MoDDM: Text-to-Motion Synthesis using Discrete Diffusion Model

Ankur Chemburkar
achemburkar@ict.usc.edu
Shuhong Lu
slu@ict.usc.edu
Andrew Feng
feng@ict.usc.edu

Institute for Creative Technologies
University of Southern California
Los Angeles, USA

Abstract

We present the motion discrete diffusion model (MoDDM) for synthesizing human motion from text descriptions that addresses challenges in cross-modal mapping and motion diversity. The previous methods that utilized variational autoencoder (VAE) to learn the latent distributions for text-to-motion synthesis tend to produce motions with less diversity and fidelity. While the diffusion models show promising results by generating high quality motions, they require higher computational costs and may produce motions less aligned with the input text. The proposed method combines the discrete latent space and diffusion models to learn an expressive conditional probabilistic mapping for motion synthesis. Our method utilizes vector quantization variational autoencoder (VQ-VAE) to learn discrete motion tokens and then applies discrete denoising diffusion probabilistic models (D3PM) to learn the conditional probability distributions for the motion tokens. The discrete classifier-free guidance is further utilized in the training process with proper guidance scale for aligning the motions and the corresponding text descriptions. By learning the denoising model in the discrete latent space, the method produces high quality motion results while greatly reducing computational costs compared to training the diffusion models on raw motion sequences. The evaluation results show that the proposed approach outperforms previous methods in both motion quality and text-to-motion matching accuracy.

1 Introduction

Synthesizing believable human motions based on input conditions is an essential task for computer vision and animation that will find many applications in gaming, simulation, and virtual reality. Various conditional inputs can be utilized to drive the motion synthesis process such as speech, music, action categories, and natural language text descriptions. Generating motions from text descriptions requires modeling of both languages and motions, which is especially challenging as the model needs to learn a cross-modal mapping from input free-form text to output motion sequences.

One of the challenges for motion synthesis from texts is that the generation process is not a one-to-one mapping problem. For example, the description "A person stands up" only

indicates the action but does not provide information about the current state of the person. Thus the person could be either sitting or lying down on the ground – both will be valid states for connecting into the "stands up" action. Similarly, if the motion styles are not explicitly defined in the description (i.e. walking fast or slowly), then motions with varying moving speed will correspond correctly to the same input text description as far as the walking actions are presented. Previous works [22, 23] have utilized variaional autoencoder (VAE) to learn a normal latent distribution for modeling such variations from the same text input. While these methods produce reasonable results, the synthesized motions have less diversity since the motions have to be sampled from a learned target distributions. To address the diversity issues, the denoising diffusion probabilistic models (DDPM) are good candidates for text-to-motion synthesis tasks. The main advantage is that the de-noising process is not tied to a particular latent space distributions (i.e. Gaussian) when learning the underlying data distributions. As shown in the previous works [32], adapting the diffusion models for motion synthesis tasks produces an expressive probabilistic mapping that are capable of generating diverse motion results. However, applying diffusion models on raw motion data requires higher computational costs due to large numbers of diffusion steps for each individual joint parameters. Moreover, from the experiments we also found that diffusion models may produce motions that are less aligned with the input text conditions. Therefore while the diffusion method generates high quality motions, the results may not be suitable for text-to-motion synthesis task.

In this paper, we propose to address the aforementioned issues for the diffusion models in human motion synthesis by utilizing discrete latent space learning. Specifically, we proposed to utilize vector quantization variational autoencoder (VQ-VAE) to learn discrete latent codes from the input motions. These discrete codes are then used by the discrete denoising diffusion probabilistic models (D3PM) to learn the denoise process. By learning the denoising model in the discrete latent space, the method not only produces high quality motion results from the diffusion process, but also greatly reduces the computational costs by requiring much fewer diffusion steps to converge. Our objective and subjective evaluations show that the proposed method performs well in both motion fidelity and correlations to the input texts.

Our contributions are summarized as the following: 1) We model the text-to-motion generation task as a discrete denoising diffusion probablistic model, which allows reduced diffusion steps for faster inferences while producing high quality results. 2) We evaluated our method in a comparison with state-of-the-art methods using both objective metrics and subjective user study. The results demonstrated that our method outperforms the previous methods in both motion quality and text-to-motion matching accuracy. The user evaluation results also showed that the proposed approach produces motions that are preferred over previous methods.

2 Related Works

2.1 Text-to-Motion Synthesis

Deterministic approaches learn a direct mapping between input text and human motions with either a sequence-to-sequence model [25] or an encoder-decoder architecture based on a gated recurrent unit [6]. However, the text-to-motion synthesis is not a one-to-one mapping and the same text description can correspond to multiple different human motions that are all

plausible. To address such stochastic mapping, other approaches have adapted probabilistic models including GANs [18], normalizing flows [11, 36], and variational autoencoder (VAE) [1, 9, 10, 22, 23] for motion synthesis, where VAE is theoretically proved to have more stability in training [4]. VAE [16] encodes the training data distribution into continuous latent variables, which can then be used to synthesize new data via latent space sampling. It has been applied successfully for both action-conditioned motion synthesis and text-to-motion synthesis tasks [22, 23] to model the text prompts and motions as parameterized normal distributions in the latent space. The recent work [9] improved the VAE sampling by adding local semantic motion contexts with time variant attention with sentence features in the motion generation stage.

The vector quantized variational autoencoder (VQ-VAE) [33] was introduced for learning the discrete representations as codebooks to address the issue of posterior collapse in autoregressive sampling stages. The synthesize tasks are modelled as a two-stage training process that uses VQ-VAE and transformer models [26]. In the first stage, the VQ-VAE is trained to learn the discrete latent space representation by learning the codebooks for reconstructing the training data. In the second stage, a transformer model is used to learn conditional priors over the discrete latent codes through autoregressive sequential prediction. Such two-staged architecture has demonstrated excellent performance in time series data, as exemplified by Videogpt [35]. It has also been adapted for text2motion and motion2text tasks by Guo et al. in TM2T [10] to allow both forward and inverse alignment between texts and motions. They leverage 1D CNN based latent quantization to encode motion features followed by autoregressive translation networks to model the mapping between text and motion. Our approach follows a similar vector quantization step but significantly differs in the second stage where we make use of discrete diffusion explained in the subsequent sections. Another very recent work T2M-GPT [38], also implemented the VQ-VAE to utilize the discrete motion features which are employed by an autoregressive Generative Pretrained Transformer (GPT) model showing promising results. Our model uses the discrete diffusion process that has the global context of the latent motion representation, while the autoregressive models will only attend to the past context.

2.2 Diffusion Models

Diffusion models [13, 27, 28, 29, 30] have recently gained attention in the field of image and human motion synthesis due to their ability to generate complex and realistic results [5, 20]. They, like VAEs are generative models based on the principle of approximate density estimation. Unlike autoregressive generative models, diffusion models are not limited by step-by-step sampling, allowing for more flexibility and reduced error accumulation during inference.

In the continuous diffusion process, a data sample of any modality (motion in our setting) is first iteratively injected with gaussian noise through a forward Markov process until pure noise is obtained. In the subsequent reverse process, the model learns to gradually denoise the sample. Diffusion transformer frameworks have been applied in motion synthesis domains such as audio conditioned gesture generation task [37] to better handle long-term dependencies in gesture sequences. For human motion synthesis, there have been several notable adaptations of diffusion models [3, 15, 32, 39]. Tevet et al. [32] applied diffusion models in continuous space for generating raw motion frames, which showed improvements in motion quality. The later work adapted time-varying weight schedule for noise estimation and addressing the jittering problem in motion generation [3] .

D3PM [2] extended the application of diffusion models to discrete data, such as categorical labels or text, using a transition matrix in the noising step. VQ-Diffusion model [7] based on a VQ-VAE whose latent space is modeled by a conditional DDPM variant has been implemented for the task of text-to-image synthesis. To our knowledge, no variant of discrete diffusion has been applied to the text-motion domain.

Motivated by the ideas of VQ-VAE and continuous diffusion, we applied the discrete diffusion model to human motion synthesis. Our approach involves a two-stage process. In the first stage, we utilize VQ-VAE for learning the discrete latent codes by reconstructing input data. In the second stage, we gradually corrupt the codes in the latent space by applying transition matrices, allowing the model to recover the discrete latent codes. To address the issue of the network ignoring text conditions in the denoising process, we also incorporate discrete classifier-free guidance [12, 31] during training.

3 Method

Our MoDDM method is summarized in Figure 1. The proposed method trains the motion synthesis models using a two-stage architecture, where the first stage consists of learning discrete motion tokens via VQ-VAE and the second stage utilizes the discrete diffusion model to learn the conditional token distributions.

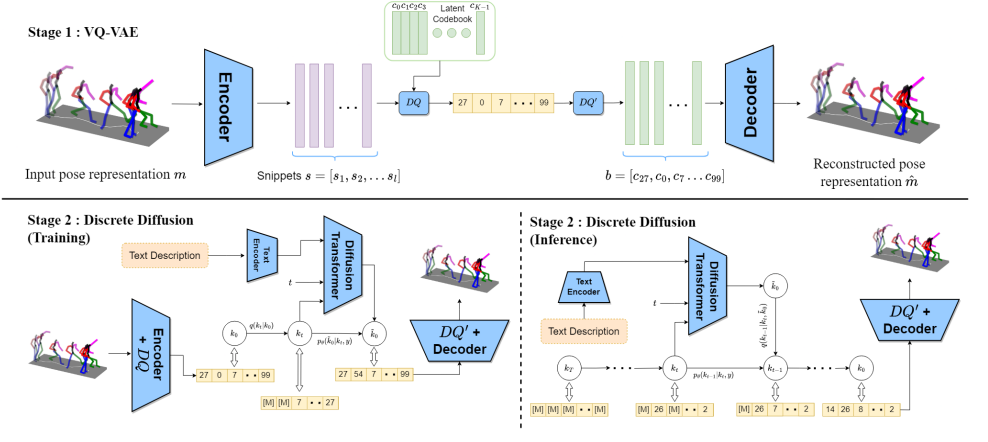

Figure 1: Architecture for VQ-Diffusion. The top half represents the VQ-VAE model framework. Bottom left figure briefly shows the forward and reverse process for training stage in Diffusion. Bottom right figure explains inference stage with the reparametrization trick.

3.1 Motion Token Learning

We make use of a latent space vector quantization model pre-trained on the domain of 3D human motions. Given a human motion represented by a sequence of poses $\mathbf{m} \in \mathbb{R}^{L \times D_m}$, where L is the sequence length and D_m is the dimensions of a single motion frame, an encoder $\mathbf{E}$ converts these poses into motion snippets $\mathbf{s} \in \mathbb{R}^{l \times h}$ with the number of snippets l being much less than L and h being the latent dimension. These snippets are then transformed into quantized vectors $\mathbf{b} \in \mathbb{R}^{l \times h}$ through the process of discrete quantization DQ with a learned

codebook C consisting of K embedding entries $(\mathbf{c}_1, ...\mathbf{c}_K)$ of dimensions $\mathbb{R}^h$. The process DQ transforms $\mathbf{s}$ by comparing each snippet $(\mathbf{s}_i)_{i=1}^t$ to all entries in C and replacing the snippet with the index of the nearest entry. Hence, the process DQ is defined as

$$\mathbf{k}_i = argmin_{\mathbf{c}_j \in C} ||\mathbf{s}_i - \mathbf{c}_j|| \tag{1}$$

The reverse process of the quantization DQ' converts the indices $\mathbf{k}$ into the corresponding entries $\mathbf{b}$ from codebook C to obtain the latent embedding for each motion snippet. Finally, a decoder D reconstructs $\mathbf{b}$ back to the 3D human motion space. Overall, this autoencoder process can be formulated as

$$\widehat{\mathbf{m}} = D(DQ'(DQ(E(\mathbf{m})))) \tag{2}$$

Following [10], this process is trained with two losses - a reconstruction loss between $\mathbf{m}$ and $\widehat{\mathbf{m}}$, and a codebook embedding loss to update the codebook entries and stabilize training. The loss equation is given by:

$$\mathbf{L}_{vq} = ||\widehat{\mathbf{m}} - \mathbf{m}||_1 + ||sg[\mathrm{E}(\mathbf{m})] - \mathbf{b}||_2^2 + \beta ||\mathrm{E}(\mathbf{m}) - sg[\mathbf{b}]||_2^2 \tag{3}$$

Here, sg[.] denotes stop gradient operation and β is a weighting factor. As the quantization process DQ is clearly non-differentiable, straight-through gradient estimator [34] was employed to enable back-propagation.

3.2 Diffusion for Discrete Motion Tokens

Discrete diffusion model is very similar to its continuous counterpart. Given a sequence of discrete motion tokens $\mathbf{k}_0 \in \mathbb{I}^l$, where the subscript denotes the diffusion step, the forward diffusion process gradually corrupts the sample through a Markov chain $q(\mathbf{k}_t|\mathbf{k}_{t-1})$. Following the discrete diffusion process [7], we use the forward process of randomly masking or replacing the tokens in $\mathbf{k}$ and obtaining increasingly noisy latent variables $\mathbf{k}_1, ..., \mathbf{k}_T \in \mathbb{I}^l$, where T is the total number of diffusion steps. Here $\mathbf{k}_T$ is pure noise, or all masked tokens in our case of discrete diffusion. The reverse diffusion process tries to reconstruct $\mathbf{k}_0$ from $\mathbf{k}_T$ by sampling from a reverse distribution $q(\mathbf{k}_{t-1}|\mathbf{k}_t, \mathbf{k}_0)$. Since $\mathbf{k}_0$ is not known during inference time, a transformer model is trained as the denoising model to approximate the reverse distribution. The distribution obtained from the transformer is denoted by $p_\theta(\mathbf{k}_{t-1}|\mathbf{k}_t, \mathbf{y})$, where $\mathbf{y}$ is the condition (text in our case). Fixed transition matrices for each timestep $Q_t \in \mathbb{R}^{(K+1)\times(K+1)}$ are used to define the transitional probabilities between codebook indices, where

$$Q_t = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \beta_t & \dots & 0 \\ \beta_t & \alpha_t + \beta_t & \beta_t & \dots & 0 \\ \beta_t & \beta_t & \alpha_t + \beta_t & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_t & \gamma_t & \gamma_t & \dots & 1 \end{bmatrix} \tag{4}$$

The extra dimension in K+1 denotes the [MASK] token. As defined in Q_t, at every diffusion step, an index in $\mathbf{k}_t$ has a probability $K\beta_t$ of being replaced by another index randomly from the K indices, γ_t probability of becoming a [MASK] index and α_t probability of remaining the same index.

The cumulative transition matrix $\overline{Q_t} = Q_t \dots Q_1$ that defines the transition probability from $\mathbf{k}_0$ to $\mathbf{k}_t$ and the corresponding forward probability distribution $q(\mathbf{k}_t|\mathbf{k}_0)$ have a closed form equation [7] that can be computed directly, which allows for an efficient forward diffusion process during training. During the reverse process, the model learns to approximate the posterior $q(\mathbf{k}_{t-1}|\mathbf{k}_t, \mathbf{k}_0)$ with $p_\theta(\mathbf{k}_{t-1}|\mathbf{k}_t, \mathbf{y})$ as mentioned above. Recent works [2, 14] perform a reparameterization trick that leads to better generation results. Instead of modelling the posterior directly, it approximates the distribution by generating the denoised motion tokens which is given by the denoising model as $p_\theta(\tilde{\mathbf{k}}_0|\mathbf{k}_t, \mathbf{y})$. During inference, we sample the $t-1^{th}$ motion from $p_\theta(\mathbf{k}_{t-1}|\mathbf{k}_t, \mathbf{y})$ using the denoised token distribution $p_\theta(\tilde{\mathbf{k}}_0|\mathbf{k}_t, \mathbf{y})$ and the posterior distribution $q(\mathbf{k_{t-1}}|\mathbf{k_t}, \tilde{\mathbf{k}}_0)$.

3.2.1 Classifier-Free Guidance

During the training process of a conditional generation task with $\mathbf{k}$ as a sample and $\mathbf{y}$ as the corresponding condition, the diffusion model tries to optimize the prior distribution $p(\mathbf{k}|\mathbf{y})$ assuming that the posterior distribution $p(\mathbf{y}|\mathbf{k})$ is satisfied. However, it is possible that this posterior probability is ignored during training. Since the model has access to both the corrupted sample and the condition, it is possible the the model only utilizes the corrupted sample to reconstruct and disregard the conditional input. This leads to the poor alignment between the generated sample and the condition, which is known as the *posterior issue* [31].

Therefore, our optimization target needs to include both $p(\mathbf{k}|\mathbf{y})$ as well as $p(\mathbf{y}|\mathbf{k})$. The simplest way is to optimize $\log p(\mathbf{k}|\mathbf{y}) + s \log p(\mathbf{y}|\mathbf{k})$, where s denotes the guidance scale which is a hyper-parameter. By using Bayes' Theorem, this optimization function can be reduced to:

$$argmax_{\mathbf{k}} = [\log p(\mathbf{k}) + (s+1)(\log p(\mathbf{k}|\mathbf{y}) - \log p(\mathbf{k}))] \tag{5}$$

where $p(\mathbf{k})$ is the unconditional distribution of $\mathbf{k}$. To handle the unconditional inputs, the model is also trained with 'null' condition [21] for a select percentage of samples. It has been shown that implementing a learnable conditional vector instead of 'null' condition is more suitable for training classifier-free guidance [31]. We adopt such technique with learnable null vector for our implementation. As shown in the ablation experiments, using classifier-free guidance with proper guidance scale heavily affects the alignment of motions to their descriptions.

4 Experiments

4.1 Datasets and Training

We train and evaluate our text-to-motion synthesis model on two popular motion datasets in this domain.

KIT Motion-Language. The dataset contains 3,911 3D Human motions with 6,278 text descriptions with 1-4 text descriptions for each motion [24]. Although the quantity and diversity of this dataset is relatively small, it has been widely used for previous works in text-to-motion research. We follow the 251 motion features [9] representation in the experiments.

HumanML3D. The dataset contains 14,616 3D human motions and 44,970 text descriptions [9]. It was created by re-annotating motion capture from AMASS [19] and HumanAct12 [8] collections. The 263 motion feature data representation additionally contains root velocity, joint positions, joint velocities, joint rotations and foot contact binary labels.

4.2 Baseline Methods

We compare our model to four state-of-the-art methods: Seq2Seq [17], Language2Pose [1], TM2T [10] and Motion Diffusion Model (MDM) [32]. Seq2seq and Language2Pose are deterministic motion generation baselines. TM2T utilizes VQ-VAE and recurrent models for text-to-motion synthesis task. MDM uses a conditional diffusion model on raw motions that showed promising motion results.

4.3 Quantitative Evaluations

We present the quantitative evaluation results on HumanML3D and KIT-ML over the set of metrics used in recent works [9, 10, 32]. *R-precision* and *Multimodal Distance* represents how aligned the motions are to their respective text conditions. *FID* calculates the distance between ground truth and predicted motion distributions. Lower *FID* correlates to higher quality motions but does not necessarily imply that the synthesized motions will align well with input text conditions. Finally, *Diversity* measures the variety in the predicted motion distributions given the same input condition.

As shown from the results in Table 1 for HumanML3D dataset, our MoDDM method produces good *FID* scores and outperforms state-of-the-art results by a large margin. While the motion diffusion model (MDM) showed a significant leap in the generated quality of motions compared to the VQ-VAE model with recurrent network (TM2T), the approach struggles in properly aligning the motions to the text condition, which is demonstrated by lower *R-precision* and *Multimodal Distance* compared to our method and TM2T. Our *R-precision* performance is similar to TM2T while our method produces much better *FID*.

On the KIT-ML dataset, Table 2 also shows MoDDM is able to capture the alignment between motions and text better than other methods as evidenced by the superior *R-precision* and *Multimodal Distance* results. For motion quality, our method also outperforms TM2T by a large margin in *FID* score and is slightly behind MDM. These results validate our proposed method that utilizes the discrete diffusion model for text-to-motion synthesis task.

We also show that our method does not only produce stronger results in motion-text alignment than MDM, but also requires only one-tenth of steps during inference. Compared to MDM, which applies diffusion model on raw motions, MoDDM achieves approximately **5x** speedup during inference with the classifier-free guidance and **10x** faster without it. Therefore, our method is capable of generating longer motion sequences that are well-aligned with the text description while using less computational resources.

Methods	R Precision ↑			FID ↓	MultiModal Dist ↓	Diversity →
	Top 1	Top 2	Top 3			
Real Motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
Seq2Seq [17]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$
Language2Pose [1]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$
MDM [32]	-	-	$0.611^{\pm.007}$	$\underline{0.544}^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$
TM2T [10]	$\underline{0.424}^{\pm.003}$	$\mathbf{0.618}^{\pm.003}$	$\mathbf{0.729}^{\pm.002}$	$1.501^{\pm.017}$	$\mathbf{3.467}^{\pm.011}$	$8.589^{\pm.076}$
MoDDM (Ours)	$\mathbf{0.425}^{\pm.004}$	$\underline{0.615}^{\pm.004}$	$\underline{0.713}^{\pm.003}$	$\mathbf{0.294}^{\pm.006}$	$\underline{3.553}^{\pm.009}$	$\underline{9.178}^{\pm.093}$

Table 1: Quantitative evaluation on the HumanML3D test set. $\pm$ indicates 95% confidence interval, and $\rightarrow$ means the closer to Real motions the better. Bold face indicates the best result, while underscore refers to the second best.

Methods	R Precision ↑			FID ↓	MultiModal Dist ↓	Diversity →
	Top 1	Top 2	Top 3			
Real Motions	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$
Seq2Seq [21]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$
Language2Pose [1]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$
MDM [33]	-	-	$0.396^{\pm.004}$	$\mathbf{0.497}^{\pm.021}$	$9.121^{\pm.022}$	$\mathbf{10.847}^{\pm.109}$
TM2T [11]	$\underline{0.280}^{\pm.005}$	$\underline{0.463}^{\pm.006}$	$\underline{0.587}^{\pm.005}$	$3.599^{\pm.153}$	$\underline{4.591}^{\pm0.26}$	$9.473^{\pm.117}$
MoDDM (Ours)	$\mathbf{0.307}^{\pm.002}$	$\mathbf{0.490}^{\pm.002}$	$\mathbf{0.630}^{\pm.003}$	$\underline{1.194}^{\pm.004}$	$\mathbf{3.860}^{\pm.009}$	$\underline{10.346}^{\pm.112}$

Table 2: Quantitative evaluation on the KIT-ML test set. ± indicates 95% confidence interval, and → means closer to Real motions the better. Bold face, underscore indicates best, second-best respectively.

4.4 Subjective Evaluations

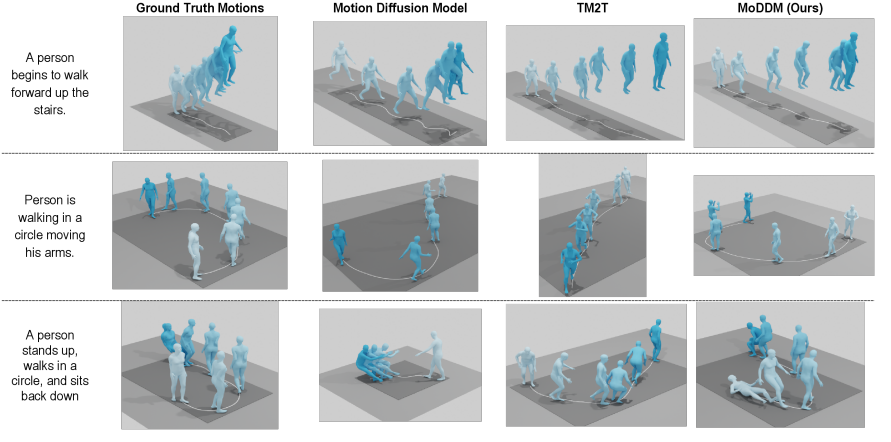

Figure 2: Qualitative Evaluations on 3 HumanML3D test samples between ground truth motion, MDM, TM2T and our Motion Discrete Diffusion Model (MoDDM) with text description in the leftmost column. Light to dark shade indicates the motion duration from start to finish.

We show a qualitative comparison between MDM, TM2T and our MoDDM on HumanML3D test dataset in Figure 2. These comparison results show the improvements of our model in motion quality and alignment to text conditions. In the first example, MDM shows downward movements before moving up, which indicates weaker condition alignment. In the second example, our model is able to produce higher quality motions by generating a circular trajectory with arm movements. In the third example, our result shows better diversity, where the person stands up from a supine position, which is not clearly defined in the description but plausible for the given condition.

We also conducted a subjective evaluation to assess the quality of the synthesized motions from our method, MDM, TM2T, and ground truth. We randomly selected 10 test sequences and generated video clips for each method. The videos from each sequence were then shown to 40 participants to compare and rank different methods. When considering the ranking, our method is the most preferred one (34.6%) behind ground truth (39.2%), followed by MDM (20.8%) and TM2T (5.4%). More details about the user study will be provided in the supplementary material.

4.5 Ablation Study

All ablation experiments are performed on HumanML3D test dataset.

Classifier free guidance: We observe that inclusion of classifier-free guidance is vital to get a good alignment between motion and text. This is indicated by the significant increase in *R-precision* and *Multimodal Distance* with classifier-free guidance in Table 3.

Guidance scale: We investigate the effect of guidance scale s on evaluation. Figure 3 shows the variance in *R-precision* and *FID* given different values of s. Increasing the guidance scale improves motion quality but excessive weight degrade the results. Similar trend is observed for alignment of motion to text. Our experiments show that a guidance scale of around 4 gives the best performance overall.

Diffusion steps: We explore the influence of diffusion steps on our evaluation metrics. Increasing the number of steps beyond 100 hurts the performance which corroborates with experimental results in previous work [7].

Methods	R Precision ↑			FID ↓
	Top 1	Top 2	Top 3	
w/o classifier-free guidance	0.303	0.454	0.558	0.297
Diffusion step = 200	0.393	0.575	0.682	0.415
Diffusion step = 100	**0.425**	**0.615**	**0.713**	**0.294**

Table 3: Quantitative ablation evaluation on the HumanML3D test set. Every value is averaged over 5 evaluation runs. Classifier-free guidance is used with guidance scale 4 unless specified otherwise.

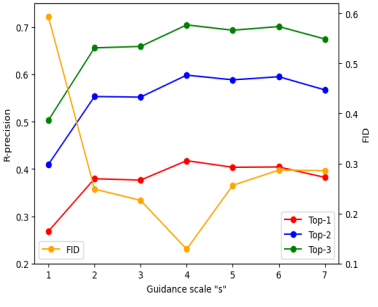

Figure 3: Ablation study on Classifier-free guidance scale. All experiments are done with batch size of 128 for 750 epochs. For each value, evaluation has been repeated 5 times and averaged over the runs.

5 Conclusion

We present MoDDM, a method for text-to-motion synthesis based on discrete diffusion models. The method utilizes vector quantization (VQ) to learn discrete motion tokens from the input motions and then trains a denoising model over the discrete token distributions. Compared to the previous diffusion methods that are applied on raw motion sequences directly [32], our method produces high quality motion results while greatly reducing the computational costs. We evaluated our approach through objective and subjective metrics, and the results demonstrated that our method produce superior results for text-to-motion synthesis task in both motion fidelity and text-to-motion matching accuracy.

References

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

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[3] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv preprint arXiv:2212.04495*, 2022.

[4] David Dehaene and Rémy Brossard. Re-parameterizing vaes for stability. *arXiv preprint arXiv:2106.13739*, 2021.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[6] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Text-based motion synthesis with a hierarchical two-stream rnn. In *ACM SIGGRAPH 2021 Posters*, pages 1–2. 2021.

[7] 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*, pages 10696–10706, 2022.

[8] 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 *Proc ACM Int. Conf. on Multimedia*, pages 2021–2029, 2020.

[9] 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 (CVPR)*, pages 5152–5161, June 2022.

[10] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 580–597. Springer, 2022.

[11] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forr'e, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *ArXiv*, abs/2102.05379, 2021.

[15] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis editing. *arXiv preprint arXiv:2209.00349*, 2022.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Angela S. Lin, Lemeng Wu, and Qixing Huang Raymond J. Mooney Rodolfo Corona, Kevin Tai. Generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS 2018*, December 2018. URL http://www.cs.utexas.edu/users/ai-labpub-view.php?PubID=127730.

[18] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018.

[19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019.

[20] 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.

[21] 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, 2022.

[22] 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*, pages 10985–10995, 2021.

[23] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

[24] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*.

[25] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018.

[26] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. pages 14866–14876, 2019.

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[31] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022.

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

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

[34] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[35] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[36] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 465–481. Springer, 2020.

[37] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, pages 231–242. Springer, 2023.

[38] 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, 2023.

[39] 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.