Text-driven Human Motion Generation with Motion Masked Diffusion Model

Xingyu Chen
University College London
xingyu.chen.23@ucl.ac.uk

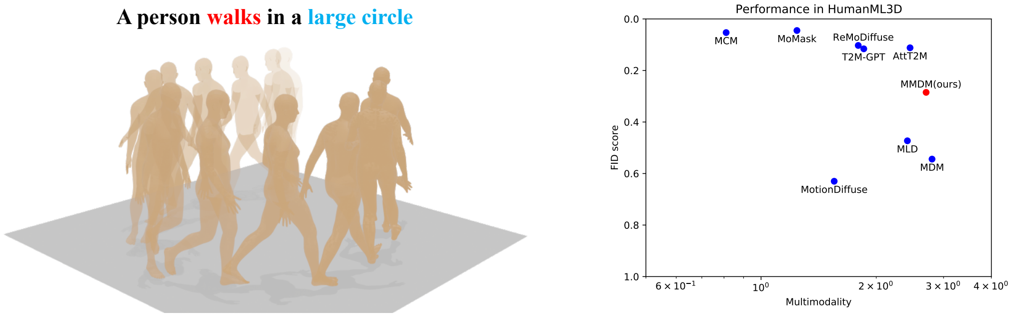
Abstract

Text-driven human motion generation is a multimodal task that synthesizes human motion sequences conditioned on natural language. It requires the model to satisfy textual descriptions under varying conditional inputs, while generating plausible and realistic human actions with high diversity. Existing diffusion model-based approaches have outstanding performance in the diversity and multimodality of generation. However, compared to autoregressive methods that train motion encoders before inference, diffusion methods lack in fitting the distribution of human motion features which leads to an unsatisfactory FID score. One insight is that the diffusion model lack the ability to learn the motion relations among spatio-temporal semantics through contextual reasoning. To solve this issue, in this paper, we proposed Motion Masked Diffusion Model **(MMDM)**, a novel human motion masked mechanism for diffusion model to explicitly enhance its ability to learn the spatio-temporal relationships from contextual joints among motion sequences. Besides, considering the complexity of human motion data with dynamic temporal characteristics and spatial structure, we designed two mask modeling strategies: **time frames mask** and **body parts mask**. During training, MMDM masks certain tokens in the motion embedding space. Then, the diffusion decoder is designed to learn the whole motion sequence from masked embedding in each sampling step, this allows the model to recover a complete sequence from incomplete representations. Experiments on HumanML3D and KIT-ML dataset demonstrate that our mask strategy is effective by balancing motion quality and text-motion consistency.

1 Introduction

Generating realistic human motions based on textual descriptions is a complex task due to its multimodal nature. This task bridges the gap between textual semantics and human motion, requires the model not only to accurately understand and translate input text into corresponding motion but also to generate coherent and naturalistic action. However, this task is inherently challenging due to the complexity and variability of human movements. The demand for such technologies is rapidly increasing in various fields, such as graphic animation and human-computer interaction, where generating diverse and contextually relevant motions from textual input can greatly enhance user immersive experience and believable digital content creation.

Existing research has extensively explored various text-driven human motion generation models and algorithms, among them one prevailing method is to construct an motion encoder for natural language inputs (Ahuja & Morency, 2019; Ghosh et al., 2021). TEMOS (Petrovich et al., 2022) trains a transformer variational autoencoder (VAE) architeture to learn the distribution parameters of text-motion latent space from KIT Motion-Language dataset (Plappert et al., 2016). MotionClip (Tevet et al., 2022) trains a encoder aligned with large pretrained CLIP (Radford et al., 2021) model. After that, T2M-GPT (Zhang et al., 2023a) introduces Vector Quantised-Variational AutoEncoder (VQ-VAE) (Van Den Oord et al., 2017) to learn a discrete motion representation and uses Generative Pretrained Transformer (GPT) for generation with an autoregressive paradigm. AttT2M (Zhong et al., 2023) proposes multi-perspective attention to learn the cross-modal relationship during text-driven motion generation stage. MoMask (Guo et al., 2024) use hierarchical quantization generative model (BERT) to predict the motion sequence based on token generation. Typically, these autoregressive



(a) Visualization

(b) Performance

Figure 1: **(a) Visualization for text-driven human motion sequence.** Our method balances the generation quality and diversity of the high-fidelity motion with the semantic consistency of the textual descriptions. **(b) Performance in HumanML3D dataset.** *FID* (lower is better) and *Multimodality* (higher is better) metrics the generation quality and average variance of motion sequences.

architectures learn an motion encoder before generation to capture the semantic representation of the input text, then, a separate decoder or generator model is trained to produce the corresponding motion sequence from these encoded features.

In addition to training motion encoder, there is another approach, Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) for human motion generation. MDM (Tevet et al., 2023) have gained prominence due to their ability to generate diverse and high-quality motion sequences in a single-stage process. MDM leverages the many-to-many nature of diffusion processes to generate diverse motion sequences from various forms of conditioning, such as text descriptions or action labels. MotionDiffuse (Zhang et al., 2022) leverages diffusion processes to achieve body part independent control with fine-grained texts by multi-level manipulation. ReMoDiffuse (Zhang et al., 2023b) combines retrieval augmented mechanism into diffusion model framework for improving text-motion consistency. MLD (Xin et al., 2023) learns a motion probabilistic mapping in the latent representation space and make the human motion generation more effective on text-to-motion and action-to-motion tasks.

Comparing to autoregressive architectures such as those built on the GPT (Radford, 2018) or BERT (Kenton & Toutanova, 2019) which follow a two-stage strategy and train a motion decoder before generation. Diffusion models, by contrast, avoid these pitfalls through their single-stage design, where the iterative refinement directly and consistently shapes the motion sequence in line with the text. This not only simplifies the training pipeline but also facilitates a more direct and coherent interaction between text and motion representations. However, recent research find diffusion architecture often struggle with contextual reasoning (Gao et al., 2023), especially in understanding and maintaining relationships among temporal and spatial semantics. This limitation arises from their architecture, which, while capable of generating diverse motions, does not inherently prioritize the learning of coherent temporal structures and contextual dependencies necessary for aligning motion with the nuances of language.

To enhance the representation learning capabilities of models, recent advancements in masked strategies (He et al., 2022; Chang et al., 2022) have shown significant progress. By masking portions of data during training, these models learn to infer robust representations of the missing parts from the visible features, thereby improving their ability to understand and predict contextual relationships. This approach has proven to be particularly effective in diffusion architecture (Gu et al., 2022; Gao et al., 2023), where learning to predict masked content fosters a deeper comprehension of the underlying data structure during denoise process. However, directly applying such a strategy into text-to-motion generation is still challenge due to the complexity of human motion data. Considering the dynamic temporal characteristics and spatial structure of motion, in this paper, we design two motion sequence embedding masked mechanism, time frames mask and body parts mask. During denoising process, MMDM masks certain tokens in the motion embedding space. Then, the decoder is designed to learn the whole motion sequence from masked embedding in each sampling step. MMDM leverages the masked strategy to explicitly improve the model's capacity to learn spatial and temporal relations. By adopting this strategy, our results indicate that MMDM not only significantly enhances the coherence and relevance of the motions to the provided textual inputs, but also balances overall quality and diversity of generated motions.

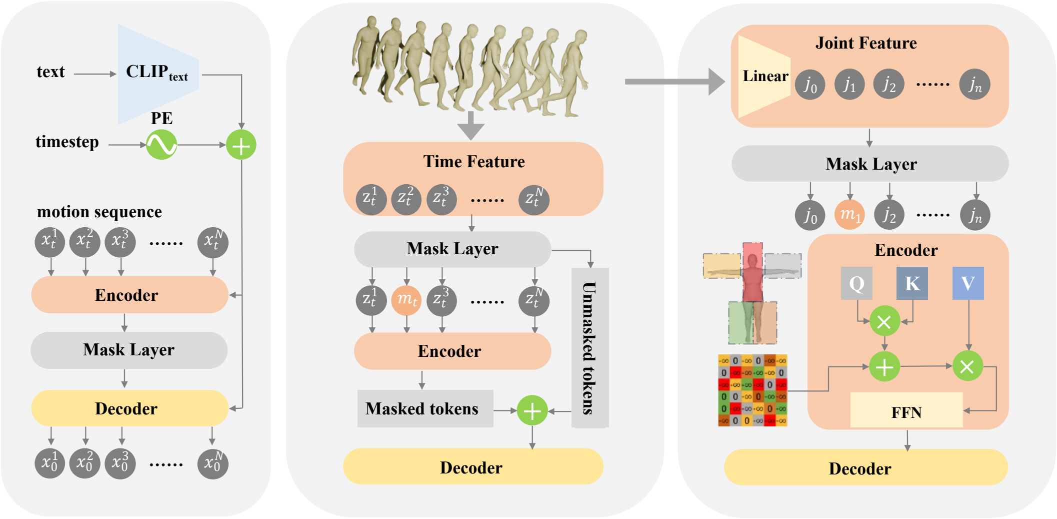
To summarize, the key contributions of our work are as follows: 1) We introduce a embedding space masking strategy into the motion diffusion process, allowing the model to focus on contextual inference of motion generation. 2) We design two Mask mechanism: time frames mask and body parts mask, in order to deal with temporal characteristics and spatial structure of human motion data. 3) Extensive experiment and evaluation on HumanML3D and KIT-ML datasets demonstrating the performance of MMDM, and verify the masked modeling is effective in motion diffusion.

2 RELATED WORKS

Human Motion Synthesis aims to generate realistic 3D human motion sequence. The early work focus on unconditional human motion generation (Yan et al., 2019; Zhang et al., 2020; Zhao et al., 2020), which random synthesis natural motion sequences from motion capture data without any constraint. In recent years, research begins to explore the human motion synthesis under various conditions, such as motion prefix (Mao et al., 2019; Liu et al., 2022), action label (Petrovich et al., 2021; Guo et al., 2020), textual description (Guo et al., 2022a;b; Tevet et al., 2023), image (Rempe et al., 2021; Chen et al., 2022) or audio signal (Siyao et al., 2022; Tseng et al., 2023; Gong et al., 2023; Zhou & Wang, 2023). Or use the pose sequences as input condition to complete incomplete motion sequences (Harvey et al., 2020; Duan et al., 2021). In this paper, we focus on text-driven motion generation, which is a sub-task of human motion synthesis under textual condition, since the textual descriptors are the convenient and easy to carve out motion details. In the early stage, Text2Action (Ahn et al., 2018) employs RNN architecture to train a text to motion mapping from short text. After that, Language2Pose (Ahuja & Morency, 2019) introduces a concept of Joint Language-to-Pose (JL2P) and applies curriculum learning approach to learn the joint embedding of language and pose. Similarly, MotionCLIP (Tevet et al., 2022) trains this joint embedding space with CLIP encoder. More recent, the current method can be divided into autoregressive architectures (Zhang et al., 2023a; Zhong et al., 2023; Guo et al., 2024) and diffusion architectures (Tevet et al., 2023; Zhang et al., 2022; 2023b; Xin et al., 2023).

Diffusion Generative Models models a stochastic diffusion process by a Markov chain, allowing the model to continuously learn the mapping between each sampling step from the inverse process, leading to denoised generation (Ho et al., 2020; Dhariwal & Nichol, 2021). It is also regarded as score-based generative modeling through stochastic differential equations perspective (Song & Ermon, 2019; 2020; Song et al., 2020b). Diffusion models combine both elegant physico-mathematical derivations and powerful generative performance, which have attracted plenty of attention from the community. Researchers proposed more efficient sampling strategies DDIM (Song et al., 2020a), DPM-solver (Lu et al., 2022) to improve diffusion models. In this paper, we focus on conditional generation for diffusion models. The early work (Dhariwal & Nichol, 2021) applies an extra classifier to guide the gradient during diffusion process. GLIDE (Nichol et al., 2021) follow this structure and condition on CLIP textual embedding feature. After that, research balance fidelity and diversity, propose Classifier-Free Guidance (Ho & Salimans, 2022), and also align it with CLIP (Ramesh et al., 2022). It becomes a dominant generative framework for visual tasks.

Generative Masked Modeling is a approach to improve the model ability in learning representations. At early stage, researchers demonstrated the effectiveness of masking methods in natural language processing in representation pretraining stage (Radford, 2018; Kenton & Toutanova, 2019) and language generation (Brown, 2020). BERT randomly masked out part of word tokens with a fixed ratio, and use incomplete langauge data to train a bi-directional transformer to predict the masked tokens. Then computer vision researchers transfer masked modeling approach from NLP area to vision area, and prove its effectiveness (He et al., 2022; Chang et al., 2022; Ji et al., 2023). In generative model, masking parts of data are beneficial in generating quality (Zhou et al., 2021), scalability (He et al., 2022), training convergence (Gao et al., 2022) and contextual reasoning ability (Gao et al., 2023). In this paper, we focus on generative masked modeling for human motion generation. Momask (Guo et al., 2024) first introduces BERT architecture and generative masked modeling into human motion synthesis and achieve the state-of-the-art FID score in humanML3D dataset (Guo et al., 2022a). Inspired by these successes, in this paper, we explore generative masked modeling approach for human motion diffusion models.


text
CLIP_text
PE
timestep
motion sequence
x_t^1 x_t^2 x_t^3 ...... x_t^N
Encoder
Mask Layer
Decoder
x_0^1 x_0^2 x_0^3 ...... x_0^N
Time Feature
z_t^1 z_t^2 z_t^3 ...... z_t^N
Mask Layer
z_t^1 m_t z_t^3 ...... z_t^N
Unmasked tokens
Encoder
Masked tokens
Decoder
Joint Feature
Linear
j_0 j_1 j_2 ...... j_n
Mask Layer
j_0 m_1 j_2 ...... j_n
Encoder
Q
K
V
FFN
Decoder

Model overview **Time frames mask** **Body parts mask**

Figure 2: **Overview and network**. We propose motion mask diffusion model (Left), including time frames mask (Middle) and body parts mask (Right) for expressive spatio-temporal features in the motion embedding. Given a natural language condition, a CLIP (Radford et al., 2021) transfer it into textual embedding and projected together with positional embedding of timestep t for classifier-free learning (Ho & Salimans, 2022). In each sampling step, our model follow MDM (Tevet et al., 2023) directly predicts the final clean motion sequence $x^{1:N}$ instead of noise, then repeats from x_t to x_0.

3 METHODOLOGY

In this paper, we demonstrate a Motion Masked Diffusion Model (MMDM). An overview framework of our method is presented in Figure 2. We introduces a masked embedding modeling scheme into the diffusion process to enhance the contextual reasoning ability for human motion generation.

3.1 HUMAN MOTION DIFFUSION MODEL

Motion Masked Diffusion Model (MMDM) proposed by Tevet et al. (2023), enables the diffusion model to learn text to motion representations for generation. Given an arbitrary text condition c, the purpose of MDM is to synthesize a human motion $X = [x^1, x^2, x^3, ..., x^N]$ with length N, where N is the number of time frames. The output human motion $x^{1:N} = \{x^i\}_{i=1}^N$ is a sequence of human poses represented by either joint positions or rotations with $x^i \in \mathbb{R}^{J\times D}$, where J is the number of joints and D is the dimension of the joint representations.

Diffusion Process. The diffusion probabilistic models is modeled by a Markov chain, and involves a forward noise adding process and a reverse denoising process. For sampling sequence $\{x_t^{1:N}\}_{t=0}^T$, where $x_0^{1:N}$ is drawn from the data distribution and

$$q(x_t^{1:N}|x_{t-1}^{1:N}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}^{1:N}, (1-\alpha_t)I), \tag{1}$$

where $\alpha_t \in (0, 1)$ are constant hyper-parameters. In human motion data, conditioned motion synthesis models the distribution $p(x_0|c)$ as the reversed diffusion process of gradually cleaning x_T.

$$\mathcal{L}_{\text{simple}} = E_{x_0\sim q(x_0|c),t\sim[1,T]}[\|x_0 - \hat{x}_0\|_2^2] \tag{2}$$

Instead of predicting ϵ_t then add it into sampling data as formulated by Ho et al. (2020), we follow Ramesh et al. (2022) and predict the human motion sequences itself (Tevet et al., 2023). This allows more geometric constraints to be added directly to the optimization function for controlling more realistic human motion generation.

Optimization function. We follow the standard geometric regularization for human motion domain from Petrovich et al. (2021); Shi et al. (2020). 1) Position Loss: This loss ensures that the predicted joint positions are consistent with the true joint positions. 2) Foot Contact Loss: To prevent foot sliding and maintain realistic foot-ground interactions, this term penalizes deviations in foot position during contact with the ground. 3) Velocity Loss: This loss encourages smooth and consistent motion by regulating the velocity of the joints.

$$\mathcal{L}_{\text{pos}} = \frac{1}{N} \sum_{i=1}^{N} \|FK(x_0^i) - FK(\hat{x}_0^i)\|_2^2, \tag{3}$$

where $FK(\cdot)$ is the forward kinematics function that converts joint rotations into positions.

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i\|_2^2, \tag{4}$$

where f_i is a binary indicator of foot contact at frame i.

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\|_2^2 \tag{5}$$

The final training loss combines these geometric losses with the standard diffusion loss:

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}. \tag{6}$$

where λ_{pos}, λ_{foot}, and λ_{vel} are weighting factors that balance the contributions of the geometric losses.

3.2 Motion embedding Masked Diffusion

To enhance the contextual reasoning ability of diffusion models, we introduce masking strategy within the embedding space during the diffusion process, termed MMDM. The core idea of Motion Masked Diffusion Model is randomly masking a subset of the input in order to force the model reasoning about missing parts based on incomplete data, which can lead to a more robust and contextually aware model. Specifically, at each time step, a certain proportion of the embedding variables corresponding to the motion sequence are masked, effectively hiding parts of the motion from the model. The model is then tasked with reconstructing the entire sequence, including the masked components, from the remaining visible parts. This forces the model to learn to infer the missing context, thereby enhancing its ability to understand and predict the complex relationships between spatial and temporal semantics in human motion.

Given a motion embedding $z_{1:N}$ of length N, where each frame z_i represents the features of the time frames or pose joints at time step i. At the start of the diffusion process, a mask M is initialized by randomly selecting with mask ratio, and a learnable embedding q is initialized by mask token adding positional encoding. The mask can be designed to either randomly select time frames to mask at each step or to focus on specific joints known to be critical for body parts. This depends on the mask type, it will be introduced in the next section. Mask token can ensure the decoder always receives same size of motion embedding during training and inference. At each diffusion step t, noise is added to the motion sequence based on the mask M. The noisy sequence $\tilde{x}_{1:N}^t$ at step t is given by:

$$\tilde{x}_{1:N}^t = M \odot q + (1 - M) \odot x_{1:N}^{t-1}$$

where $\odot$ denotes element-wise multiplication. The mask M ensures that mask token is only applied to the selected parts of the motion embedding, allowing the model to focus on both denoising and contenxtual reasoning.

Traditional mask modeling apply mask token to the parts of input data, however, the human motion sequence is high dimensions data with three dimensions in space and one dimension in time frames. The key innovation in MMDM lies in the masked strategies for human motion data. Consider its complexity in spatial structure and dynamic temporal characteristics, we design two mask modeling: time frames mask and body parts mask.

3.2.1 Time frames mask

For time frames mask, we introduces the asymmetric diffusion transformer(Gao et al., 2023) which including an encoder and a decoder for joint training of mask embedding modeling and diffusion process. During training, the encoder processes masked embedding and predict the full tokens, and the masked tokens from encoder prediction will be added together with the unmasked portion to be fed into the decoder for diffusion training shown as Figure2 middle part.

Time frames encoder. Firstly, the motion embedding takes masking operation on the time dimension, if a time frame is selected, it will be substituted by a learnable mask token. Secondly, to enhance the relative position information during learning, positional encoding is necessary for all motion embedding. In our model, the encoder adds the conventional learnable global position embedding into the noisy embedding latent input. Thirdly, a transformer block is designed for computing the attention score of self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_r\right)V \tag{7}$$

where Q, K, and V, respectively denote the query, key, and value in the self-attention module, d_k is the dimension of the key, and B_r is the relative positional bias (Liu et al., 2021). The relative positional bias is helpful to capture the token relationship, facilitating the prediction on masked tokens.

Time frames decoder. Similar to encoder, the decoder also adds learnable position embedding into its input motion tokens to enhance positional information. But the difference between them is that the encoder is concerned only with the unmasked portion, while the decoder needs to be concerned with all tokens, so this structure is asymmetric diffusion transformer architecture(Gao et al., 2023).

3.2.2 Body parts mask

For body parts mask, we introduces Body-Part attention-based Spatio-Temporal(BPST) encoder (Zhong et al., 2023) for learning the body parts features. Then we mask out a portion of the body parts features for diffusion training shown as Figure2 right part.

Body parts encoder. Firstly, we follow the BPST encoder (Zhong et al., 2023) divide the human body skeleton with n joints into five body parts: *Torso, Left Arm, Right Arm, Left Leg, Right Leg*, each containing its own set of joints. Secondly, a linear layer is applied for mapping all tokens into the same dimension before computing self-attention. Thirdly, we construct a adjacency matrix M from the joint information. For example, if a joint i and joint j belong to the same body parts $m_{i,j} = 0$, otherwise $m_{i,j} = -\infty$. Then, we concatenate the mapping results of different tokens after obtaining the feature from parts to compute the body-part attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top + M}{\sqrt{d_k}}\right)V \tag{8}$$

where Q, K, and V, respectively denote the query, key, and value in the self-attention module, d_k is the dimension of the key, and M is body parts adjacency matrix (Zhong et al., 2023). After obtaining the final spatial feature of body parts, we takes mask operation for randomly select portion of body parts feature and substitute with learnable mask tokens.

Body parts decoder. Similar to time frames decoder, we also adds position encoding to learn the position information. However, unlike the masking operation on the previous time frames mask, the masking operation on the body part features is executed after encoder. So the body parts decoder needs to complete the mask motion data and take diffusion learning at the same time.

4 EXPERIMENTS

We implement MMDM using the PyTorch framework. The time-frames mask transformers are initialized with 2 layers encoder and 6 layers decoder with a hidden dimension of 512. The body-parts mask transformer decoder is initialized with 6 layers and a hidden dimension of 640. We train

Methods	R Precision↑			FID.↓	MM-D.↓	Div.→	MM.↑
	Top-1	Top-2	Top-3				
Real	0.511 ± .003	0.703 ± .003	0.797 ± .002	0.002 ± .000	2.974 ± .008	9.503 ± .065	-
T2M	0.455 ± .003	0.636 ± .003	0.736 ± .002	1.087 ± .021	3.347 ± .008	9.175 ± .083	2.219 ± .074
MDM	0.320 ± .005	0.498 ± .004	0.611 ± .007	0.544 ± .044	5.566 ± .027	**9.559** ± .086	**2.799** ± .072
MotionDiffuse	**0.491** ± .001	**0.681** ± .001	**0.782** ± .001	0.630 ± .001	**3.113** ± .001	<u>9.410</u> ± .049	1.553 ± .042
MLD	<u>0.481</u> ± .003	0.673 ± .003	0.772 ± .002	0.473 ± .013	3.196 ± .010	9.724 ± .082	2.413 ± .079
T2M-GPT	**0.491** ± .003	<u>0.680</u> ± .003	<u>0.775</u> ± .002	**0.116** ± .004	<u>3.118</u> ± .011	9.761 ± .081	1.856 ± .011
MMDM-t	0.464 ± .006	0.654 ± .007	0.754 ± .005	0.319 ± .026	3.288 ± .023	9.299 ± .064	<u>2.741</u> ± .112
MMDM-b	0.435 ± .006	0.627 ± .006	0.733 ± .007	<u>0.285</u> ± .032	3.363 ± .029	9.398 ± .088	2.701 ± .083

Table 1: **Quantitative evaluation on the testset of HumanML3D.** We report the metrics following T2M (Guo et al., 2022a) and repeat 20 times to get the average results with 95% confidence interval. The $\downarrow, \uparrow, and \rightarrow$ denote the lower, higher, and closer to Real are better, respectively. The best results are marked in bold and the second best is underlined. Our method achieves significant improvement on almost all metrics.

Methods	R Precision↑			FID.↓	MM-D.↓	Div.→	MM.↑
	Top-1	Top-2	Top-3				
Real	0.424 ± .005	0.649 ± .006	0.779 ± .006	0.031 ± .004	2.788 ± .012	11.08 ± .097	-
T2M	0.361 ± .006	0.559 ± .007	0.681 ± .007	3.022 ± .107	3.488 ± .028	10.72 ± .145	2.052 ± .107
MDM	0.164 ± .004	0.291 ± .004	0.396 ± .004	0.497 ± .021	9.191 ± .022	10.85 ± .109	1.907 ± .214
MotionDiffuse	**0.417** ± .004	<u>0.621</u> ± .004	<u>0.739</u> ± .004	1.954 ± .064	<u>2.958</u> ± .005	**11.10** ± .143	0.730 ± .013
MLD	0.390 ± .008	0.609 ± .008	0.734 ± .007	<u>0.404</u> ± .027	3.204 ± .027	10.80 ± .117	<u>2.192</u> ± .071
T2M-GPT	0.402 ± .006	0.619 ± .005	0.737 ± .006	0.717 ± .041	3.053 ± .026	<u>10.86</u> ± .094	1.912 ± .036
MMDM-t	<u>0.432</u> ± .006	**0.643** ± .007	**0.760** ± .006	**0.237** ± .013	**2.938** ± .025	10.84± .125	1.457 ± .129
MMDM-b	0.386 ± .007	0.603 ± .006	0.729 ± .006	0.408 ± .022	3.215 ± .026	10.53 ± .100	**2.261** ± .144

Table 2: **Quantitative evaluation on the testset of KIT-ML.** The experimental settings are the same as Table 1.We report the metrics following T2M Guo et al. (2022a) and repeat 20 times to get the average results with 95% confidence interval. The best results are marked in bold and the second best is underlined.

the model using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 64. The cosine noise schedule is applied during the diffusion process, with 1,000 noising steps. For the masking mechanism, the mask token is dynamically updated based on the gradients of the loss function during training, allowing the model to learn the most important frames or joints for accurate generation.

4.1 SET UP

Dataset. We conduct our experiments on two widely used human motion datasets:

- HumanML3D (Guo et al., 2022a): This dataset contains 14,616 motion sequences annotated with 44,970 textual descriptions. It includes a wide variety of human actions and motions, providing a comprehensive benchmark for text-to-motion generation tasks.
- KIT Motion-Language (Plappert et al., 2016): The KIT dataset consists of 3,911 motion sequences paired with textual descriptions. Although smaller in size compared to HumanML3D, it is commonly used in text-to-motion research, making it an important benchmark for evaluating model performance.

Evaluate metrics. To quantitatively evaluate the performance of our model, we use the following metrics: Fréchet Inception Distance (FID): This metric measures the similarity between the distribution of generated motions and the ground truth motions. Lower FID scores indicate better performance. R-Precision: This metric evaluates the relevance of generated motions to the input textual descriptions by computing the top-k accuracy of the retrieval results. Higher R-Precision scores indicate better alignment with the input text. Diversity: This metric assesses the variability in the generated motion sequences, ensuring that the model does not collapse to generating a limited set of motions. Multimodality: This metric measures the average variance of generated motions given a single text prompt, reflecting the model's ability to generate diverse outputs from the same input.

Quantitative evaluation. Table 1 and Table 2 demonstrate MMDM performance in the text-driven human motion diffsuion task on the HumanML3D and KIT datasets respectively. We conduct 20 evaluations, with 1000 samples in each, and report their average and a 95% confidence interval.

4.2 Results and Analysis

4.3 Ablation Study on Masking Mechanism

Dataset	Mask Type	Ratio	FID↓	Top-3 R Precision↑	MM-D.↓	Div.→	MM.↑
HumanML3D	MMDM-t	0.1	0.286	**0.743**	3.383	**9.361**	**2.795**
		0.2	**0.276**	0.742	**3.355**	9.285	2.741
		0.3	0.302	0.734	3.426	9.288	2.674
		0.4	0.349	0.733	3.422	9.144	2.661
	MMDM-b	0.1	**0.252**	**0.744**	**3.338**	**9.442**	2.701
		0.2	0.614	0.712	3.588	8.765	**2.832**
		0.3	1.539	0.689	3.723	8.409	2.803
		0.4	1.542	0.677	3.810	8.424	2.829
KIT	MMDM-t	0.1	**0.234**	0.767	2.937	10.77	1.535
		0.2	0.278	**0.772**	**2.925**	**10.84**	1.500
		0.3	0.328	0.770	3.004	10.74	**1.567**
		0.4	0.366	0.749	3.038	10.72	1.550
	MMDM-b	0.1	0.923	0.664	3.755	10.21	**2.695**
		0.2	**0.449**	**0.735**	**3.196**	10.49	2.261
		0.3	0.481	0.714	3.312	10.40	2.440
		0.4	0.697	0.693	3.550	**10.57**	2.399

Table 3: **Effect of different masking ratios.** We repeat 5 times to get the average results with 95% confidence interval under different mask ratio. The $\downarrow, \uparrow, and \rightarrow$ denote the lower, higher, and closer to Real are better, respectively.

To analyze the impact of the masking mechanism, we conduct an ablation study where we compare the performance of MMDM with different mask ratio. As the mask rate increases, the model focuses more on the mask portion rather than diffusion generation. In Table 3, Diffusion masking on time frames or on body part features, the optimal masking ratio are both in the interval 0.1-0.2.

4.4 Ablation Study on Model Architecture

Arch	FID↓	Top-3 R Precision↑	MM-D.↓	Div.→	MM.↑
04 Encoder+2 Decoder	0.461	0.721	3.469	9.136	2.684
06 Encoder+2 Decoder	**0.232**	**0.746**	**3.333**	9.335	2.596
08 Encoder+4 Decoder	0.296	0.742	3.405	9.090	**2.694**
12 Encoder+4 Decoder	0.369	0.731	3.399	**9.442**	2.638

Table 4: **Effect of different architecture.** We repeat 5 times to get the average results under 95% confidence interval with mask ratio 0.2 in HumanML3D dataset. The $\downarrow, \uparrow, and \rightarrow$ denote the lower, higher, and closer to Real are better, respectively.

To analyze the impact of the masking mechanism, we also conduct an ablation study where we compare the performance under different number of layers. As shown in Table 4, We tested four different sets of model capacities for time frames mask. From the FID score, the model size of a 6-layer encoder with a 2-layer decoder is optimal for the HumanML3D dataset. For a fair comparison, we followed this layer structure for our experiments on the body parts mask and KIT dataset.

5 Conclusion

In this paper, we introduced Motion Masked Diffusion model (MMDM), a novel approach that integrates the generative masking strategy into the Human Motion Diffusion Model. Our proposed model addresses key challenges in human motion diffusion model to improve its FID-score performance, while maintaining diversity and realism in generated motions.

References

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

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

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Xin Chen, Zhuo Su, Lingbo Yang, Pei Cheng, Lan Xu, Bin Fu, and Gang Yu. Learning variational motion prior for video-based motion capture. *arXiv preprint arXiv:2210.15134*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Towards sustainable self-supervised learning. *arXiv preprint arXiv:2210.11016*, 2022.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23164–23173, 2023.

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1396–1406, 2021.

Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9942–9952, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022a.

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

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

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion inbetweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. Masked vision-language transformer in fashion. *Machine Intelligence Research*, 20(3):421–434, 2023.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhenguang Liu, Shuang Wu, Shuyuan Jin, Shouling Ji, Qi Liu, Shijian Lu, and Li Cheng. Investigating pose representations and motion contexts modeling for 3d motion prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):681–697, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9489–9497, 2019.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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

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

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

Alec Radford. Improving language understanding by generative pre-training. 2018.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.

Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *Acm transactions on graphics (tog)*, 40(1):1–15, 2020.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=SJ1kSyO2jwu`.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458, 2023.

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

Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4394–4402, 2019.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023a.

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

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 364–373, 2023b.

Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020.

Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6225–6234, 2020.

Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 509–519, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5632–5641, 2023.