the time dimension. The method introduces a motion amplitude computation module and a dynamic masking allocation mechanism, enhancing the model's ability to adaptively adjust joint motion amplitude and temporal characteristics in human motion generation. This approach effectively overcomes the limitations of traditional methods in generating complex dynamic motions.

The experimental results validate the effectiveness of the proposed method, demonstrating a significant reduction in generation error, improved motion accuracy, and enhanced model performance in handling dynamic motion features. This method provides a novel research direction for motion generation tasks, with practical significance in improving the accuracy and realism of motion generation.

Acknowledgments

The research was supported by the Hainan Provincial Joint Project of Li'an International Education Innovation pilot Zone (624 LALH 004).

References

1. Ahn H, Ha T, Choi Y, Yoo H, Oh S. Text2action: generative adversarial synthesis from language to action. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2018;5915–5920. Available from: https://rllab.snu.ac.kr/publications/papers/2018_icra_text2action.pdf
2. Ahuja C, Morency L-P. Language2pose: natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). IEEE; 2019;719–728. Available from: https://arxiv.org/abs/1907.01108
3. Athanasiou N, Petrovich M, Black MJ, Varol G. Teach: temporal action composition for 3D humans. In: 2022 International Conference on 3D Vision (3DV). IEEE; 2022;414–423. Available from: https://arxiv.org/abs/2209.04066
4. Chen L-H, Zhang J, Li Y, Pang Y, Xia X, Liu T. Humanmac: masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;9544–9555. Available from: https://arxiv.org/abs/2302.03665
5. Chen X, Su Z, Yang L, Cheng P, Xu L, Fu B, Yu G. Learning a variational motion prior for video-based motion capture. arXiv. 2022. arXiv:2210.15134. Available from: https://arxiv.org/abs/2210.15134
6. Dai W, Chen L-H, Wang J, Liu J, Dai B, Tang Y. Motionlcm: real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision. Springer; 2024;390–408. Available from: https://arxiv.org/abs/2404.19759
7. Gao H, Wan Y, Xu H, Chen L, Xiao J, Ran Q. Swinbtc: transfer learning to brain tumor classification for healthcare electronics using augmented MR images. IEEE Transactions on Consumer Electronics. 2025. Available from: https://ui.adsabs.harvard.edu/link_gateway/2025ITCE...71.2297G/doi:10.1109/TCE.2025.3527061
8. Ghosh A, Dabral R, Golyanik V, Theobalt C, Slusallek P. IMoS: intent-driven full-body motion synthesis for human-object interactions. In: Computer Graphics Forum. 2023;42:1–12. Available from: https://doi.org/10.1111/cgf.14739
9. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems. 2014;27. Available from: https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html
10. Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, Cheng L. Generating diverse and natural 3D human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022;5152–5161. Available from: https://openaccess.thecvf.com/content/CVPR2022/papers/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.pdf
11. Guo C, Zuo X, Wang S, Zou S, Sun Q, Deng A, Gong M, Cheng L. Action2motion: conditioned generation of 3D human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020;2021–2029. Available from: https://doi.org/10.1145/3394171.3413635
12. Hassan M, Ceylan D, Villegas R, Saito J, Yang J, Zhou Y, Black MJ. Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021;11374–11384. Available from: https://arxiv.org/abs/2108.08284
13. Huang S, Wang Z, Li P, Jia B, Liu T, Zhu Y, Liang W, Zhu S-C. Diffusion-based generation, optimization, and planning in 3D scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;16750–16761. Available from: https://arxiv.org/abs/2301.06015
14. Huang Y, Wan W, Yang Y, Callison-Burch C, Yatskar M, Liu L. Como: controllable motion generation through language-guided pose code editing. In: European Conference on Computer Vision. Springer; 2024;180–196. Available from: https://arxiv.org/abs/2403.13900
15. Jiang N, Liu T, Cao Z, Cui J, Zhang Z, Chen Y, Wang H, Zhu Y, Huang S. Full-body articulated human-object interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;9365–9376. Available from: https://openaccess.thecvf.com/content/ICCV2023/html/Jiang_Full-Body_Articulated_Human-Object_Interaction_ICCV_2023_paper.html
16. Jiang X, Chu M, Wang X, Huang R. A survey on human motion generation tasks: consistency, diversity, and customization. In: 2024 7th Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII). IEEE Computer Society; 2024;269–277. Available from: https://doi.ieeecomputersociety.org/10.1109/CEII65291.2024.00059
17. Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics. 2020;8:64–77. Available from: https://doi.org/10.1162/tacl_a_00300
18. Karunratanakul K, Preechakul K, Aksan E, Beeler T, Suwajanakorn S, Tang S. Optimizing diffusion noise can serve as universal motion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024;1334–1345. Available from: https://arxiv.org/abs/2312.11994
19. Karunratanakul K, Preechakul K, Suwajanakorn S, Tang S. Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;2151–2162. Available from: https://arxiv.org/abs/2305.12577
20. Kim J, Kim J, Choi S. Flame: free-form language-based motion synthesis & editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37:8255–8263. Available from: https://doi.org/10.1609/aaai.v37i7.25996
21. Kulkarni N, Rempe D, Genova K, Kundu A, Johnson J, Fouhey D, Guibas L. Nifty: neural object interaction fields for guided human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024;947–957. Available from: https://arxiv.org/abs/2307.07511
22. Li B, Zhao Y, Zhelun S, Sheng L. Danceformer: music-conditioned 3D dance generation with a parametric motion transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2022;36:1272–1279. Available from: https://doi.org/10.1609/aaai.v36i2.20014
23. Li J, Wu J, Liu CK. Object motion guided human motion synthesis. ACM Transactions on Graphics. 2023;42(6):1–11. Available from: https://doi.org/10.1145/3618333
24. Li R, He C, Li S, Zhang Y, Zhang L. DynaMask: dynamic mask selection for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;11279–11288. Available from: https://openaccess.thecvf.com/content/CVPR2023/html/Li_DynaMask_Dynamic_Mask_Selection_for_Instance_Segmentation_CVPR_2023_paper.html
25. Li R, Yang S, Ross DA, Kanazawa A. AI choreographer: music-conditioned 3D dance generation with AIST++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021;13401–13412. Available from: https://arxiv.org/abs/2101.08779
26. Li R, Zhang Y, Zhang Y, Zhang H, Guo J, Zhang Y, Liu Y, Li X. Lodge: a coarse-to-fine diffusion network for long dance generation guided by characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024;1524–1534. Available from: https://arxiv.org/abs/2403.10518
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019. arXiv:1907.11692. Available from: https://arxiv.org/abs/1907.11692

28. Lu S, Chen L-H, Zeng A, Lin J, Zhang R, Zhang L, Shum H-Y. HumanToMaTo: text-aligned whole-body motion generation. arXiv. 2023. arXiv:2310.12978. Available from: https://arxiv.org/abs/2310.12978

29. Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ. AMASS: archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019;5442–5451. Available from: https://openaccess.thecvf.com/content_ICCV_2019/papers/Mahmood_AMASS_Archive_of_Motion_Capture_As_Surface_Shapes_ICCV_2019_paper.pdf

30. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–533. Available from: https://doi.org/10.1038/nature14236

31. 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

32. Petrovich M, Black MJ, Varol G. TEMOS: generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. Springer; 2022;480–497. Available from: https://arxiv.org/abs/2204.14109

33. Pi H, Peng S, Yang M, Zhou X, Bao H. Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;15061–15073. Available from: https://arxiv.org/abs/2310.02242

34. Pinyoanuntapong E, Saleem MU, Karunratanakul K, Wang P, Xue H, Chen C, Guo C, Cao J, Ren J, Tulyakov S. ControlMM: controllable masked motion generation. arXiv. 2024. arXiv:2410.10780. Available from: https://arxiv.org/html/2410.10780v1

35. Rempe D, Luo Z, Peng XB, 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. 2023;13756–13766. Available from: https://arxiv.org/abs/2304.01893

36. Siyao L, Yu W, Gu T, Lin C, Wang Q, Qian C, Loy CC, Liu Z. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022;11050–11059. Available from: https://arxiv.org/abs/2203.13055

37. Starke S, Zhang H, Komura T, Saito J. Neural state machine for character-scene interactions. ACM Transactions on Graphics. 2019;38(6):178. Available from: https://doi.org/10.1145/3355089.3356505

38. Tevet G, Gordon B, Hertz A, Bermano AH, Cohen-Or D. MotionCLIP: exposing human motion generation to CLIP space. In: European Conference on Computer Vision. Springer; 2022;358–374. Available from: https://arxiv.org/abs/2203.08063

39. Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH. Human motion diffusion model. arXiv. 2022. arXiv:2209.14916. Available from: https://arxiv.org/abs/2209.14916

40. Tseng J, Castellon R, Liu K. EDGE: editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;448–458. Available from: https://openaccess.thecvf.com/content/CVPR2023/papers/Tseng_EDGE_Editable_Dance_Generation_From_Music_CVPR_2023_paper.pdf

41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30. Available from: https://arxiv.org/abs/1706.03762

42. Wan W, Dou Z, Komura T, Wang W, Jayaraman D, Liu L. TLControl: trajectory and language control for human motion synthesis. In: European Conference on Computer Vision. Springer; 2024;37–54. Available from: https://link.springer.com/chapter/10.1007/978-3-031-72913-3_3

43. Wang J, Rong Y, Liu J, Yan S, Lin D, Dai B. Towards diverse and natural scene-aware 3D human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022;20460–20469. Available from: https://arxiv.org/abs/2205.13001

44. Wang Z, Chen Y, Liu T, Zhu Y, Liang W, Huang S. Humanise: language-conditioned human motion generation in 3D scenes. Advances in Neural Information Processing Systems. 2022;35:14959–14971. Available from: https://arxiv.org/abs/2210.09729

45. Xie Y, Jampani V, Zhong L, Sun D, Jiang H. OmniControl: control any joint at any time for human motion generation. arXiv. 2023. arXiv:2310.08580. Available from: https://arxiv.org/abs/2310.08580

46. Xu S, Li Z, Wang Y-X, Gui L-Y. InterDiff: generating 3D human-object interactions with physics-informed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;14928–14940. Available from: https://openaccess.thecvf.com/content/ICCV2023/papers/Xu_InterDiff_Generating_3D_Human-Object_Interactions_with_Physics-Informed_Diffusion_ICCV_2023_paper.pdf

47. 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. 2023;3836–3847. Available from: https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.pdf

48. Zhao K, Zhang Y, Wang S, Beeler T, Tang S. Synthesizing diverse human motions in 3D indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;14738–14749. Available from: https://openaccess.thecvf.com/content/ICCV2023/papers/Zhao_Synthesizing_Diverse_Human_Motions_in_3D_Indoor_Scenes_ICCV_2023_paper.pdf

49. Zhong C, Hu L, Zhang Z, Xia S. AttT2M: text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023;509–519. Available from: https://openaccess.thecvf.com/content/ICCV2023/papers/Zhong_AttT2M_Text-Driven_Human_Motion_Generation_with_Multi-Perspective_Attention_Mechanism_ICCV_2023_paper.pdf

50. Zhou Z, Wang B. UDE: a unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023;5632–5641. Available from: https://openaccess.thecvf.com/content/CVPR2023/html/Zhou_UDE_A_Unified_Driving_Engine_for_Human_Motion_Generation_CVPR_2023_paper.html