M2D2M: Multi-Motion Generation from Text with Discrete Diffusion Models

Seunggeun Chi[1,2]* , Hyung-gun Chi[2]* , Hengbo Ma[‡], Nakul Agarwal[1],
Faizan Siddiqui[1], Karthik Ramani[2†] , and Kwonjoon Lee[1†]

[1]Honda Research Institute USA [2]Purdue University
{chi65, chi45}@purdue.edu,
ramani@purdue.edu, kwonjoon_lee@honda-ri.com

Abstract. We introduce the **M**ulti-**M**otion **D**iscrete **D**iffusion **M**odels
(M2D2M), a novel approach for human motion generation from textual
descriptions of multiple actions, utilizing the strengths of discrete dif-
fusion models. This approach adeptly addresses the challenge of gener-
ating multi-motion sequences, ensuring seamless transitions of motions
and coherence across a series of actions. The strength of M2D2M lies in
its dynamic transition probability within the discrete diffusion model,
which adapts transition probabilities based on the proximity between
motion tokens, encouraging mixing between different modes. Comple-
mented by a two-phase sampling strategy that includes independent and
joint denoising steps, M2D2M effectively generates long-term, smooth,
and contextually coherent human motion sequences, utilizing a model
trained for single-motion generation. Extensive experiments demonstrate
that M2D2M surpasses current state-of-the-art benchmarks for motion
generation from text descriptions, showcasing its efficacy in interpreting
language semantics and generating dynamic, realistic motions.

Keywords: Text-to-Motion · Multi-Motion Generation · VQ-Diffusion

1 Introduction

The generation of human motion is a rapidly advancing field with profound appli-
cations in areas such as animation [3,5], VR/AR [27,28], and human-computer
interaction [33,62]. Particularly, the ability to accurately convert textual de-
scriptions into realistic, fluid human motions is not just a remarkable technical
achievement but also a crucial step towards more immersive digital experiences.

Recent progress in human motion generation has seen a surge in the use of
diffusion models [11,32,54,63]. These advancements have been critical in aligning
textual descriptions with corresponding human motions.

Previous research on human motion generation have mainly focused on single-
motion sequences (*i.e.*, sequences that feature a single action), but the ability to
generate *multi-motion sequences*, which involve a series of actions, from a set of

---

* Co-first authors. † Senior authors. ‡ Work done at Honda Research Institute.
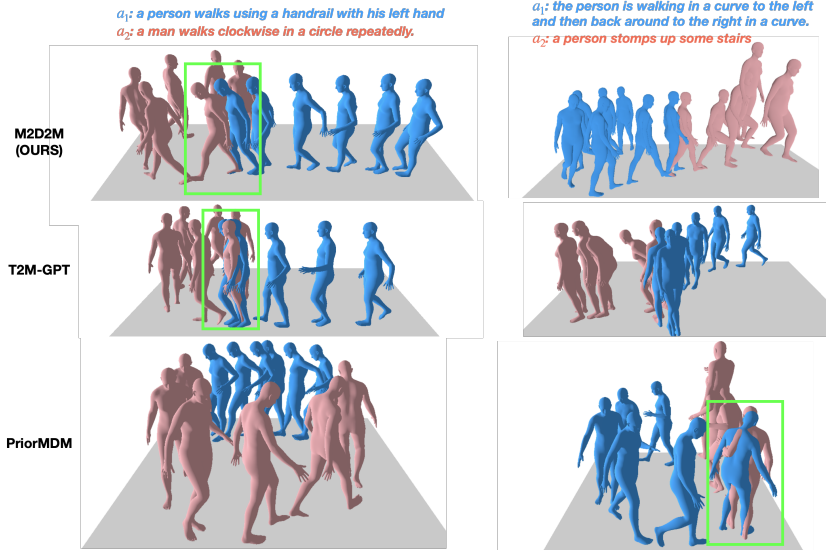
Fig. 1: Qualitative Comparison of Multi-Motion Sequences. In the transitions highlighted by the green boxes, our model shows a consistent and gradual progression of poses compared to others. This indicates that our model not only produces more realistic and smooth motions but also maintains the fidelity of each motion segment, aligning accurately with the corresponding action descriptions on top.

action descriptions, which is illustrated in Fig. 1, is crucial for many applications. This capability is particularly important in scenarios where a series of actions must be depicted in a continuous and coherent manner, such as in storytelling, interactive gaming, or complex training simulations. However, generating such sequences presents unique challenges where models often struggle to maintain continuity and coherence throughout a series of actions. Previous methods [14, 23, 30, 49, 52, 63], which generate motion for each action description separately and then attempt to connect them, frequently result in motions with abrupt transitions at action boundaries or distorted individual motions, lacking fidelity to textual descriptions for individual motion.

To tackle this challenge, we introduce the **M2D2M** (**M**ulti **M**otion **D**iscrete **D**iffusion **M**odels), a novel approach for human motion generation from textual descriptions of multiple actions. We devise a novel sampling mechanism to generate coherent and faithful multi-motion sequences using discrete diffusion models [17] trained on single-motion sequences. Furthermore, to encourage mixing between different modes (especially at multi-motion boundaries), we introduce a unique transition probability mechanism that considers the proximity between motion tokens.

A key contribution of our work is the introduction of a dynamic transition probability model within the discrete diffusion framework. This model adjusts the transition probabilities based on principles of exploration and exploitation.

Initially, it emphasizes broad exploration of diverse motions by choosing elements far apart in the codebook. As the process progresses, the focus shifts to selecting closer elements, refining the probabilities to facilitate convergence towards accurate individual motions, embodying the principle of exploitation.

Our Two-Phase Sampling (TPS) strategy represents another key contribution, addressing the challenge of generating extended human motion sequences. This method starts by sketching a coarse outline of multi-motion sequences through joint sampling, and then it refines each motion with independent sampling. We posit that the exploration of diverse motions is crucial in the initial stage of TPS, where we establish the rough layout of multi-motion sequences. Our ablation results (Table 5) demonstrate that the synergy between the dynamic transition probability model and TPS is essential for converging to optimal solutions. TPS allows for the generation of multi-motion using models *trained on single-motion generation* without additional training for multi-motion generation, which is particularly advantageous given the *scarcity of datasets containing multiple actions*. TPS enhances the natural flow of the motion, ensuring that transitions between actions are both smooth and realistic.

To quantify the transition behavior of different multi-motion generation models (as seen in Fig. 1), we introduce a novel evaluation metric, **Jerk**, to measure the smoothness of multi-motion sequences at the transitions between actions. Although similar metrics have been employed in various fields [18,41,48,60], to the best of our knowledge, our work is the first to use Jerk for evaluating transition smoothness in multi-motion generation, establishing a specialized benchmark for this task. Experiments show the effectiveness of our approach in enhancing transition smoothness, while maintaining **fidelity** to **individual motions** within motion boundaries.

The contribution of our work is summarized as follows:

- We present a two-phase sampling method for creating multi-motion sequences from text. This method enables multi-motion generation without additional training and offers a better tradeoff between the fidelity of individual motions and the smoothness of transitions between actions [49,63].
- We introduce a dynamic transition probability for the discrete diffusion model, specifically designed for human motion generation from text.
- We introduce a new evaluation metric, **Jerk**, designed to assess the *smoothness* of generated motions at action boundaries.
- Extensive experiments confirm that our methods establish state-of-the-art performance in both single-motion generation and multi-motion generation.

2 Related Works

2.1 Human Motion Generation from Text

Generating 3D human motion from textual descriptions is a growing area within human motion generation [1,2,4,16,31,39,40,58,59,65–67]. Recent advancements include using CLIP [45] for text encoding and aligning text and motion embeddings [42,53] alongside the development of motion-language datasets [19,43,44].

(a) Motion VQ-VAE   (b) Denoising Transformer   (c) Single Motion Generation   (d) Multi Motion Generation with TPS
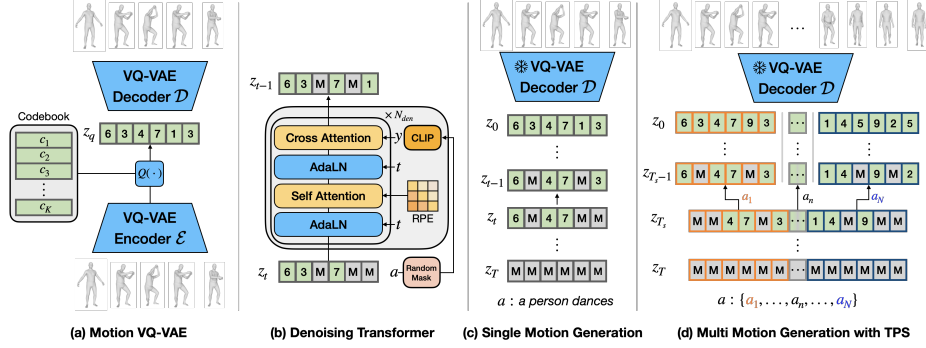
Fig. 2: Overview of M2D2M. We train a **(a)** VQ-VAE to obtain motion tokens, which is subsequently used to train a **(b)** Denoising Transformer for the discrete diffusion model. In generating human motion, we follow the **(c)** standard denoising process for single-motion generation and **(d)** employ Two-Phase Sampling (TPS) for multi-motion generation. A `<MASK>` token is denoted as 'M' in the figure.
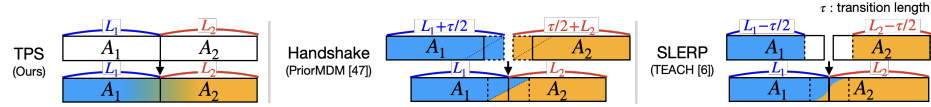
Fig. 3: Comparison of Multi-Motion Generation Algorithms. Unlike heuristic post-processing methods for combining independent motions such as Handshake [49] and SLERP [6], TPS is a single-stage algorithm for a multi-motion generation that does not require completed individual motions or a hyper-parameter for transition length.

Notably, diffusion-based models [11, 54, 64–66] have been increasingly used for text-to-motion generation.

Multi-motion generation is crucial for creating realistic and continuous human motion sequences. Recent work, such as "motion in-betweening" [14, 22, 23, 30, 52], addresses this by interpolating between motions. TEACH [6] enhanced this approach with an unfolding method using SLERP interpolation. Additionally, PriorMDM [49] introduces a handshake algorithm for smoother transitions in long sequences. Despite their effectiveness in bridging gaps, these methods require an additional stage for multi-motion generation to merge independently generated motions. As illustrated in Fig. 3, these methods not only modify the length of individual motions but also require an extra hyper-parameter for transition length to achieve smoothness, affecting the outcomes and evaluation metrics. VAE-based methods like Multi-Act [36] and TEACH [6] attempt to generate motion conditioned on a previous motion and a text but face limitations in long-term generation due to their iterative process. FineMoGen [66] utilzes spatial-temporal attention for fine-grained text description to generate multi-motion. Our method overcomes these challenges by maintaining fidelity and smoothness in the generated motion, significantly enhancing multi-motion generation without the trade-offs inherent in traditional approaches.

2.2   Discrete Diffusion Models

Diffusion models [50], defined by forward and reverse Markov processes, are integral to advancements in generative models. These models, known for transforming data into increasingly noisy variables and subsequently denoising them, benefit from stability and rapid sampling capabilities. Enhanced by neural networks learning the reverse process [13, 17, 24, 51], they are particularly effective in continuous spaces like images. Latent diffusion models [47], operating in a latent space before returning to the original data space, adeptly handle complex data distributions.

In discrete spaces such as text, diffusion models also excel. D3PM [7] and VQ-Diffusion [17] have introduced methods like structured categorical corruption and mask-and-replace to minimize errors in iterative models. This showcases the broad applicability of diffusion models. Drawing inspiration from recent studies demonstrating VQ-VAE's [56] effectiveness in human pose modeling in discrete spaces [32, 63], we apply discrete diffusion to human motion generation. Similar to our approach Kong et al [32] also introduce discrete diffusion for human motion generation from the text. However, different from this work, we introduce a new transition matrix considering the relationship between action tokens and is specifically tailored for multi-motion generation.

3   Preliminary: Discrete Diffusion Models

Discrete diffusion models [7, 17, 26] are a class of diffusion models which work by gradually adding noise to data and learning to reverse this process. Unlike continuous models like latent diffusion models [47] which operate on data represented in a continuous space, discrete diffusion models work with data representation in discrete state spaces.

Forward Diffusion Process. Since the first introduction of the discrete diffusion model [50], VQ-Diffusion [17] has improved the approach by incorporating a mask-and-replace strategy. VQ-Diffusion entails a forward diffusion process by transitioning from one token to another token. The forward Markov diffusion process for step $t-1$ to t is given by:

$$q(z_t|z_{t-1}) = \boldsymbol{v}^\top(z_t)\mathbf{Q}_t\boldsymbol{v}(z_{t-1}), \tag{1}$$

where $\boldsymbol{v}(z_t) \in \mathbb{R}^{(K+1)\times 1}$ denotes the one-hot encoded vector for the token index of z_t, and $\mathbf{Q}_t[i,j]$ is the transition probability from a token z_i to z_j at diffusion step t. The transition probability matrix $\mathbf{Q}_t \in \mathbb{R}^{(K+1)\times(K+1)}$ is written as:

$$\mathbf{Q}_t = \left[\begin{array}{c|c} \hat{\mathbf{Q}}_t & 0 \\ \hline \gamma_t \cdot \mathbf{1}^\top & 1 \end{array} \right], \text{ where } \hat{\mathbf{Q}}_t = \alpha_t\mathbf{I} + \beta_t\mathbf{1}\mathbf{1}^\top. \tag{2}$$

Here, $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is a column vector of ones, β_t represents the probability of transitioning between the different tokens, γ_t denotes the probability of transitioning to a `<MASK>` token, and $\alpha_t = 1 - K\beta_t - \gamma_t$. Due to the

Markov property, the probabilities of z_t at arbitrary diffusion time step can be derived $q(z_t|z_0) = \boldsymbol{v}^\top(z_t)\overline{\mathbf{Q}}_t\boldsymbol{v}(z_0)$, where $\overline{\mathbf{Q}}_t = \mathbf{Q}_t\mathbf{Q}_{t-1}\cdots\mathbf{Q}_1$. The matrix is constructed such that the <MASK> token always maintains its original state so that z_t converges to <MASK> token with sufficiently large t.

Conditional Denoising Process. The conditional denoising process through a neural network p_θ. This network predicts the noiseless token z_0 when provided with a corrupted token and its corresponding condition, such as a language token. For training the network p_θ, beyond the denoising objective, the training incorporates the standard variational lower bound objective [50], denoted as $\mathcal{L}_{\text{vlb}}$. The training objective with a coefficient for the denoising loss λ is:

$$\mathcal{L} = \mathcal{L}_{\text{vlb}} + \lambda\mathbb{E}_{z_t\sim q(z_t|z_0)}[-\log p_\theta(z_0|z_t, y)], \tag{3}$$

Here, the reverse transition distribution can be written as follow:

$$p_\theta(z_{t-1}|z_t, y) = \sum_{\tilde{z}_0=1}^K q(z_{t-1}|z_t, \tilde{z}_0)p_\theta(\tilde{z}_0|z_t, y). \tag{4}$$

By iteratively denoising tokens from T down to 1, we can obtain the generated token z_0 conditioned on y. The tractable posterior distribution of discrete diffusion can be expressed as:

$$\begin{aligned}q(z_{t-1}|z_t, z_0) &= \frac{q(z_t|z_{t-1}, z_0)q(z_{t-1}|z_0)}{q(z_t|z_0)}\\ &= \frac{\big(\boldsymbol{v}^\top(z_t)\mathbf{Q}_t\boldsymbol{v}(z_{t-1})\big)\big(\boldsymbol{v}^\top(z_{t-1})\overline{\mathbf{Q}}_{t-1}\boldsymbol{v}(z_0)\big)}{\boldsymbol{v}^\top(z_t)\mathbf{Q}_t\boldsymbol{v}(z_0)}.\end{aligned} \tag{5}$$

4   Multi-Motion Discrete Diffusion Model

We introduce the **M2D2M** (**M**ulti **M**otion **D**iscrete **D**iffusion **M**odel), as illustrated in Fig. 2. M2D2M is a discrete diffusion model designed specifically for generating human motion from textual descriptions of multiple actions. This model utilizes a VQ-VAE-based discrete encoding [56] which has proven effective in representing human motion [29, 32, 63]. The following sections will provide a detailed overview of our approach.

4.1   Motion VQ-VAE

To establish a codebook for discrete diffusion, we first trained a VQ-VAE model [56]. Our training approach and the model's architectural design closely follow [63]. The model comprises of an encoder $\mathcal{E}(\cdot)$, a decoder $\mathcal{D}(\cdot)$, and a quantizer $\mathcal{Q}(\cdot)$ (see Fig. 2 (a)). The encoder processes human motion, represented by $\mathbf{x}\in\mathbb{R}^{L\times D}$, converting it into motion tokens, $\mathbf{z}=\mathcal{E}(\mathbf{x})\in\mathbb{R}^{\frac{L}{4}\times D}$. Here, L signifies the length of the motion sequence and D denotes the dimensionality of each codebook. The quantizer's role is to map the motion token at any timeframe τ to the nearest codebook entry, determined by $\mathbf{z}_q[\tau]=\mathcal{Q}(\mathbf{z}[\tau])=\text{argmin}_{c_i\in\mathcal{C}}||\mathbf{z}[\tau] - c_i||_2$. Here, $\mathcal{C} = \{c_1,\ldots,c_K\}$ represents the codebook, where K signifies the total

(a) Latent Codebook Representation   **(b)** Transition Probability
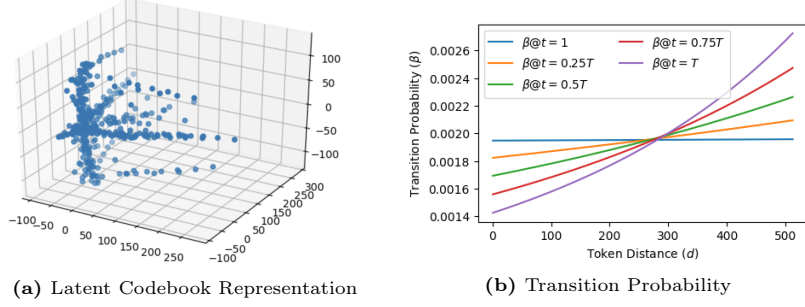
Fig. 4: (a) PCA plot representing motion tokens from the codebook of Motion VQ-VAE visualized in 3D space. (b) Plot of the dynamic transition probability function $\beta(t, d)$ across various diffusion steps t.

number of codebooks. Subsequently, the decoder utilizes these motion tokens to reconstruct the human motion as $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$. We train the motion VQ-VAE with

$$\mathcal{L}_{\text{VQ}} = ||\mathbf{x} - \hat{\mathbf{x}}||_2 + ||\mathbf{z}_q - \text{sg}[\mathbf{z}]||_2 + \lambda_{\text{VQ}}||\text{sg}[\mathbf{z}_q] - \mathbf{z}||_2, \tag{6}$$

where $\text{sg}[\cdot]$ represents stop gradient and λ_{VQ} is coefficient for commitment loss.

4.2   Dynamic Transition Probability

In the VQ-Diffusion model [17], the transition matrix $\mathbf{Q}_t$ employs a uniform transition probability β_t across different tokens, as described in Eq. (8). This method overlooks the varying proximity between motion tokens, which is essential for capturing the context of human motion. To address this, we introduce a dynamic transition probability that accounts for the distance between tokens. During the initial stages of diffusion, when the diffusion step t is large, our model adopts an exploratory approach, allowing for a wide range of transitions to foster diversity. As t progresses towards 0, the model gradually reduces the favorability of transitions between more distantly related tokens, eventually becoming uniform, identical to the original VQ-Diffusion model [17], as shown in Fig. 4. The strategy is to begin with broad exploration and progressively narrow the focus as diffusion steps decrease, thereby improving the precision and coherence in generating extended motion sequences.

Our approach mathematically formulates the transition probability at each diffusion step t as $\beta(t, d)$, where d signifies the distance between codebook tokens. The transition probability is defined by the following equation:

$$\beta(t, d) = (1 - \gamma_t - \alpha_t) \cdot \text{softmax}_d \left(\eta \cdot \frac{t}{T} \cdot \frac{d}{K} \right), \tag{7}$$

where η is a scale factor that modulates the influence of the softmax function on the relative distances between tokens. The essence of this equation lies in its softmax function over distances, which progressively assigns higher probabilities to greater distances between tokens as the diffusion step t advances. This allocation adheres to the transition probability constraint $\gamma_t + \alpha_t + \sum_{d=1}^{K} \beta(t, d) = 1$.

Algorithm 1 Two-Phase Sampling (TPS).

Given: Action sentences $\mathbf{a} = \{a^1, \ldots, a^N\}$
Hyperparameter: T_s

1: $\mathbf{z}_T \sim p(\mathbf{z}_T)$
2: $\mathbf{y} \leftarrow$ CLIP-TextEncoder($\mathbf{a}$)
3: **for** $t = T, T-1, \ldots, T_s+1$ **do**   ▷ Joint Sampling
4:     $\mathbf{z}_{t-1} \sim p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y})$
5: **end for**
6: **for** $t = T_s, T_s-1, \ldots, 1$ **do**   ▷ Independent Sampling
7:     **for** $i = 1, \ldots, N$ **do**   ▷ Parallel Process
8:         $\mathbf{z}_{t-1}^i \sim p_\theta(\mathbf{z}_{t-1}^i|\mathbf{z}_t^i, y^i)$
9:     **end for**
10: **end for**
11: $\mathbf{z}_0 \leftarrow$ Concat($\{\mathbf{z}_0^1, \ldots, \mathbf{z}_0^N\}$)
12: $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_0)$
13: **return** $\hat{\mathbf{x}}$

The distance-based modulation, scaled by $\eta \cdot \frac{t}{T} \cdot \frac{d}{K}$, ensures that as the diffusion process unfolds, the selection of token transitions is increasingly influenced by the distance metric. This approach boosts early stage exploration of denoising by favoring transitions to distant tokens. The transition matrix is defined as

$$\mathbf{Q}_t = \left[\begin{array}{c|c} \hat{\mathbf{Q}}_t & 0 \\ \hline \gamma_t \cdot \mathbf{1}^\top & 1 \end{array}\right], \text{ where } \hat{\mathbf{Q}}_t = \alpha_t \mathbf{I} + \beta_{(t, d_{i,j})} \mathbf{1}\mathbf{1}^\top. \qquad (8)$$

In this matrix, each element $d_{i,j}$ represents the distance $d(z_i, z_j)$ for indices i, j ranging from 1 to K, where i and j denote the row and column indices, respectively. Here, $d(\cdot, \cdot)$ is a distance metric specifically chosen as the rank index of codebook entries, which are sorted by their L2 distances. The dynamic and context-sensitive nature of this matrix formulation allows for an adaptive approach to the diffusion process, modifying transition probabilities in response to the evolving state of the diffusion and the relative distances between tokens.

4.3   Sampling for Multi-Motion Generation

We introduce a Two-Phase Sampling (TPS) method for the discrete diffusion model, designed to generate long-term human motion sequences from a series of action descriptions $\mathbf{a} = a^1, \ldots, a^N$. This approach enables the creation of multi-motion sequences using a model trained for single-motion generation. The overview of TPS is presented in Algorithm 1 and visually depicted in Fig. 2 **(d)**. In the algorithm, subscripts represent diffusion steps, while superscripts denote action indices. TPS effectively overcomes the challenge of ensuring smooth transitions between distinct actions, while preserving the distinctiveness of each motion segment as per its action description.

The denoising process begins by outlining the basic contours of the entire action sequence, subsequently refining these outlines to achieve semantic coherence with the textual descriptions. Inspired by this approach, our two-phase

sampling starts with *joint sampling*, where mask tokens from different actions are merged and collectively denoised using a denoising Transformer. This allows self-attention mechanism within the Transformer to integrate contextual information from action descriptions, ensuring that tokens influence each other to achieve seamless motion transitions. This step is followed by *independent sampling*, wherein each action is individually denoised within its designated boundaries to align accurately with its specific description.

The number of joint denoising steps, denoted by T_s, is carefully adjusted to achieve smooth transitions without losing the distinctiveness of each action.

4.4 Denoising Transformer

Motivated by the work of VQ-Diffusion [17], we design a denoising transformer that estimates the distribution $p_\theta(\tilde{z}_0|z_t, y)$ using the Transformer architecture [57]. An overview of our proposed model is depicted in Fig. 2 (b). To incorporate the diffusion step t into the network, we employ the adaptive layer normalization (AdaLN) [9,35]. The action sentence a is encoded into the action token y using the CLIP [45] text encoder. The Transformer's cross-attention mechanism then integrates this action information with motion, providing a nuanced conditioning with the action sentence. To enhance the human motion generation of our Transformer architecture, we added the following features:

Relative Positional Encoding. One of our primary objectives is the generation of long-term motion sequences. During the training phase, models exclusively trained on single-motion struggle to generate longer sequences. This limitation is observed when using traditional absolute positional encodings [57] that assign a static vector to each position, confining the model's capability to the maximum sequence length encountered during its training. By leveraging Relative Positional Encoding (RPE) [46], we equip our models with the ability to extrapolate beyond the sequence lengths experienced in training, thus significantly enhancing their proficiency in generating extended motion sequences.

Classifier-Free Guidance. We adopt classifier-free guidance [25]. This approach facilitates a balance between diversity and fidelity, allowing both conditional and unconditional sampling from the same model. For unconditional sampling, a learnable null token, denoted as $\varnothing$, may be substituted for the action token y. The action token y is replaced by $\varnothing$ with a probability of 10%. When inference, the denoising step is defined using s as follows:

$$\log p_\theta(z_{t-1}|z_t, y) = (s+1)\log p_\theta(z_{t-1}|z_t, y) - s \log p_\theta(z_{t-1}|z_t, \varnothing). \quad (9)$$

5 Experiments

Our experiments are designed to assess the capabilities of our model on two tasks: 1) multi-motion generation (Sec. 5.2) and 2) single-motion generation (Sec. 5.3). These experiments are conducted using the following motion-language datasets. **HumanML3D** [20] is the largest dataset in the domain of language-annotated 3D human motion, boasting 14,616 sequences of human motion, each

meticulously aligned to a standard human skeleton template and recorded at
20 FPS. Accompanying these sequences are 44,970 textual descriptions, with an
average length of 12 words. Notably, each motion sequence is associated with
a minimum of three descriptive texts. **KIT-ML** [43] comprises 3,911 human
motion sequences, each annotated with one to four natural language descriptions,
culminating in a total of 6,278 descriptions, averaging 8 words in length. The
dataset collated motion sequences sourced from both the KIT [38] and CMU [12]
datasets, with each sequence downsampled to 12.5 FPS. Each motion clip in this
collection is paired with one to four corresponding textual descriptions.

In our experiments, we perform 10 evaluations for each model, providing
a comprehensive analysis of their performance. We report both the mean and
standard deviation of these evaluations, along with a 95% confidence interval to
ensure statistical robustness. Detailed information about our model implementation can be found in Appendix A.

5.1   Evaluation of Generated Motions

Single-Motions. To evaluate generated single-motions, we adopt evaluation
metrics from previous works [19,32]. 1) **R-Top3** measures the model's precision
in correlating motion sequences with their corresponding textual descriptions,
highlighting the importance of accurate retrieval from a range of options. 2)
FID (Frechet Inception Distance) evaluates the realism of generated motions
by comparing the distribution of generated data with real data. 3) **MM-Dist**
(Multi-Modal Distance) gauges the average closeness between features of generated motions and the text features of their respective descriptions, ensuring
effective synchronization across different modalities. 4) **Diversity** assesses the
range of generated motions, reflecting the natural variation in human movement.
5) **MModality** (Multi-Modality) examines the model's capacity to produce a
wide array of plausible motions from a single text prompt, a key aspect for
versatile motion generation.
Multi-Motions. To evaluate the fidelity of each individual motion within our
generated multi-motion sequences, we compute key metrics used to evaluate
single-motion: **R-Top3**, **FID**, **MM-Dist**, and **Diversity**. This is measured by
segmenting the generated multi-motion into distinct parts, each corresponding
to a specific action description. Through this process, we can precisely analyze
the alignment between each motion segment and its corresponding action. To
evaluate the transition part for multi-motion generation, **FID** and **Diversity** are
evaluated for 40 frames of motion around boundaries, following the methodology
explained by [49]. Additionally, we defined a metric for evaluating the smoothness of multi-motion generation near motion boundaries. **Jerk** [15,18,41,48] is
based on the derivative of acceleration with respect to time. Similar to [18], we
use the logarithm of a dimensionless jerk to achieve scale invariance with respect
to velocity. Contrary to the *Jitter* described in the prior study [61], which measures the smoothness of individual motions, we calculate the integral of the time
derivative for each joint in a transitional motion. We then average these values
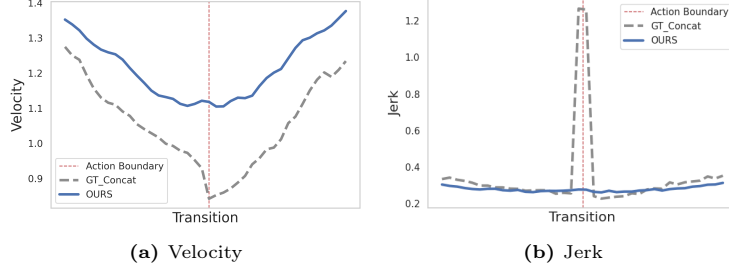
(a) Velocity

(b) Jerk

Fig. 5: Plots for **average motion transition** produced by our approach versus a concatenation of two randomly selected real motions: **(a)** Velocity. **(b)** Jerk (time derivative of acceleration normalized by peak velocity).

to derive a single statistic that represents the smoothness of the transition.

$$Jerk = \sum_p \ln \frac{1}{v_{p,\text{peak}}^2} \int_{t_1}^{t_2} \|\frac{d}{dt}\mathbf{a}_p(t)\|_2^2 \, dt, \tag{10}$$

where p denotes each joint, $v_{p,\text{peak}}$ is the maximum speed of the joint p and $[t_1, t_2]$ is the time interval for the motion transition. The metric's intuitive significance becomes clear when comparing a **single-motion transition** generated by our method to the concatenation of two arbitrarily chosen real motions. As depicted in Fig. 5, directly joining different real motions results in considerable jerk. Conversely, our method, which integrates two motions seamlessly using a two-phase sampling strategy, greatly diminishes jerk. This method facilitates a smooth transition between motions, resulting in minimal jerk.

5.2 Multi-Motion Generation

The objective of multi-motion generation is to produce continuous human motion sequences from a series of action descriptions. To achieve this, we adapt our model, initially trained for single-motion generation, to handle extended sequences. This adaptation is facilitated by our two-phase sampling strategy, as elaborated in Sec. 4.3. To assess the model's effectiveness in multi-motion generation, we created a test set by randomly combining N action sentences in the test set. This process yielded a total of 1,448 test instances for the HumanML3D and 532 instances for the KIT-ML for $N = 4$ scenario. We provide further details about test set generation for multi-motion generation in Appendix B. This approach ensures a thorough evaluation of the model's capability to generate coherent long-term motion sequences from multiple action descriptions.

Baselines. We selected PriorMDM [49] and T2M-GPT [63] as baseline models for the multi-motion generation task. For the PriorMDM model, we adhered to the original configuration without making any modifications. For the T2M-GPT, we extend the approach by concatenating the codebook for each motion and feeding it into the decoder to accommodate the auto-regressive nature of GPT

Table 1: Multi-motion generation performance on HumanML3D. 'Individual Motion' denotes individual motions within our motion boundaries. For transitions, ground truth (single) motions are independently sampled to match the transition length, while ground truth (concat) involves concatenating ground truth motions sampled from the same textual condition with generated motions.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	$0.791^{\pm.002}$	$0.002^{\pm.000}$	$2.707^{\pm.008}$	$9.820^{\pm.065}$	$0.003^{\pm.002}$	$9.574^{\pm.054}$	$1.192^{\pm.005}$
Ground Truth (Concat)	-	-	-	-	-	-	$1.371^{\pm.004}$
PriorMDM [49]	$0.586^{\pm.003}$	$0.832^{\pm.017}$	$5.901^{\pm.021}$	$9.543^{\pm.005}$	$3.351^{\pm.034}$	$\mathbf{8.801}^{\pm.098}$	$0.476^{\pm.004}$
T2M-GPT [63]	$0.719^{\pm.003}$	$0.342^{\pm.019}$	$3.512^{\pm.014}$	$9.692^{\pm.003}$	$3.412^{\pm.027}$	$8.716^{\pm.135}$	$1.321^{\pm.005}$
M2D2M	$\mathbf{0.733}^{\pm.003}$	$\mathbf{0.253}^{\pm.016}$	$\mathbf{3.165}^{\pm.019}$	$9.806^{\pm.005}$	$\mathbf{3.276}^{\pm.024}$	$8.599^{\pm.154}$	$\mathbf{1.238}^{\pm.008}$

Table 2: Multi-motion generation performance on KIT-ML.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	$0.775^{\pm.008}$	$0.034^{\pm.004}$	$2.779^{\pm.019}$	$11.055^{\pm.122}$	$0.041^{\pm.005}$	$10.434^{\pm.044}$	$1.231^{\pm.002}$
Ground Truth (Concat)	-	-	-	-	-	-	$1.469^{\pm.003}$
PriorMDM [49]	$0.292^{\pm.217}$	$3.311^{\pm.106}$	$5.451^{\pm.045}$	$\mathbf{10.842}^{\pm.067}$	$21.231^{\pm.844}$	$\mathbf{7.281}^{\pm.045}$	$0.594^{\pm.002}$
T2M-GPT [63]	$0.667^{\pm.006}$	$0.907^{\pm.059}$	$3.421^{\pm.026}$	$10.587^{\pm.089}$	$\mathbf{14.494}^{\pm.547}$	$7.059^{\pm.042}$	$1.388^{\pm.003}$
M2D2M	$\mathbf{0.711}^{\pm.006}$	$\mathbf{0.817}^{\pm.058}$	$\mathbf{3.272}^{\pm.021}$	$10.337^{\pm.122}$	$15.843^{\pm.742}$	$7.156^{\pm.048}$	$\mathbf{1.351}^{\pm.003}$

models. Implementation details of these models can be found in Appendix A. Importantly, we opted not to include iterative multi-motion generation models such as TEACH [6] and Multi-Act [36] in our comparison. This decision was based on the fact that these models rely on the relationships between subsequent actions, which are not represented in the HumanML3D and KIT-ML datasets and would require additional annotations for proper evaluation.

Results. In our comparative analysis, detailed in Tables 1 and 2, we established the number of actions for generation at $N = 4$. Our model demonstrates superior performance in terms of FID, R-Top3, and MMdist. It exhibits a Jerk value smaller than that of concatenated real motions and approaches the value of individual real motions. These results suggest that our model is capable of generating multi-motion with high fidelity while maintaining continuity and consistency throughout the entire motions. In Table 1, PriorMDM falls short in FID and R-precision metrics. It also tends to oversmoothens at the transition, resulting in Jerk metric lower than both real single-motion and concatenated real motions. This indicates that the motions generated by PriorMDM lack the depth and subtle characteristics of real motions. A qualitative comparison in Fig. 1 further demonstrates our model's superiority, showcasing more natural, continuous motion compared to T2M-GPT. This quantitative and qualitative evidence underscores our model's advanced capability in producing realistic, coherent long-term human motions.

5.3   Single-Motion Generation

Results. The task of single-motion generation involves generating human motion from individual action descriptions. In Tables 3 and 4, we present a compara-

Table 3: Single-motion generation performance on HumanML3D. The figures highlighted in **bold** and **blue** denote the best and second-best results, respectively.

Methods	R-Top 3↑	FID↓	MM-Dist↓	Diversity→	MModality↑
Ground Truth	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
VQ-VAE (reconstruction)	$0.785^{\pm.002}$	$0.070^{\pm.001}$	$3.072^{\pm.009}$	$9.593^{\pm.079}$	-
TEMOS [42]	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$0.725^{\pm.071}$	$0.368^{\pm.018}$
TM2T [21]	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.973^{\pm.076}$	$2.424^{\pm.093}$
MotionDiffuse [64]	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$\mathbf{9.410}^{\pm.049}$	$1.553^{\pm.042}$
MLD [11]	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$8.589^{\pm.083}$	$2.219^{\pm.074}$
T2M-GPT [63]	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
AttT2M [67]	$0.786^{\pm.006}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$\textcolor{blue}{2.452}^{\pm.051}$
M2DM [32]	$0.763^{\pm.003}$	$0.352^{\pm.005}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$	$\mathbf{3.587}^{\pm.072}$
M2D2M (w/ β_t)	$\textcolor{blue}{0.796}^{\pm.002}$	$\textcolor{blue}{0.115}^{\pm.006}$	$\mathbf{3.036}^{\pm.008}$	$9.680^{\pm.074}$	$2.193^{\pm.077}$
M2D2M (w/ $\beta(t,d)$)	$\mathbf{0.799}^{\pm.002}$	$\mathbf{0.087}^{\pm.004}$	$3.018^{\pm.008}$	$\textcolor{blue}{9.672}^{\pm.086}$	$2.115^{\pm.079}$

Table 4: Single-motion generation performance on KIT-ML. The figures highlighted in **bold** and **blue** denote the best and second-best results, respectively.

Methods	R-Top3↑	FID↓	MM-Dist↓	Diversity→	MModality↑
Ground Truth	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
VQ-VAE (reconstruction)	$0.740^{\pm.006}$	$0.472^{\pm.011}$	$2.986^{\pm.027}$	$10.994^{\pm.120}$	-
TEMOS [42]	$0.687^{\pm.002}$	$3.717^{\pm.028}$	$3.417^{\pm.008}$	$10.84^{\pm.004}$	$0.532^{\pm.018}$
TM2T [21]	$0.587^{\pm.005}$	$3.599^{\pm.051}$	$4.591^{\pm.019}$	$9.473^{\pm.100}$	$\mathbf{3.292}^{\pm.034}$
MotionDiffuse [64]	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$\mathbf{2.958}^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
MLD [11]	$0.734^{\pm.007}$	$\textcolor{blue}{0.404}^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
T2M-GPT [63]	$0.737^{\pm.006}$	$0.717^{\pm.041}$	$3.053^{\pm.026}$	$\textcolor{blue}{10.862}^{\pm.094}$	$1.912^{\pm.036}$
AttT2M [67]	$\textcolor{blue}{0.751}^{\pm.006}$	$0.870^{\pm.039}$	$3.309^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
M2DM [32]	$0.743^{\pm.004}$	$0.515^{\pm.029}$	$3.015^{\pm.017}$	$11.417^{\pm.097}$	$\textcolor{blue}{3.325}^{\pm037}$
M2D2M (w/ β_t)	$0.743^{\pm.006}$	$\textcolor{blue}{0.404}^{\pm.022}$	$3.018^{\pm.019}$	$10.749^{\pm.102}$	$2.063^{\pm.066}$
M2D2M (w/ $\beta(t,d)$)	$\mathbf{0.753}^{\pm.006}$	$\mathbf{0.378}^{\pm.023}$	$\textcolor{blue}{3.012}^{\pm.021}$	$10.709^{\pm.121}$	$2.061^{\pm.067}$

tive analysis of single-motion generation performance against selected baselines. We present comprehensive comparison in Appendix D. Our approach outperforms current state-of-the-art methods on both the HumanML3D and KIT-ML datasets, particularly excelling in FID and R-Top 3 metrics. While not leading but closely competitive in other metrics such as MM-Distance, Diversity our method demonstrates robust near-best performance. In terms of multi-modality, our model exhibits lower performance compared to the top-performing models. However, there appears to be a trade-off between multi-modality and FID, suggesting that models with higher FID scores may achieve better multi-modality, as observed in the case of M2DM [32] and TM2T [21].

5.4 Ablation Studies

We conduct ablation studies to assess the effects of Dynamic Transition Probability and Two-Phase Sampling on our model. Due to space constraints, we have included further ablation studies in Appendix C.

Dynamic Transition Probability. We first investigate the impact of dynamic transition probability presented in wcrefsec:transition. In Table 5, we compare the performance of our model with a dynamic token transition probability, denoted as $\beta(t,d)$, as opposed to a static one, β_t. Dynamic transition probability substantially enhances our model's performance, particularly in terms of FID.

Table 5: Ablation studies on multi-motion generation performance on HumanML3D. 'Individual Motion' denotes individual motions within our motion boundaries.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	0.791	0.002	2.707	9.820	0.003	9.574	1.192
Ground Truth (Concat)	-	-	-	-	-	-	1.371
$\beta_t, T_s = 100$	0.749	0.212	3.015	9.990	3.324	8.681	1.248
$\beta_t, T_s = 90$	0.738	0.253	3.164	9.822	3.483	**8.625**	1.265
$\beta(t, d), T_s = 100$	0.751	0.196	3.012	9.894	3.340	8.751	1.248
$\beta(t, d), T_s = 90$	0.733	0.253	3.165	9.806	**3.276**	8.599	**1.238**

Table 6: Multi-motion generation on different smoothing methods on HumanML3D.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	0.791	0.002	2.707	9.820	0.003	9.574	1.192
Ground Truth (Concat)	-	-	-	-	-	-	1.371
Handshake [49]	0.635	1.279	4.182	8.939	**3.039**	8.566	1.097
SLERP [6]	0.549	1.402	4.679	8.535	4.873	7.912	−3.554
TPS (Ours)	**0.733**	**0.254**	**3.165**	**9.806**	3.276	**8.599**	**1.238**

In addition, the table shows that it enhances the smoothness proven by the lowest jerk when it is combined with two-phase sampling. This improvement underscores the importance of the synergistic effect of TPS and $\beta(t, d)$.

Two-Phase Sampling (TPS). To compare multi-motion generation algorithms illustrated in Fig. 3, we evaluate the performance of these algorithms on VQ-Diffusion model in Table 6, highlighting the effectiveness of TPS in multi-motion generation. TPS significantly improves the smoothness of long-term motion sequences while maintaining fidelity to individual motions, as evidenced by the improved Jerk and FID metrics. However, the results for Handshake and SLERP show an over-smoothing effect when compared to the original dataset, with their Jerk values being lower than that of the single ground truth. Notably, SLERP even exhibits a negative Jerk value, indicating excessive smoothing.

6   Conclusion & Discussion

We present M2D2M, a model designed for generating multi-motion sequences from a set of action descriptions. Incorporating a Dynamic Transition Matrix and Two-Phase Sampling, M2D2M achieves state-of-the-art performance in generating human motion from text tasks. For multi-motion generation, a ground truth is absent as we create extended sequences from multiple action descriptions. Consequently, we introduce a new evaluation metric to assess the smoothness of the motion. However, these metrics do not comprehensively evaluate the generated multi-motion sequences, as they do not account for all possible scenarios. Addressing this limitation remains for future work. Additionally, our research aims to enhance virtual reality and assistive technologies but could raise privacy and security concerns, requiring strict data policies and transparent monitoring.

Acknowledgements

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 5915–5920. IEEE (2018)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
3. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
4. Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5223–5232 (2020)
5. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. arXiv preprint arXiv:2303.14613 (2023)
6. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action compositions for 3d humans. In: International Conference on 3D Vision (3DV) (September 2022)
7. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems **34**, 17981–17993 (2021)
8. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: Large-scale text-conditional 3d human motion generation. arXiv preprint arXiv:2305.09662 (2023)
9. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
10. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)
11. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
12. CMU Graphics Lab: Motion capture database. http://mocap.cs.cmu.edu (2016)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B., Yuan, Y.: Single-shot motion completion with transformer. arXiv preprint arXiv:2103.00776 (2021)

15. Flash, T., Hogan, N.: The coordination of arm movements: an experimentally confirmed mathematical model. Journal of neuroscience **5**(7), 1688–1703 (1985)

16. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)

17. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)

18. Gulde, P., Hermsdörfer, J.: Smoothness metrics in complex movement tasks. Frontiers in Neurology **9**, 615 (09 2018). https://doi.org/10.3389/fneur.2018.00615

19. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)

20. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)

21. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)

22. Harvey, F.G., Pal, C.: Recurrent transition networks for character locomotion. In: SIGGRAPH Asia 2018 Technical Briefs, pp. 1–4 (2018)

23. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion inbetweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020)

24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

26. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions (2021), https://arxiv.org/abs/2102.05379

27. Huang, G., Qian, X., Wang, T., Patel, F., Sreeram, M., Cao, Y., Ramani, K., Quinn, A.J.: Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2021)

28. Ipsita, A., Li, H., Duan, R., Cao, Y., Chidambaram, S., Liu, M., Ramani, K.: Vrfromx: from scanned reality to interactive virtual experience with human-in-the-loop. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–7 (2021)

29. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023)

30. Kaufmann, M., Aksan, E., Song, J., Pece, F., Ziegler, R., Hilliges, O.: Convolutional autoencoders for human motion infilling. In: 2020 International Conference on 3D Vision (3DV). pp. 918–927. IEEE (2020)

31. Komura, T., Habibie, I., Holden, D., Schwarz, J., Yearsley, J.: A recurrent variational autoencoder for human motion synthesis. In: The 28th British Machine Vision Conference (2017)

32. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14806–14816 (2023)
33. Kucherenko, T., Hasegawa, D., Henter, G.E., Kaneko, N., Kjellström, H.: Analyzing input and output representations for speech-driven gesture generation. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. pp. 97–104 (2019)
34. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in neural information processing systems **32** (2019)
35. Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: ViTGAN: Training GANs with vision transformers. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=dwg5rXg1WS_
36. Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: AAAI Conference on Artificial Intelligence (AAAI) (2023)
37. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. Learning **2018**(1) (2018)
38. Mandery, C., Terlemez, Ö., Do, M., Vahrenkamp, N., Asfour, T.: The kit whole-body human motion database. In: 2015 International Conference on Advanced Robotics (ICAR). pp. 329–336. IEEE (2015)
39. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 474–489. Springer (2020)
40. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019)
41. Mobini, A., Behzadipour, S., Foumani, M.: Test–retest reliability of kinect's measurements for the evaluation of upper body recovery of stroke patients. Biomedical engineering online **14**, 75 (08 2015). https://doi.org/10.1186/s12938-015-0070-0
42. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
43. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
44. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
46. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

48. Roren, A., Mazarguil, A., Vaquero-Ramos, D., Deloose, J.B., Vidal, P.P., Nguyen, C., Rannou, F., Wang, D., Oudre, L., lefevre colau, m.m.: Assessing smoothness of arm movements with jerk: A comparison of laterality, contraction mode and plane of elevation. a pilot study. Frontiers in Bioengineering and Biotechnology **9** (01 2022). https://doi.org/10.3389/fbioe.2021.782740

49. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=dTpbEdN9kr

50. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)

51. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)

52. Tang, X., Wang, H., Hu, B., Gong, X., Yi, R., Kou, Q., Jin, X.: Real-time controllable motion transition for characters. ACM Transactions on Graphics (TOG) **41**(4), 1–10 (2022)

53. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)

54. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu

55. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)

56. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

58. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4394–4402 (2019)

59. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European conference on computer vision (ECCV). pp. 265–281 (2018)

60. Yi, X., Zhou, Y., Xu, F.: Transpose: real-time 3d human translation and pose estimation with six inertial sensors. ACM Trans. Graph. **40**(4) (jul 2021). https://doi.org/10.1145/3450626.3459786, https://doi.org/10.1145/3450626.3459786

61. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions On Graphics (TOG) **40**(4), 1–13 (2021)

62. Yin, T., Hoyet, L., Christie, M., Cani, M.P., Pettré, J.: The one-man-crowd: Single user generation of crowd motions using virtual reality. IEEE Transactions on Visualization and Computer Graphics **28**(5), 2245–2255 (2022)

63. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

64. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
65. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
66. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems **36** (2024)
67. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 509–519 (2023)