# Priority-Centric Human Motion Generation in Discrete Latent Space

Hanyang Kong[1], Kehong Gong[1], Dongze Lian[1], Michael Bi Mi[2], Xinchao Wang[1*]
[1]National University of Singapore [2]Huawei International Pte Ltd
{hanyang.k, gongkehong}@u.nus.edu {dongze, xinchao}@nus.edu.sg

## Abstract

*Text-to-motion generation is a formidable task, aiming to produce human motions that align with the input text while also adhering to human capabilities and physical laws. While there have been advancements in diffusion models, their application in discrete spaces remains underexplored. Current methods often overlook the varying significance of different motions, treating them uniformly. It is essential to recognize that not all motions hold the same relevance to a particular textual description. Some motions, being more salient and informative, should be given precedence during generation. In response, we introduce a Priority-Centric Motion Discrete Diffusion Model (M2DM), which utilizes a Transformer-based VQ-VAE to derive a concise, discrete motion representation, incorporating a global self-attention mechanism and a regularization term to counteract code collapse. We also present a motion discrete diffusion model that employs an innovative noise schedule, determined by the significance of each motion token within the entire motion sequence. This approach retains the most salient motions during the reverse diffusion process, leading to more semantically rich and varied motions. Additionally, we formulate two strategies to gauge the importance of motion tokens, drawing from both textual and visual indicators. Comprehensive experiments on the HumanML3D and KIT-ML datasets confirm that our model surpasses existing techniques in fidelity and diversity, particularly for intricate textual descriptions.*

## 1. Introduction

Generating realistic and diverse 3D human motions under various conditions *e.g.*, action labels [34, 18], natural language descriptions [55, 17, 5, 56], and musical cues [27, 28, 29], presents a significant challenge across multiple domains, including gaming, filmmaking, and robotic animation. Notably, motion generation based on language descriptions has garnered substantial interest, given its promise for enhancing realism and broadening practical applications.

However, generating a high-quality motion is not trivial due to the inherent modality gap and the complex mapping between text and motion modalities. Previous works [35, 43] align the latent feature space between text and motion modalities. For instance, TEMOS [35] aligns the text and motion feature by learning text and motion encoders. MotionClip [43] aligns the human motion manifold to CLIP [37] space for infusing semantic knowledge of CLIP into the motion extractor. Despite these advancements, they might encounter performance degradation when dealing with complex textual descriptions.

To process these complex textual descriptions, some text-to-motion methods are proposed. State-of-the-art approaches such as TM2T [17] and T2M-GPT [55] use vector-quantized autoencoder (VQ-VAE) [46] to learn a discrete and compact motion representation, followed by a translation model [38, 47] to map the text modality to the motion modality. With the popularity and the superior performance in the generation tasks of diffusion models [20], MDM [44] and MLD [5] are proposed to learn conditioned diffusion models on the raw motion representation space and the latent feature space, respectively.

While there have been promising advancements, two primary issues remain unresolved: i) The aforementioned diffusion methods, namely MDM [44] and MLD [5], predominantly address the latent feature within a continuous space. Although VQ-VAE-inspired architectures have made considerable strides in motion generation [55, 17], particularly with the support of discrete and compact motion representations, the integration of the diffusion model into a discrete space remains inadequately explored. ii) Discrete diffusion models employed in prior studies [15, 14, 22] treat all tokens uniformly. This approach presupposes that every token conveys an equivalent amount of information, neglecting the inherent disparities among tokens within a sequence. A more intuitive generative approach for humans would involve a progressive hierarchy, commencing with overarching concepts and gradually delving into finer details.

To address the aforementioned challenges, we introduce a priority-centric motion discrete diffusion model (M2DM)

*Corresponding author: xinchao@nus.edu.sg

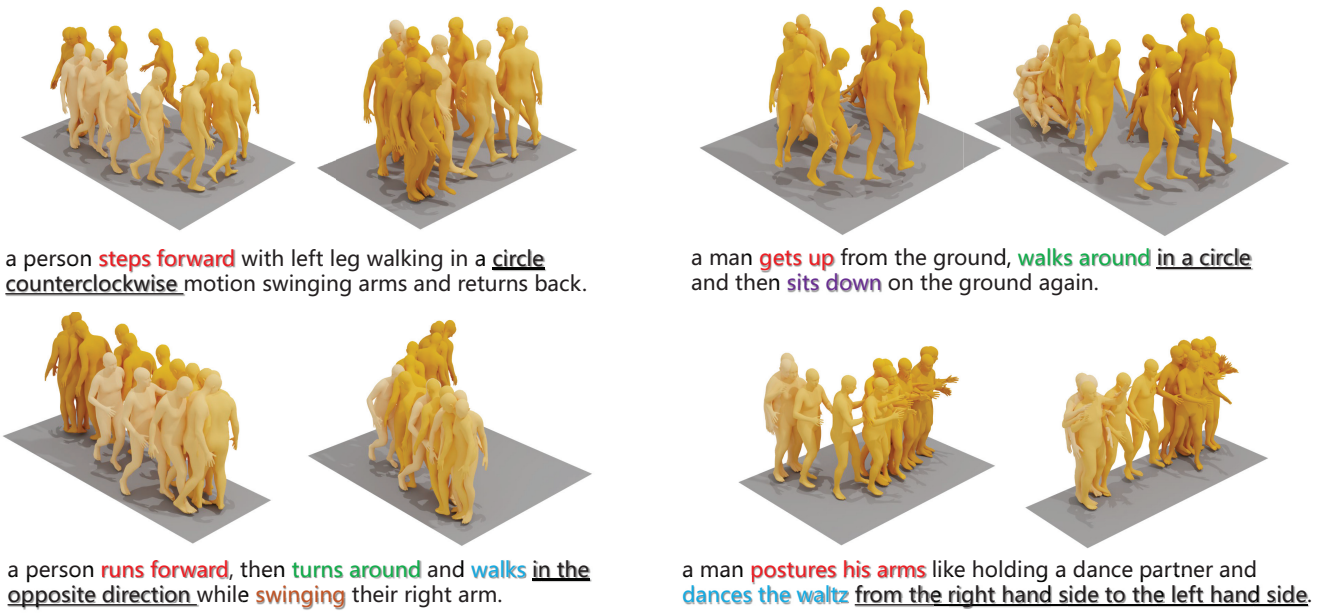



Figure 1: Our Priority-Centric Motion Discrete Diffusion Model (M2DM) generates diverse and precise human motions given complex textual descriptions. The color of human mesh goes from light to dark over time.

designed to generate motion sequences from textual descriptions, progressing in a primary to secondary manner. Initially, we employ a Transformer-based VQ-VAE, which is adept at learning a concise, discrete motion representation through the global self-attention mechanism. To circumvent code collapse and guarantee the optimal utilization of each motion token within the codebook, a regularization term is incorporated during the VQ-VAE training phase. Furthermore, we craft a noise schedule wherein individual tokens within a sequence are allocated varying corruption probabilities, contingent on their respective priorities during the forward process. Specifically, tokens of higher priority are slated for corruption towards the latter stages of the forward process. This ensures that the subsequent learnable reverse process adheres to a primary-to-secondary sequence, with top-priority tokens being reinstated foremost.

To discern the significance of motion tokens within a sequence, we introduce two evaluative strategies: static assessment and dynamic assessment. 1) Static assessment: Drawing inspiration from the neural language domain, where the significance of individual words is gauged by their entropy across datasets, we calculate the entropy of each motion token across the entire motion dataset. 2) Dynamic assessment: We cultivate an agent specifically to dynamically gauge the importance of tokens within a sequence. Given a discrete motion token sequence, the agent masks one token at each iteration. These masked token sequences are then fed into the VQ decoder for continuous motion reconstruction. The agent's objective is to curtail the cumulative reconstruction discrepancy between the original and the reconstructed motions at every phase. This is achieved using a reinforcement learning [23] (RL) strategy, facilitating the agent's identification of motion tokens that convey minimal information within a sequence—a process we term dynamic analysis. Our priority-centric discrete diffusion model has showcased commendable generative prowess, especially when dealing with intricate textual descriptions. Through rigorous testing, our method has proven its mettle, consistently matching or surpassing the performance of prevailing text-to-motion generation techniques on the HumanML3D and KIT-ML datasets.

We summarize our contributions as follows:

- To capture long-range dependencies in the motion sequences, we apply Transformer as the architecture of VQ-VAE. Besides, a regularization term is applied to increase the usage of tokens in the codebook.
- We design a priority-centric scheme for human motion generation in the discrete latent space. To enhance the performance given complex descriptions, we design a novel priority-centric scheme for the discrete diffusion model.
- Our proposed priority-centric M2DM achieves state-of-the-art performance on the HumanML3D [16] and KIT-ML [36] datasets.

## 2. Related Work

**VQ-VAE.** Vector Quantized Variational Autoencoder (VQ-VAE) [46] aims to minimize reconstruction error with dis-

crete representations by training a variational autoencoder. VQ-VAE achieves promising performance on generative tasks, such as text-to-image generation [39, 52], music generation [7], unconditional image synthesis [9], etc. For motion generation tasks, most of the methods [17, 55] tokenize the continuous motion representation by the 1D-CNN network. Since the motion style varies from person to person and the number of action combinations is almost infinite, it is essential to extract discriminative and distinctive motion features from the raw motion representations. Moreover, raw motion representation is much more redundant compared with neural language, it is feasible to represent a motion sequence using few motion tokens from a global perspective. To this end, we introduce a transformer-based VQ-VAE and several techniques are applied to balance the usage of each motion token in the codebook and prevent the codebook from collapsing, *i.e.*, only a small proportion of codes in the codebook are activated.

**Diffusion models.** Diffusion generative models are first proposed in [42] and achieve promising results on image generation [8, 33, 21, 20, 50, 48, 53, 11, 10]. Discrete diffusion model is first proposed by [42] and several works [22, 14] apply the discrete diffusion model for text generation. D3PM [3] then applies the discrete diffusion model to the image generation task. For motion generation tasks, several works [24, 56, 44, 5] introduce diffusion models to generate diverse motion sequences from textual descriptions. In this work, we first introduce the discrete diffusion model for text-to-motion generation tasks which achieves promising results, especially for long text descriptions with several action commands.

**Conditional human motion synthesis** Compared with unconditional Human motion synthesis, human motion synthesis conditioned on various conditions, such as textual description [55, 17, 5, 56], music [27, 28, 29, 13], and actions [34, 18] is more challenging due to the huge gap between two different domains. A common strategy of conditioned human motion synthesis is to employ generative models, for instance, conditional VAE [26], and learn a latent motion representation. Text-to-motion task is more challenging because neural language is a type of information highly refined by human beings, whereas human motion is much more redundant. Most recently, several promising works [5, 16, 17, 56, 55] are proposed for text-to-motion tasks. T2M-GPT [55] and TM2T [17] regard the text-to-motion task as a translation problem. They first train a VQ-VAE [46] to tokenize the motion sequences into motion tokens, and Recurrent Neural Networks (RNN) [32], Transformer [47, 51, 49, 41, 40, 54], or GPT-like [38] models are further applied to 'translate' textual descriptions to motions. Moreover, diffusion-based [20] models are introduced for text-to-motion generation by [56, 6, 44, 5] in the contentious space. In this work, to the best of our knowledge, we are the first ones to diffuse the motion representation in the discrete latent space. Moreover, we notice that different motion clips in a whole motion sequence play various roles, for instance, stepping forward is more important than swinging arms for a running motion. We further design a diffuse schedule such that the generation process follows a from-primary-to-secondary manner.

# 3. Method

## 3.1. Motion Tokenizer

The motion tokenizer aims to learn a codebook that can represent a wide range of motion sequences with the priority awareness of motion tokens. To achieve this, we propose a transformer-based VQ-VAE that leverages self-attention to capture long-range dependencies between motion frames and learns a stable and diverse codebook.

Fig. 2 shows the overview of our motion tokenizer. Given a motion sequence $X \in \mathbb{R}^{T\times d}$ with $T$ frames and $d$ dimensional motion representation, we reconstruct the motion sequence using a transformer autoencoder and learn an informative codebook $Z = \{z_k\}_{k=1}^{K}$ with $K$ entries. We use the transformer encoder $E$ to compute the latent motion features $\boldsymbol{b} = \{b_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$, where $d$ is the dimension of the latent features. Then for each feature $b_t$, we query the codebook $Z$ by measuring the distance between feature $b_t$ and each entry in the codebook $Z$. To improve the codebook usage and avoid dead codes, we first normalize $b_t$ and $z_t$ and then measure the Euclidean distance between two normalized vectors in the unit sphere coordinate, which is formulated as:

$$\hat{b}_t = \underset{b_t\in\boldsymbol{b}}{\arg\min} \|\ell_2(b_t) - \ell_2(z_k)\|_2^2, \tag{1}$$

where $\ell_2$ is the $l_2$ normalized function.

The standard optimization goal consists of three parts: the reconstruction loss, the embedding loss, and the commitment loss, which is formulated as

$$\begin{aligned}\mathcal{L}_{vq} = \left\|X - \tilde{X}\right\|_2^2 + ||Z - sg[\hat{Z}]||_2^2 \\ + \eta||sg[Z] - \hat{Z}||_2^2.\end{aligned} \tag{2}$$

Moreover, to achieve an informative codebook with diverse and high usage of its entries, we apply an orthogonal regularization. Specifically, we calculate the distance between each pair of codebook entries to enforce orthogonality among all the codebook entries:

$$\mathcal{L}_{orth}(Z) = \left\|\ell_2(Z)^\top \ell_2(Z) - I_K\right\|_2^2, \tag{3}$$

where $\mathcal{L}_2$ is the $l_2$ normalized function.

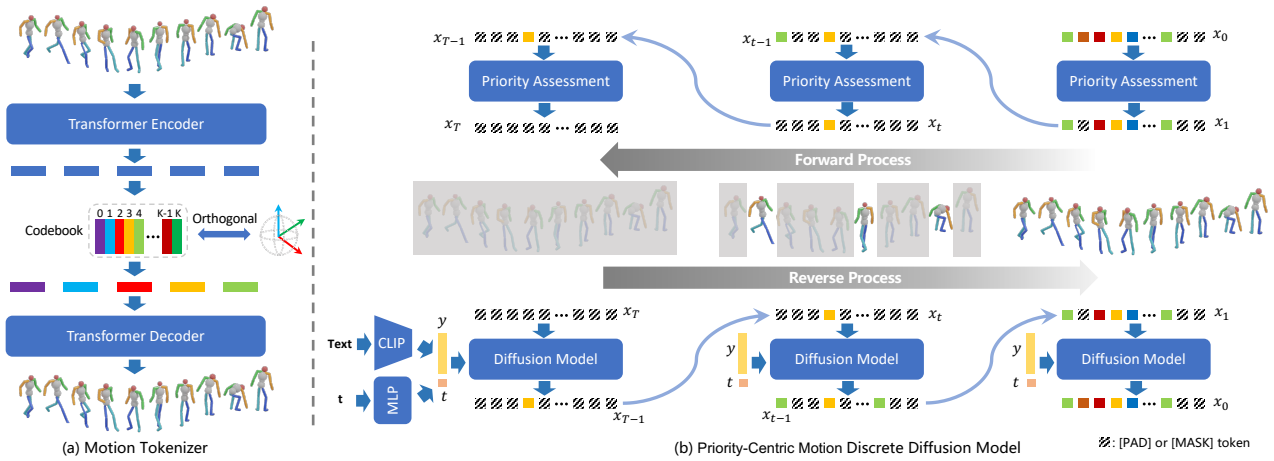



Figure 2: **Overview of our framework for our Motion Discrete Diffusion Model (M2DM).** M2DM consists of a Transformer-based VQ-VAE (Sec. 3.1) and a discrete diffusion model (Sec. 3.2). Firstly, the Transformer-based motion tokenizer is learned to discretize motion sequences into tokens and reconstruct original motion sequences from tokens queried from the learnable codebook. Then the discrete diffusion model is learned to generate diverse motions conditioned on the textual descriptions. Two priority assessment strategies (Sec. 3.3) are applied to mask or replace motion tokens with less information at the beginning of the forward diffusion process and the most important motion tokens are corrupted in the last few diffusion steps. The diffusion model learns to restore the original motion token index conditioned on the given textual descriptions.

Therefore, the overall training objective of vector quantization is defined as follows:

$$\begin{aligned}\mathcal{L}_{vq} = \left\|X - \tilde{X}\right\|_2^2 + ||Z - sg[\hat{Z}]||_2^2 \\ + \eta||sg[Z] - \hat{Z}||_2^2 + \delta\mathcal{L}_{orth}(Z),\end{aligned} \tag{4}$$

where $\tilde{X} = D(\boldsymbol{b})$, $D$ is the decoder, $sg[\cdot]$ denotes the stop-gradient operation, $\eta$ and $\delta$ are the weighting scalars.

### 3.2. Discrete Diffusion Models

Generally speaking, the forward discrete diffusion process corrupts each element $x_t$ at timestep $t$ with $K$ categories via fixed Markov chain $q(x_t|x_{t-1})$, for instance, replace some tokens of $x_{t-1}$ or mask the token directly. After a fixed number of $T$ timesteps, the forward process yields a sequence of increasingly noisy discrete tokens $z_1, ..., z_T$ of the same dimension of $z_0$ and $z_T$ becomes pure noise token. For the denoising procedure, the real $x_0$ is restored sequentially based on the pure noisy token $z_T$. To be specific, for scalar random variable with $K$ categories $x_t, x_{t-1} \in 1, ...K$, the forward transition probabilities can be represented by the transition matrices $[\boldsymbol{Q}_t]_{ij} = q(x_t = i|x_{t-1} = j) \in \mathbb{R}^{(K+1)\times(K+1)}$. $[\boldsymbol{Q}_t]_{i,j}$ is the Markov transition matrix from state $t$ to state $t-1$ that is applied to each token in the sequence independently, which can be written as:

$$\boldsymbol{Q}_t = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \beta_t & \dots & 0 \\ \beta_t & \alpha_t + \beta_t & \beta_t & \dots & 0 \\ \beta_t & \beta_t & \alpha_t + \beta_t & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma_t & \gamma_t & \gamma_t & \dots & 1 \end{bmatrix}, \tag{5}$$

where $\alpha_t \in [0, 1]$ is the probability of retaining the token, $\gamma_t$ is the probability of replacing the original token to [MASK] token, leaving the probability $\beta_t = (1 - \alpha_t - \gamma_t)/K$ to be diffused. The forward Markov diffusion process for the whole toke sequence is written as,

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \text{Cat}(\boldsymbol{x}; \boldsymbol{p} = \boldsymbol{x}_{t-1}\boldsymbol{Q}_t) = \boldsymbol{x}_t\boldsymbol{Q}_t\boldsymbol{x}_{t-1}, \tag{6}$$

where $\boldsymbol{x}$ is the one-hot row vector identifying the token index and $\text{Cat}(\boldsymbol{x}; \boldsymbol{p})$ is a categorical distribution over the one-hot vector $\boldsymbol{x}_t$ with probabilities $\boldsymbol{p}$.

Starting from the initial step $\boldsymbol{x}_0$, the posterior diffusion process is formulated as:

$$\begin{aligned} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\ &= \frac{(\boldsymbol{x}^\top\boldsymbol{Q}_t\boldsymbol{x}_{t-1})(\boldsymbol{x}_{t-1}^T\bar{\boldsymbol{Q}}_{t-1}\boldsymbol{x}_0)}{\boldsymbol{x}_t^\top\bar{\boldsymbol{Q}}_t\boldsymbol{x}_0} \end{aligned} \tag{7}$$

with $\bar{\boldsymbol{Q}} = \boldsymbol{Q}_1\boldsymbol{Q}_2 \dots \boldsymbol{Q}_t$.

Recording to Markov chain, the intermediate step can be marginalized out and the derivation of the probability of $\boldsymbol{x}_t$

at arbitrary timestep directly from $\boldsymbol{x}_0$ can be formulated by

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \boldsymbol{x}_t^\top \bar{\boldsymbol{Q}} \boldsymbol{x}_0. \tag{8}$$

The cumulative transition matrix $\bar{\boldsymbol{Q}}$ and the probability $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ are then computed as

$$\bar{\boldsymbol{Q}}\boldsymbol{x}_0 = \bar{\alpha}_t \boldsymbol{x}_0 + (\bar{\gamma} - \bar{\beta}_t)\boldsymbol{m} + \bar{\beta}_t \tag{9}$$

where $\boldsymbol{m}$ is the one-hot vector for [MASK] and [PAD], $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\bar{\gamma}_t = 1 - \prod_{i=1}^{t}(1-\gamma_i)$, and $\bar{\beta}_t = (1-\bar{\alpha}_t - \bar{\gamma}_t)/K$.

Regarding the reverse diffusion process, we train a transformer-like [47] denoising network $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})$ to estimate the posterior transition distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$. The network is trained to minimize the variational lower bound (VLB) [42]:

$$\begin{aligned}\mathcal{L}_{VLB} =& \mathbb{E}_q[D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_T))] \\ &+ \mathbb{E}_q[\sum_{t=2}^{T} D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, t))] \\ &- \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\end{aligned} \tag{10}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation over the joint distribution $q(\boldsymbol{x}_{0:T})$.

### 3.3. Priority-Centric Denoising Process

The designation of the noise schedule in the continuous domain, for instance, the linear schedule [20] and the cosine schedule [33], achieves the excellent performance of diffusion models. The noise can be easily controlled by the variation of Gaussian noise. For the discrete diffusion models, several noise schedules [22, 3] have been explored to control the data corruption and denoise the reverse process by choosing the transition matrix $\boldsymbol{Q}_t$. *For text-to-motion generation, however, such a schedule assumes all motion tokens carry the same amount of information and do not consider the difference among the discrete tokens extracted from the continuous motion sequence.* Intuitively, the motion generation process would be in a progressive manner, where the most important motion tokens should appear in earlier steps for laying a solid foundation when denoising the noise in the discrete domain.

Specifically, we apply the mask-and-replace strategy in Eq. (5) to corrupt each ordinary discrete token with a probability of $\gamma_t$ for masking [MASK] token and a probability of $\beta_t$ for uniform diffusion to a random token $\boldsymbol{x}_k$, resulting in a remaining probability of $\alpha_t = 1 - K\beta_t - \gamma_t$ for retention. When corrupting across the forward process in the discrete domain, we expect that $\bar{\alpha}_t^i < \bar{\alpha}_t^j$ if token $\boldsymbol{x}_i$ is more informative than $\boldsymbol{x}_j$ such that the tokens with high information emerge earlier than less informative motion tokens in the reverse process. Instead of using a normal corruption strategy with linearly increasing $\bar{\gamma}_t$ and $\bar{\beta}_t$, we propose a priority-score function $F$ to assess the relative importance of information for a motion token sequence, which enables the recovery of the most informative tokens during the reverse process.

Given an original motion token sequences $\boldsymbol{x}_0$ at timestep 0 with length $N$, we aim to analyze the importance of each token $x_0^i$ in $\boldsymbol{x}_0$. Based on the information entropy theory, we calculate the information entropy for each motion token $x_0^i$ as the importance score

$$F(x_0^i) = \frac{NH(x_0^i)}{\sum_{j=1}^{N} H(x_0^j)}, \tag{11}$$

where

$$H(x) = -\sum_{i=1}^{K} p(x_i) \log p(x_i), \tag{12}$$

and $K$ denotes the number of categories belonging to $x_i$.

After obtaining the importance score function and setting linearly increasing schedule $\bar{\gamma}_t^i$ and $\bar{\beta}_t^i$ for all tokens, $\bar{\gamma}_t^i$ and $\bar{\beta}_t^i$ for specific token $x_0^i$ can be further updated by

$$\bar{\gamma}_t^i \Leftarrow \bar{\gamma}_t^i \cdot \sin\frac{t\pi}{T} \cdot F(x_0^i), \tag{13}$$

and

$$\bar{\beta}^i \Leftarrow \bar{\beta}^i \cdot \sin\frac{t\pi}{T} \cdot F(x_0^i), \tag{14}$$

where $\boldsymbol{x}_0^i$ is the $i$-th token at timestep 0.

Now we introduce two solutions for obtaining $p(x_i)$.

**Static assessment.** Analogously to computing the entropy for each word in natural language processing, we estimate the probability $p(x_i)$ for each motion token by counting its frequency in the quantized motion sequences over the entire dataset.

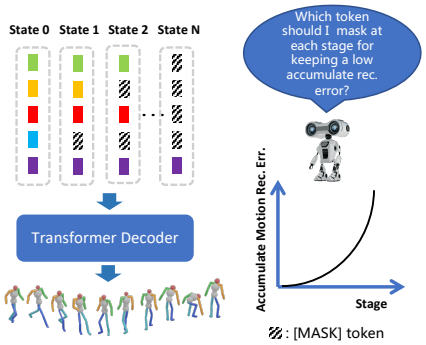



Figure 3: Illustration for dynamic assessment strategy. The gray blocks are masked tokens and other colorful blocks are motion tokens.

**Dynamic assessment.** Rather than counting motion token frequency and calculating the importance of each token, we further propose to learn an agent to estimate the importance of each token in the token sequences by reinforcement learning strategy. As shown in Fig. 3, given a series of complete motion tokens $\boldsymbol{x} = \{x_n\}_{n=1}^{N}$ that are queried from the codebook at current state $\boldsymbol{s}_t \in \mathcal{S}$, the scorer samples the motion tokens $\tilde{\boldsymbol{x}} = \{x_n\}_{n=1}^{N'}$ and tries to minimize the reconstruction error between the original motion sequences $\boldsymbol{m}$ and the motion sequences decoded from the sampled tokens $\tilde{\boldsymbol{m}}$.

State includes the currently selected motion tokens $\tilde{\boldsymbol{x}} = \{x_n\}_{n=1}^{N'}$, the complete motion tokens $\boldsymbol{x} = \{x_n\}_{n=1}^{N}$, and the corresponding reconstructed continuous motions $\tilde{\boldsymbol{m}}$ and $\boldsymbol{m}$. The reconstructed continuous motion sequences $\tilde{\boldsymbol{m}}$ and $\boldsymbol{m}$ are obtained by decoding the motion tokens by VQ decoder. The action is to sample one motion token from the complete motion token sequence at each stage. To be specific, given a motion token sequence, the agent $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$ tries to sample a motion token that can recover the continuous motion sequences with the minimum reconstruction error. The reward measures the reconstruction difference between $\tilde{\boldsymbol{m}}$ and $\boldsymbol{m}$ with the minimum sample times.

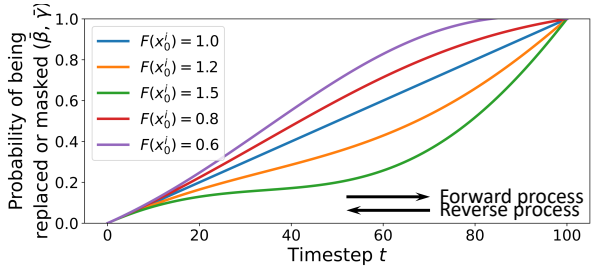



Figure 4: Priority-centric denoising schedule for motion tokens.

After the agent is well-trained, we calculate and restore the sampling order for each motion token sequence by the agent $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$ and compute the noise schedule for each motion token in the codebook by Eq. (13) and (14). Fig. 4 shows the noise schedule for different motion tokens *w.r.t.* the corresponding priorities. As shown in Fig. 4, high-priority tokens (orange and green curves) are encouraged to be corrupted at the end of the forward process such that the learnable reverse process follows a primary-to-secondary manner and high-priority tokens will be restored first. It is worth noting that both strategies are applied to design noise schedules for each token, which can be calculated and stored in advance such that the agent will not join the forward and reverse diffusion processes. Thus, there is no extra computational cost for the diffusion model during the training and inference stages.

# 4. Experiments

## 4.1. Datasets and Evaluation Metrics

Our experiments are conducted on two standard text-to-motion datasets for text-to-motion generation: HumanML3D [16] and KIT-ML [36]. We follow the evaluation metrics provided in [16].

**HumanML3D.** HumanML3D [16] is currently the largest 3D human motion dataset that covers a broad range of daily human actions, for instance, exercising and dancing. The dataset contains 14,616 motion sequences and 44970 text descriptions. The motion sequences are processed to 20 frame-per-second (FPS) and cropped to 10 seconds if the motion sequences are longer than 10 seconds, resulting in duration ranges from 2 to 10 seconds. For each motion clip, the corresponding number of descriptions is at least three.

**KIT Motion-Language (KIT-ML).** KIT-ML dataset [36] contains 3,911 3D human motion clips with 6,278 text descriptions. For each motion clip, one to four textual descriptions are provided. The motion sequences are collected from the KIT dataset [31] and the CMU dataset [1] with the down-sampled 12.5 FPS.

Both of the aforementioned datasets are split into training, validation, and testing datasets with the proportions of 80%, 5%, and 15%.

**Implementation details.** Both our VQ-encoder and decoder consist of 4 transformer layers with 8 heads and the dimension is 512. For the HumanML3D [16] and KIT-ML [36] datasets, the motion sequences are cropped to 64 frames for training. We apply Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay is 1e-4. The learning rate is 1e-4 after warmup linearly for 5 epochs. The number of tokens in the codebook is 8192. We train our VQ-VAE with a batch size of 1024 across 8 Tesla T4 for a total 100 training epochs for 8 hours.

For the discrete diffusion model, we set timesteps $T = 100$ and the network is trained using Adam [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate reaches 2e-4 after 2000 iterations of a linear warmup schedule. The transformer-based discrete diffusion model consists of 12 transformer layers with 8 heads, and the dimension is 512. The text features are extracted by pre-trained CLIP [37] model. The diffusion model is trained with a batch size of 64 across 8 Tesla T4 for 100k iterations for 36 hours.

**Evaluation metrics.** Following the evaluation protocols provided in previous works [16, 55, 44, 5], we evaluate the correspondences between motion and language using deep multimodal features with the pre-trained models in [16] by the following metrics: 1) Generation diversity: We randomly

| Methods | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → | MModality ↑ |
|---|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | | |
| **Real motion** | $0.511^{\pm.003}$ | $0.703^{\pm.003}$ | $0.797^{\pm.002}$ | $0.002^{\pm.000}$ | $2.974^{\pm.008}$ | $9.503^{\pm.065}$ | - |
| Our VQ-VAE | $0.508^{\pm.002}$ | $0.691^{\pm.002}$ | $0.791^{\pm.003}$ | $0.63^{\pm.001}$ | $3.015^{\pm.010}$ | $9.577^{\pm.081}$ | - |
| Our VQ-VAE *w/o* $\ell_2$ | $0.494^{\pm.003}$ | $0.685^{\pm.002}$ | $0.772^{\pm.003}$ | $0.70^{\pm.001}$ | $3.251^{\pm.008}$ | $9.525^{\pm.092}$ | - |
| Our VQ-VAE *w/o* $\ell_2$, $\mathcal{L}_{\text{orth}}$ | $0.485^{\pm.002}$ | $0.671^{\pm.003}$ | $0.752^{\pm.003}$ | $0.79^{\pm.001}$ | $3.378^{\pm.012}$ | $9.511^{\pm.095}$ | - |
| Seq2Seq [30] | $0.180^{\pm.002}$ | $0.300^{\pm.002}$ | $0.396^{\pm.002}$ | $11.75^{\pm.035}$ | $5.529^{\pm.007}$ | $6.223^{\pm.061}$ | - |
| Language2Pose [2] | $0.246^{\pm.002}$ | $0.387^{\pm.002}$ | $0.486^{\pm.002}$ | $11.02^{\pm.046}$ | $5.296^{\pm.008}$ | $7.676^{\pm.058}$ | - |
| Text2Gesture [4] | $0.165^{\pm.001}$ | $0.267^{\pm.002}$ | $0.345^{\pm.002}$ | $5.012^{\pm.030}$ | $6.030^{\pm.008}$ | $6.409^{\pm.071}$ | - |
| Hier [12] | $0.301^{\pm.002}$ | $0.425^{\pm.002}$ | $0.552^{\pm.004}$ | $6.532^{\pm.024}$ | $5.012^{\pm.018}$ | $8.332^{\pm.042}$ | - |
| MoCoGAN [45] | $0.037^{\pm.000}$ | $0.072^{\pm.001}$ | $0.106^{\pm.001}$ | $94.41^{\pm.021}$ | $9.643^{\pm.006}$ | $0.462^{\pm.008}$ | $0.019^{\pm.000}$ |
| Dance2Music [27] | $0.033^{\pm.000}$ | $0.065^{\pm.001}$ | $0.097^{\pm.001}$ | $66.98^{\pm.016}$ | $8.116^{\pm.006}$ | $0.725^{\pm.011}$ | $0.043^{\pm.001}$ |
| TM2T [17] | $0.424^{\pm.003}$ | $0.618^{\pm.003}$ | $0.729^{\pm.002}$ | $1.501^{\pm.017}$ | $3.467^{\pm.011}$ | $8.589^{\pm.076}$ | $2.424^{\pm.093}$ |
| Guo *et al.* [16] | $0.455^{\pm.003}$ | $0.636^{\pm.003}$ | $0.736^{\pm.002}$ | $1.087^{\pm.021}$ | $3.347^{\pm.008}$ | $9.175^{\pm.083}$ | $2.219^{\pm.074}$ |
| MDM [44]§ | - | - | $0.611^{\pm.007}$ | $0.544^{\pm.044}$ | $5.566^{\pm.027}$ | $9.559^{\pm.086}$ | $2.799^{\pm.072}$ |
| MotionDiffuse [56]§ | $0.491^{\pm.001}$ | $0.681^{\pm.001}$ | $0.782^{\pm.001}$ | $0.630^{\pm.001}$ | $3.113^{\pm.001}$ | $9.410^{\pm.049}$ | $1.553^{\pm.042}$ |
| T2M-GPT [55] | $0.492^{\pm.003}$ | $0.679^{\pm.002}$ | $0.775^{\pm.002}$ | $0.141^{\pm.005}$ | $3.121^{\pm.009}$ | $9.722^{\pm.082}$ | $1.831^{\pm.048}$ |
| MLD [5] | $0.481^{\pm.003}$ | $0.673^{\pm.003}$ | $0.772^{\pm.002}$ | $0.473^{\pm.013}$ | $3.196^{\pm.010}$ | $9.724^{\pm.082}$ | $2.413^{\pm.079}$ |
| M2DM (Ours) *w.* linear schedule | $0.452^{\pm.003}$ | $0.678^{\pm.002}$ | $0.752^{\pm.003}$ | $0.417^{\pm.004}$ | $3.167^{\pm.008}$ | $9.972^{\pm.089}$ | $3.562^{\pm.071}$ |
| M2DM (Ours) *w.* static assessment | $0.492^{\pm.003}$ | $0.671^{\pm.003}$ | $0.775^{\pm.003}$ | $0.395^{\pm.005}$ | $3.116^{\pm.008}$ | $9.937^{\pm.075}$ | $3.413^{\pm.035}$ |
| M2DM (Ours) *w.* dynamic assessment | $0.497^{\pm.003}$ | $0.682^{\pm.002}$ | $0.763^{\pm.003}$ | $0.352^{\pm.005}$ | $3.134^{\pm.010}$ | $9.926^{\pm.073}$ | $3.587^{\pm.072}$ |

§ reports results using ground-truth motion length.

Table 1: **Comparison with the state-of-the-art methods on HumanML3D [16] test set.** The evaluation metrics are evaluated by the motion encoder from Guo *et al.* [16]. The right row → means the closer to the real motion the better.

sample 300 pairs of motions from the motion pool and extracted the selected motion features by the extractor provided by [16]. Euclidean distances of the motion pairs are computed for measuring the motion diversity. 2) R-Precision: Provided one motion representation and 32 textual descriptions, the Euclidean distances between the description feature and each motion feature in the candidates are calculated and ranked. Top-k ($k = 1, 2, 3$) average accuracy of motion-to-text retrieval is reported. 3) Frechet Inception Distance (FID) [19]: FID is the principal metric to evaluate the feature descriptions between the generated motions and the ground-truth motions by the feature extractor [16]. 4) Multimodal Distance (MM-Dist): MM-Dist measures the average Euclidean distances between each text feature and the generated motion feature. 5) MModality measures the generation diversity within the same textual descriptions. Please refer to the supplementary material for a detailed introduction.

### 4.2. Comparisons with state-of-the-art approaches

We compare our methods with other state-of-the-art methods [30, 2, 4, 12, 45, 27, 17, 16, 44, 56, 55, 5] on the HumanML3D [16] and KIT-ML [36] datasets.

**Quantitative comparisons.** For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval, followed by [16]. Most of the results are borrowed from [55]. Tab. 1 and Tab. 2 summarize the comparison results on the HumanML3D [16] and the KIT-ML [36] datasets, respectively. Firstly, the metrics scores for our reconstruction motion achieve close performance compared with real motion, which suggests the distinctive and high-quality discrete motion tokens in the codebook learned by our Transformer-based VQ-VAE. The performance of our VQ-VAE degrades when the normalized function $\ell_2$ and the orthogonal regularization term $\mathcal{L}_{\text{orth}}$ are removed. For text-to-motion generation, our approach achieves compatible performance (R-Precision, FID, and MModality) compared to other state-of-the-art methods. Moreover, compared with MDM [44] and MotionDiffuse [56] which evaluate their performance with the ground-truth motion lengths, our model can generate motion sequences with arbitrary lengths conditioned on the given textual descriptions since the [MASK] and [PAD] tokens are introduced in the discrete diffusion process.

Furthermore, we conduct more experiments to evaluate the generation performance conditioned on different lengths of textual descriptions. The testing dataset is split into three parts according to the length of the textual descriptions: less than 15 words, between 15 words and 30 words, and more than 30 words. Tab. 3 shows the quantitative results on the aforementioned three sub-sets. Both TM2T [17] and MLD [5] suffer degradation when the textual descriptions are long and complex, especially the descriptions are more than 30 words. Our method significantly outperforms the other two methods with various metrics and there is not much degradation when the descriptions are more complex. The quantitative results demonstrate the effectiveness of our proposed primary-to-secondary diffusion manner.

| Methods | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → | MModality ↑ |
|---|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | | |
| **Real motion** | $0.424^{\pm.005}$ | $0.649^{\pm.006}$ | $0.779^{\pm.006}$ | $0.031^{\pm.004}$ | $2.788^{\pm.012}$ | $11.08^{\pm.097}$ | - |
| Our VQ-VAE (Recons.) | $0.417^{\pm.004}$ | $0.621^{\pm.003}$ | $0.741^{\pm.006}$ | $0.413^{\pm.009}$ | $2.772^{\pm.018}$ | $10.851^{\pm.105}$ | - |
| Seq2Seq [30] | $0.103^{\pm.003}$ | $0.178^{\pm.005}$ | $0.241^{\pm.006}$ | $24.86^{\pm.348}$ | $7.960^{\pm.031}$ | $6.744^{\pm.106}$ | - |
| Language2Pose [2] | $0.221^{\pm.005}$ | $0.373^{\pm.004}$ | $0.483^{\pm.005}$ | $6.545^{\pm.072}$ | $5.147^{\pm.030}$ | $9.073^{\pm.100}$ | - |
| Text2Gesture [4] | $0.156^{\pm.004}$ | $0.255^{\pm.004}$ | $0.338^{\pm.005}$ | $12.12^{\pm.183}$ | $6.964^{\pm.029}$ | $9.334^{\pm.079}$ | - |
| Hier [12] | $0.255^{\pm.006}$ | $0.432^{\pm.007}$ | $0.531^{\pm.007}$ | $5.203^{\pm.107}$ | $4.986^{\pm.027}$ | $9.563^{\pm.072}$ | - |
| MoCoGAN [45] | $0.022^{\pm.002}$ | $0.042^{\pm.003}$ | $0.063^{\pm.003}$ | $82.69^{\pm.242}$ | $10.47^{\pm.012}$ | $3.091^{\pm.043}$ | $0.250^{\pm.009}$ |
| Dance2Music [27] | $0.031^{\pm.002}$ | $0.058^{\pm.002}$ | $0.086^{\pm.003}$ | $115.4^{\pm.240}$ | $10.40^{\pm.016}$ | $0.241^{\pm.004}$ | $0.062^{\pm.002}$ |
| TM2T [17] | $0.280^{\pm.005}$ | $0.463^{\pm.006}$ | $0.587^{\pm.005}$ | $3.599^{\pm.153}$ | $4.591^{\pm.026}$ | $9.473^{\pm.117}$ | $3.292^{\pm.081}$ |
| Guo *et al.* [16] | $0.361^{\pm.006}$ | $0.559^{\pm.007}$ | $0.681^{\pm.007}$ | $3.022^{\pm.107}$ | $3.488^{\pm.028}$ | $10.72^{\pm.145}$ | $2.052^{\pm.107}$ |
| MDM [44]§ | - | - | $0.396^{\pm.004}$ | $0.497^{\pm.021}$ | $9.191^{\pm.022}$ | $10.847^{\pm.109}$ | $1.907^{\pm.214}$ |
| MotionDiffuse [56]§ | $0.417^{\pm.004}$ | $0.621^{\pm.004}$ | $0.739^{\pm.004}$ | $1.954^{\pm.062}$ | $2.958^{\pm.005}$ | $11.10^{\pm.143}$ | $0.730^{\pm.013}$ |
| T2M-GPT [55] | $0.416^{\pm.006}$ | $0.627^{\pm.006}$ | $0.745^{\pm.006}$ | $0.514^{\pm.029}$ | $3.007^{\pm.023}$ | $10.921^{\pm.108}$ | $1.570^{\pm.039}$ |
| MLD [5] | $0.390^{\pm.003}$ | $0.609^{\pm.003}$ | $0.734^{\pm.002}$ | $0.404^{\pm.013}$ | $3.204^{\pm.010}$ | $10.8^{\pm.082}$ | $2.192^{\pm.079}$ |
| M2DM (Ours) *w.* linear schedule | $0.405^{\pm.003}$ | $0.629^{\pm.005}$ | $0.739^{\pm.004}$ | $0.502^{\pm.049}$ | $3.012^{\pm.015}$ | $11.375^{\pm.079}$ | $3.273^{\pm.045}$ |
| M2DM (Ours) *w.* static assessment | $0.417^{\pm.006}$ | $0.625^{\pm.003}$ | $0.741^{\pm.006}$ | $0.521^{\pm.041}$ | $3.024^{\pm.018}$ | $11.373^{\pm.081}$ | $3.317^{\pm.031}$ |
| M2DM (Ours) *w.* dynamic assessment | $0.416^{\pm.004}$ | $0.628^{\pm.004}$ | $0.743^{\pm.004}$ | $0.515^{\pm.029}$ | $3.015^{\pm.017}$ | $11.417^{\pm.97}$ | $3.325^{\pm.37}$ |

§ reports results using ground-truth motion length.

Table 2: **Comparison with the state-of-the-art methods on KIT-ML [36] test set.** The evaluation metrics are evaluated by the motion encoder from Guo *et al.* [16]. The right row → means the closer to the real motion the better.

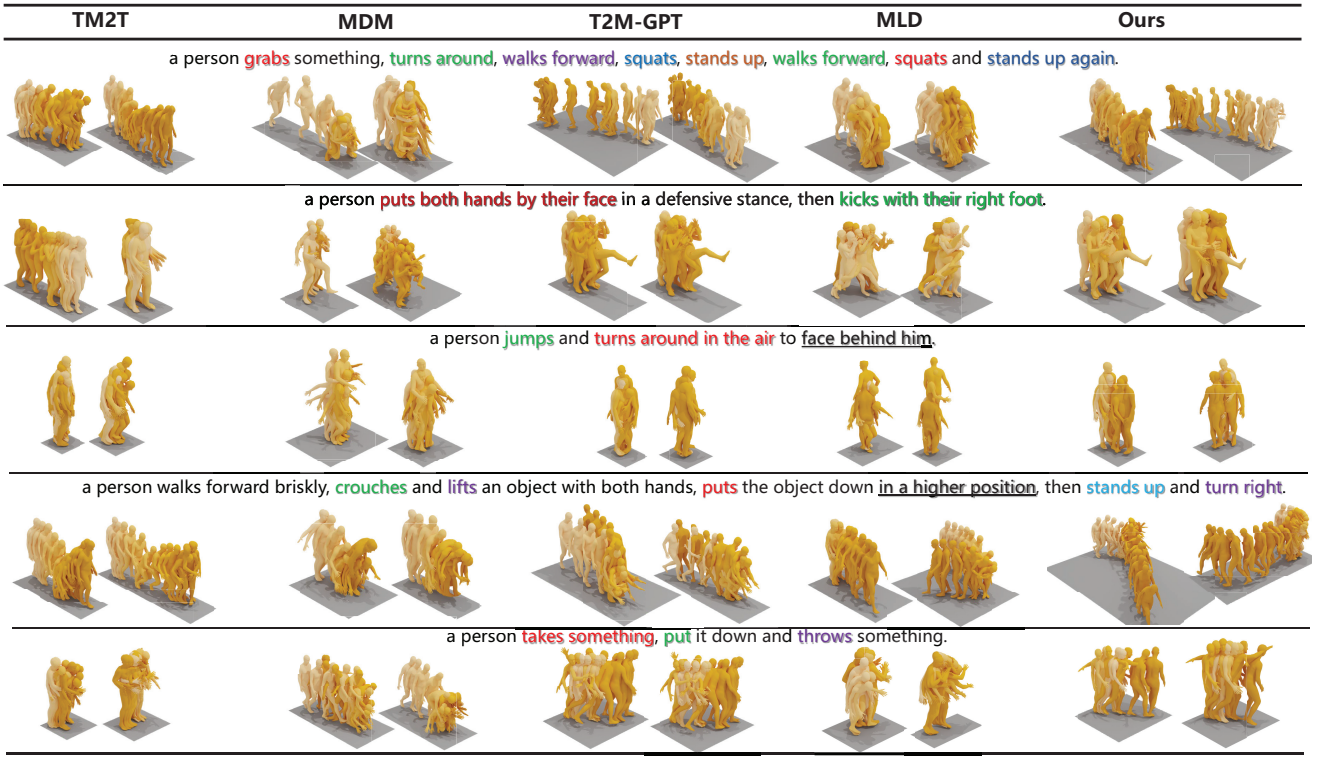



Figure 5: Qualitative comparison of the state-of-the-art methods on the HumanML3D [16] dataset. We compare our generations with TM2T [17], MDM [44], T2M-GPT [55], and MLD [5]. The color of human mesh goes from light to dark over time.

**Qualitative comparisons.** Fig. 5 qualitatively compares the generated motions from the same textual descriptions. Most of the methods suffer from motion freezing if the textual descriptions consist of several complex commands. For example, the first row in Fig. 5 shows the generation results from a long textual description with various duplicate action

| Text Length | Methods | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → | MModality ↑ |
|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-2 | Top-3 | | | | |
| All | TM2T [17] | $0.424^{\pm.003}$ | $0.618^{\pm.003}$ | $0.729^{\pm.002}$ | $1.501^{\pm.017}$ | $3.467^{\pm.011}$ | $8.589^{\pm.076}$ | $2.424^{\pm.093}$ |
| | MLD [5] | $0.481^{\pm.003}$ | $0.673^{\pm.003}$ | $0.772^{\pm.002}$ | $0.473^{\pm.013}$ | $3.196^{\pm.010}$ | $9.724^{\pm.082}$ | $2.413^{\pm.079}$ |
| | M2DM (Ours) | $0.497^{\pm.003}$ | $0.682^{\pm.002}$ | $0.763^{\pm.003}$ | $0.352^{\pm.005}$ | $3.134^{\pm.010}$ | $9.926^{\pm.073}$ | $3.587^{\pm.072}$ |
| < 15 words | TM2T [17] | $0.433^{\pm.003}$ | $0.627^{\pm.003}$ | $0.738^{\pm.003}$ | $1.592^{\pm.017}$ | $3.446^{\pm.008}$ | $8.677^{\pm.077}$ | $2.520^{\pm.051}$ |
| | MLD [5] | $0.492^{\pm.003}$ | $0.680^{\pm.004}$ | $0.779^{\pm.004}$ | $0.469^{\pm.014}$ | $3.191^{\pm.011}$ | $9.770^{\pm.010}$ | $2.586^{\pm.086}$ |
| | M2DM (Ours) | $0.508^{\pm.003}$ | $0.685^{\pm.003}$ | $0.767^{\pm.004}$ | $0.347^{\pm.015}$ | $3.129^{\pm.013}$ | $9.915^{\pm.068}$ | $3.594^{\pm.077}$ |
| 15-30 words | TM2T [17] | $0.409^{\pm.003}$ | $0.596^{\pm.002}$ | $0.708^{\pm.002}$ | $1.425^{\pm.014}$ | $3.528^{\pm.007}$ | $8.000^{\pm.051}$ | $2.477^{\pm.057}$ |
| | MLD [5] | $0.406^{\pm.005}$ | $0.592^{\pm.005}$ | $0.701^{\pm.006}$ | $0.480^{\pm.020}$ | $3.591^{\pm.020}$ | $9.089^{\pm.083}$ | $2.727^{\pm.094}$ |
| | M2DM (Ours) | $0.493^{\pm.003}$ | $0.681^{\pm.004}$ | $0.761^{\pm.004}$ | $0.353^{\pm.031}$ | $3.135^{\pm.014}$ | $9.928^{\pm.073}$ | $3.588^{\pm.081}$ |
| > 30 words | TM2T [17] | $0.353^{\pm.009}$ | $0.536^{\pm.009}$ | $0.650^{\pm.007}$ | $1.779^{\pm.061}$ | $3.577^{\pm.024}$ | $7.411^{\pm.071}$ | $2.612^{\pm.055}$ |
| | MLD [5] | $0.289^{\pm.010}$ | $0.466^{\pm.012}$ | $0.581^{\pm.015}$ | $0.947^{\pm.066}$ | $3.870^{\pm.039}$ | $8.347^{\pm.100}$ | $2.916^{\pm.063}$ |
| | M2DM (Ours) | $0.491^{\pm.004}$ | $0.677^{\pm.004}$ | $0.758^{\pm.004}$ | $0.356^{\pm.038}$ | $3.138^{\pm.025}$ | $9.930^{\pm.088}$ | $3.585^{\pm.071}$ |

Table 3: **Comparison with the state-of-the-art methods on HumanML3D [16] test set.** The evaluation metrics are evaluated by the motion encoder from Guo *et al.* [16]. The right row → means the closer to the real motion the better. The first three rows are the evaluation results on the whole testing dataset. The rest are the evaluation results from textual descriptions with different lengths.

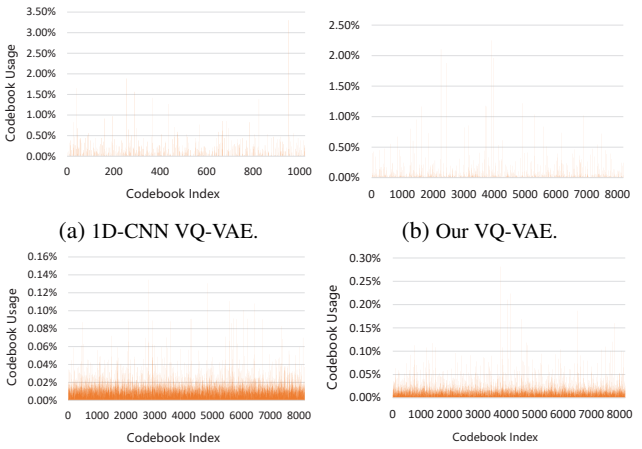
commands. MDM [44] and MLD [5] generate fewer semantic motions compared with the given description. With the help of compact discrete motion tokens, TM2T [17] and T2M-GPT [55] achieve better results compared with MDM [44] and MLD [5]. Our motion generation results conform to the textual descriptions and exhibit a high degree of diversity compared with other state-of-the-art methods, which validates the efficacy of our proposed method.

### 4.3. Ablation studies

We delved into the significance of various quantization techniques, as delineated in the initial four rows of Tab.1. The findings indicate that a rudimentary Transformer-based VQ-VAE struggles to reconstruct credible motion sequences, primarily due to the limited efficacy of the codebook. However, the incorporation of the $\ell_2$ norm and $\mathcal{L}_{\text{orth}}$ ensures that each motion token acquires a distinct motion representation. We further scrutinized the utilization of each motion token on the HumanML3D [16] test set, with the outcomes depicted in Fig.6. As evidenced by Fig.6a and 6b, both the vanilla CNN-based VQ-VAE and Transformer-based VQ-VAE grapple with sub-optimal codebook usage. This is because the tokens in the codebook aren't inherently driven to develop a distinguishing motion representation. Yet, with the aid of the $\ell_2$ norm and $\mathcal{L}_{\text{orth}}$, there's a marked surge in the codebook token utilization. Moreover, when juxtaposing the codebook usage in Fig.6c with that in Fig.6d, the latter exhibits a more equitably distributed frequency, attributable to the orthogonal regularization term $\mathcal{L}_{\text{orth}}$.

## 5. Conclusion

In this study, we introduced the priority-centric motion discrete diffusion model (M2DM) tailored for text-to-motion generation. Our model excels in producing varied and life-like human motions, aligning seamlessly with the input text. This is achieved by mastering a discrete motion representation through a Transformer-based VQ-VAE and implementing a priority-informed noise schedule within the motion discrete diffusion framework. Additionally, we unveiled two innovative techniques for gauging the priority or significance of each motion token. Experimental evaluations across two datasets underscore our model's competitive edge, outpacing existing methodologies in R-Precision and diversity, particularly with intricate textual narratives.




(a) 1D-CNN VQ-VAE. (b) Our VQ-VAE.

(c) Our VQ-VAE w. $\ell_2$ norm. (d) Our VQ-VAE w. $\ell_2$ and $\mathcal{L}_{\text{orth}}$.

Figure 6: **Codebook usage for different training strategies on HumanML3D [16] test set.** We train our VQ-VAE (*Reconstruction*) with different training streategy on the HumanML3D [16] dataset.


## Acknowledgement

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).


## References


[1] Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. 6

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 7, 8

[3] 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, 5

[4] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021. 7, 8

[5] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *arXiv preprint arXiv:2212.04048*, 2022. 1, 3, 6, 7, 8, 9

[6] 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. 3

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 3

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[10] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16091–16101, 2023. 3

[11] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *arXiv preprint arXiv:2305.10924*, 2023. 3

[12] 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*, pages 1396–1406, 2021. 7, 8

[13] 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 *IEEE/CVF International Conference on Computer Vision*, 2023. 3

[14] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 1, 3

[15] 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. 1

[16] 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*, pages 5152–5161, 2022. 2, 3, 6, 7, 8, 9

[17] 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. 1, 3, 7, 8, 9

[18] 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*, pages 2021–2029, 2020. 1, 3

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3, 5

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 3

[22] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *arXiv preprint arXiv:2102.05379*, 3(4):5, 2021. 1, 3, 5

[23] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996. 2

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 1, 3, 7, 8

[28] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279, 2022. 1, 3

[29] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 3

[30] Angela S. Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J. Mooney. 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. 7, 8

[31] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 329–336. IEEE, 2015. 6

[32] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001. 3

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3, 5

[34] 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. 1, 3

[35] 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. 1

[36] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 6, 7, 8

[37] 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*, pages 8748–8763. PMLR, 2021. 1, 6

[38] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1, 3

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[40] Sucheng Ren, Xingyi Yang, Songhua Liu, and Xinchao Wang. SG-Former: Self-guided Transformer with Evolving Token Reallocation. In *IEEE International Conference on Computer Vision*, 2023. 3

[41] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted Self-Attention via Multi-Scale Token Aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[42] 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. 3, 5

[43] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 1

[44] 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. 1, 3, 6, 7, 8, 9

[45] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 7, 8

[46] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3, 5

[48] Xingyi Yang and Xinchao Wang. Diffusion Model as Representation Learner. In *IEEE/CVF International Conference on Computer Vision*, 2023. 3

[49] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, 2022. 3

[50] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion Probabilistic Model Made Slim. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[51] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. In *Advances in neural information processing systems*, 2022. 3

[52] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[53] Runpeng Yu, Songhua Liu, Xingyi Yang, and Xinchao Wang. Distribution Shift Inversion for Out-of-Distribution Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[54] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer Baselines for Vision. *arXiv preprint arXiv:2210.13452*, 2022. 3

[55] 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. *arXiv preprint arXiv:2301.06052*, 2023. 1, 3, 6, 7, 8, 9

[56] 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. 1, 3, 7, 8