This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Deterministic-to-Stochastic Diverse Latent Feature Mapping for Human Motion Synthesis

Yu Hua[1,3] Weiming Liu[2*] Gui Xu [4] Yaqing Hou [3]
Yew-Soon Ong [1,5] Qiang Zhang [3]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]ByteDance Inc. [3]Dalian University of Technology, China [4]Dalian University, China
[5]Centre for Frontier AI Research, Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

yu_hua,asysong@ntu.edu.sg lwming95@gmail.com guixu@s.dlu.edu.cn houyq,zhangq@dlut.edu.cn

Abstract

Human motion synthesis aims to generate plausible human motion sequences, which has raised widespread attention in computer animation. Recent score-based generative models (SGMs) have demonstrated impressive results on this task. However, their training process involves complex curvature trajectories, leading to unstable training process. In this paper, we propose a Deterministic-to-Stochastic Diverse Latent Feature Mapping (DSDFM) method for human motion synthesis. DSDFM consists of two stages. The first human motion reconstruction stage aims to learn the latent space distribution of human motions. The second diverse motion generation stage aims to build connections between the Gaussian distribution and the latent space distribution of human motions, thereby enhancing the diversity and accuracy of the generated human motions. This stage is achieved by the designed deterministic feature mapping procedure with DerODE and stochastic diverse output generation procedure with DivSDE. DSDFM is easy to train compared to previous SGMs-based methods and can enhance diversity without introducing additional training parameters. Through qualitative and quantitative experiments, DSDFM achieves state-of-the-art results surpassing the latest methods, validating its superiority in human motion synthesis.

1. Introduction

Human motion synthesis task aims to generate diverse and high quality 3D human motion sequences. This task has wide-ranging applications, such as human motion understanding [7, 14, 18], human-robot interactions [49, 61], and computer graphics [44]. Recent efforts mainly focus on conditional and unconditional human motion generation. Conditional human motion generation aims to generate human motion sequences under some limiting factors, such as music [16, 17], audio [1, 15, 28], and action labels [33, 33, 61]. Unconditional human motion generation intends to generate diverse human motions [32, 35] from diverse data, which still presents a significant challenge, especially when the human motion datasets are unstructured. In this paper, we focus on conditional (under the action labels) and unconditional human motion generations, as shown in Figure 1. Efficiently generating diverse and accurate human motions remains a tremendous challenge, which has led to the development of many different generative models.

Recent advancements in deep generative models, including Variational Autoencoders (VAEs) [53, 57, 60], Generative Adversarial Networks (GANs) [31], score-based generative models (SGMs), and related techniques [12, 39, 46, 51, 58], emerge as the dominant approaches for capturing the data distribution. Specifically, VAEs leverage an encoder-decoder network to learn the latent representation of human motion distribution. VAEs require approximate variational or Monte Carlo inference techniques, which tend to be intractable for complex models. GANs utilize a generator and discriminator to generate real-like motions from random noise. Unfortunately, GANs are known to suffer from numerical instability and mode collapse issues. SGMs define a forward diffusion process that maps data to noise by gradually perturbing the input data. Generation corresponds to a reverse process that synthesizes novel data via iterative denoising process. Even though they have presented high fidelity in generation, it is important to note that these methods have the challenge of curve trajectory modeling within diffusion models, as their forward pass is inherently designed to exhibit curvature in SDE, following either a Variance Preserving SDE (VPSDE) or a Variance Exploding

*Corresponding author

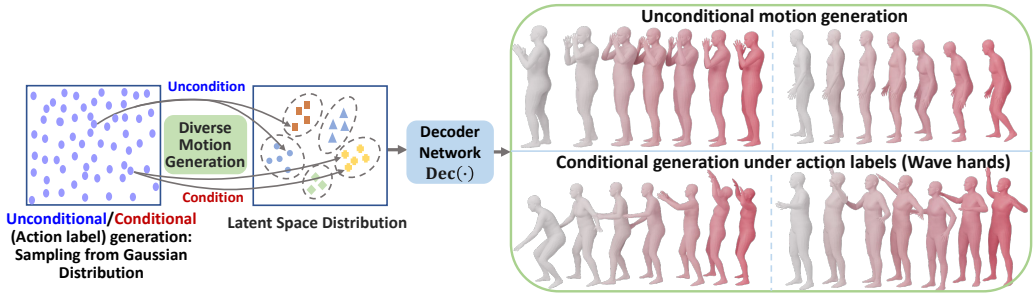

Figure 1. Examples of the inference process for human motion synthesis. Our method aims to generate diverse and accurate human motion sequences through the designed generative model.

SDE (VESDE) [39], leading to unstable training process and slow inference process. Recent methods, like DDIM [37], aim to accelerate the inference process by one-step or few-step generator, nevertheless, these methods lead to an obvious performance drop [5], and the training process is still unstable.

To synthesize diverse and accurate human motions, we propose a novel method called DSDFM for human motion synthesis. The proposed method has straight trajectories and is easy to train compared to previous SGMs methods, while guaranteeing the diversity and accuracy of the generated human motions. The proposed DSDFM consists of two stages. In the first stage, a human motion reconstruction process is designed to learn the latent space distribution of human motions and motion representation. This process is implemented by the Vector Quantized Variational Autoencoders (VQVAE) [47] network. In the second stage, we design a diverse motion generation module, including deterministic feature mapping procedure and stochastic diverse output generation procedure. Deterministic feature mapping procedure aims to explore the optimal solution for building the connections between the Gaussian distribution and the latent space distribution of human motions using the designed Deterministic Ordinary Equation (DerODE) operation. DerODE has a straight training trajectory compared to previous diffusion generative methods [12, 39] and Flow Matching [27]. Although DerODE is easy to train, it is hard to generate highly diverse human motion patterns since DerODE could only provide deterministic output. Therefore, the designed stochastic diverse output generation procedure aims to enhance the diversity of generated human motions through Diverse Stochastic Differential Equations (DivSDE). It is noted that DivSDE operates during the sampling process of the model without introducing additional training processes.

In summary, our main contributions are as follows:

- We propose a novel method called Diverse Latent Feature Mapping (DSDFM) for human motion synthesis. DSDFM is efficient to train and to utilize at sampling process, and can be used for conditional and unconditional generation.
- We propose an optimal solution to build the connection between the Gaussian distribution and the latent space distribution of human motions. In addition, we provide a stochastic diverse output generation process during the sampling process without reintroducing additional training processes.
- The proposed method DSDFM is evaluated on widely-used human motion datasets in the comprehensive experiments. The obtained results demonstrate the effectiveness of the proposed method over the state-of-the-art approaches for conditional and unconditional human motion generation tasks.

2. Related Work

2.1. Human Motion Synthesis

Conditional human motion synthesis aims to generate diverse and realistic human motions [42, 50, 54, 55] according to various conditional inputs, such as action labels [3, 41, 59] and music [11, 43]. For example, MDM [45] utilized a diffusion-based generative model for action-conditioned human motion generations, and reported a trading-off between diversity and fidelity of human motions due to the curve trajectory of training and sampling process. MLD [3] proposed to utilize the DDPM in latent space for human motion generations given an input action label, which also encountered the same problem as DDPM. In addition, the unconditional human motion synthesis [2, 35, 56] task also encounters the same issues although a series of achievements have been made in this field. For example, Holden et al. [13] presented a pioneer work in deep unconditional human motion synthesis. Modi [35] employed the style of StyleGAN to synthesize human motions. Unfortunately, these methods usually suffer from mode collapse or mode mixture. In contrast, we propose a novel method for conditional and unconditional human motion synthesis, which is easy to train compared to previous diffusion-based methods while guaranteeing the diversity of generated motions.

2.2. Diffusion Generative Models

Recent years have witnessed a promising potential in modeling data distributions with diffusion generative models using Denoising diffusion probabilistic modeling (DDPM) [12]

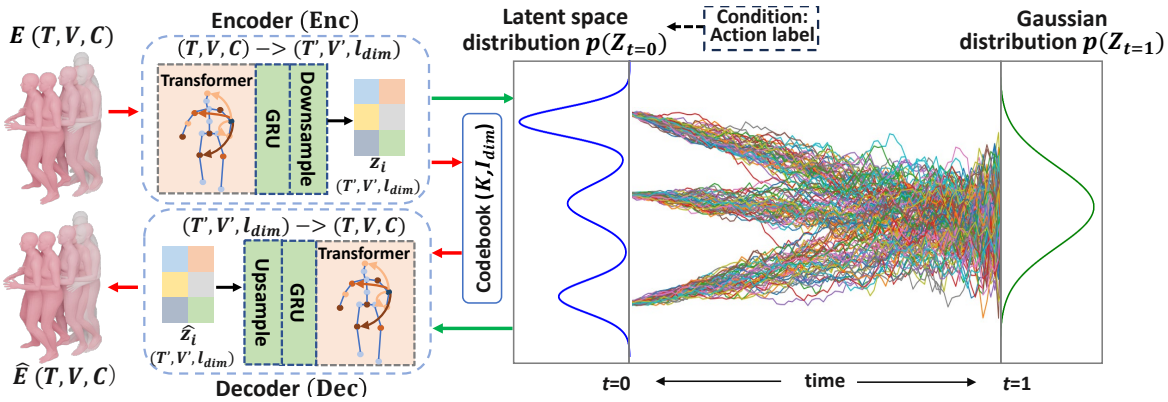

Figure 2. The overview of the proposed method DSDFM. The red arrow denotes the first stage and the green arrow denotes the second stage of DSDFM.

and score-based generative models (SGMs) [39], which define a forward diffusion process that maps data to noise by gradually perturbing the input data. Variants of SGMs and techniques have been applied to images [6], audio [30]. For example, Robin et al. [8] proposed latent diffusion models (LDMs) that work on a compressed latent space of lower dimensionality for high-resolution image synthesis. LSGM [46] proposed to train SGMs in a latent space, which relies on the variational autoencoder framework to generate diverse images. However, it is important to note that these methods have the challenge of curve trajectory modeling within diffusion models, as their complex forward and backward processes are inherently designed to exhibit curvature, leading to unstable training process and slow sampling process. Although DDIM and related techniques [37] can shorten the sampling process, they often result in a performance drop [5, 52]. Flow Matching-based methods [20, 27] offer a more robust and stable alternative to diffusion models during the training process. However, the trajectories between source and target distributions remain relatively curved, and more importantly, these methods struggle to produce highly diverse samples, often sacrificing diversity in the training process. In contrast, we propose a generative model DSDFM for human motion generation tasks. This model utilizes straight trajectories, making it easier to train compared to other diffusion models. Additionally, it is capable of generating diverse human motion sequences.

3. Preliminary

Score-based diffusion models gradually perturb data by a forward diffusion process, and then reverse it to recover the data [38, 39]. Under the stochastic differential equation (SDE) framework proposed in [39], diffusion models construct a process $x(t)_{t=0}^{T}$ indexed by a continuous time variable $t \in [0, T]$, such that $x(0) \sim p_0$, for which we have a dataset of i.i.d. samples, and $x(T) \sim p_T$, we have a tractable form to generate samples efficiently. p_0 is the data distribution, p_T is the prior distribution. The forward diffusion process can be modeled as the solution to an Itô SDE:

$$dx_t = f(x, t)dt + g(t)dw_t, \tag{1}$$

where w is the standard Wiener process (a.k.a., Brownian motion), $f(\cdot, t) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a vector valued function called the $drift$ coefficient of $x(t)$, and $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is a scalar function known as the diffusion coefficient of $x(t)$. There are various ways of designing the SDE such that it diffuses the data distribution into a fixed prior distribution p_T. By starting from samples of $x(T) \sim p_T$ and reversing the process, we can obtain samples $x(0) \sim p_0$. The reverse of a diffusion process is also a diffusion process, running backwards in time and given by the reverse-time SDE:

$$dx_t = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w}_t, \tag{2}$$

where $\bar{w}$ is a standard Wiener process when time flows backwards from T to 0, dt is an infinitesimal negative timestep. Once the score $\nabla_x \log p_t(x)$ is learned, we can derive the reverse diffusion process and simulate it to sample from p_0.

4. The Proposed Method

This paper aims to synthesize diverse and realistic human motion sequences. The overview of the proposed method is shown in Figure 2. The conditional motion generation is performed under the action labels (Action-to-Motion task). Once the action labels are removed, the entire process becomes unconditional motion generation. The training process of the DSDFM mainly involved in two stages. The first stage is the human motion reconstruction process (Section 4.1), which aims to learn the human motion representation and capture the latent space distribution of human motions. The second stage (Section 4.2) aims to build the connections between the Gaussian distribution and the latent space distribution using the designed deterministic feature mapping procedure (DerODE) (Section 4.2.1). Moreover, we employ the stochastic diverse generation process (DivSDE) to enhance the diversity of generated human motions (Section 4.2.2).

4.1. Human Motion Reconstruction

The human motion reconstruction network aims to learn the representation and the latent space distribution of human motions. In this process, we utilize VQVAE [47] to capture dynamic spatio-temporal features of human motions. Specifically, the input is a sequence of human motion sequence $\boldsymbol{E} = \{e_1, e_2, \cdots, e_T\}$ with the length of T, where $e_t \in \mathbb{R}^{V \times C}$ is denoted by 3D coordinates at time t, C denotes the 3D coordinates of human joints ($C = 3$), V is the used number of human joints. The encoder of VQ-VAE aims to transform motion sequence into latent features, i.e., $\mathrm{Enc}(\boldsymbol{E}) \rightarrow z_i \in \boldsymbol{Z}$. z_i is substituted by its closest vector k_j using a quantization codebook, where $\hat{z}_i = \arg\min_{k_j \in K} ||z_i - k_j||$. The quantized feature $\hat{z}_i$ is decoded to $\hat{\boldsymbol{E}}$ by the decoder network, i.e., $\mathrm{Dec}(\hat{z}_i) \rightarrow \hat{\boldsymbol{E}}$.

In this work, the encoder $\mathrm{Enc}(\cdot)$ and decoder $\mathrm{Dec}(\cdot)$ networks are implemented by the Transformer [48] and GRU [4] module. For the Transformer process, we project the input human motion sequences into matrices Q, K, and V by W_Q, W_K, W_V. The summary of the spatial joints $\tilde{M}_t$ is calculated by aggregating all the joint information using the multi-head mechanism (head_i). The GRU_ϕ with parameter ϕ intends to capture the smoothness property of human motions, and then encode the human motions into latent space $\boldsymbol{Z}$. In addition, the decoder $\mathrm{Dec}(\cdot)$ aims to map the latent space $\boldsymbol{Z}$ back to the reconstructed human motion sequence. The VQVAE is optimized by minimizing the following loss function:

$$\mathcal{L}_{VQ} = \mathcal{L}(\boldsymbol{E}, \hat{\boldsymbol{E}}) + ||\hat{z} - \mathrm{sg}(z)||_2^2 + \beta || \mathrm{sg}(\hat{z}) - z||_2^2, \tag{3}$$

where $\mathrm{sg}[\cdot]$ is the stop gradient operator and β is the hyper parameter. The first term $\mathcal{L}(\boldsymbol{E}, \hat{\boldsymbol{E}}) = \sum_{t=1}^{T} \sum_{v=1}^{V} \parallel e_t^{(v)} - \hat{e}_t^{(v)} \parallel_2$, represents the reconstruction error. The second term aims to optimize the codebook, and the last term is to optimize the encoder by pushing z close to its nearest latent vector in the codebook. The human motion reconstruction process aims to learn the human motion representation and map the human motions into latent space $\boldsymbol{Z}$.

4.2. Diverse Motion Generation

Although we have established the human motion reconstruction in Section 4.1, we still cannot generate diverse human motion accordingly. The main reason is that the latent feature space for human motion is rather complicated and hard to sample. Therefore, it is essential to model the latent feature space for human motion by establishing the relationship between a Gaussian distribution and the latent space distribution. Previous diffusion-based generative methods [39] and flow matching methods [20], suffer from instability during training, exhibiting curved trajectories or difficulty in generating diverse samples. To tackle this issue, we innovatively propose a diverse motion generation module to enhance the diversity and accuracy of generated human motion sequences. Diverse motion generation module consists of two steps, i.e., *deterministic feature mapping procedure* and *stochastic diverse output generation procedure*. We will introduce the details of our proposed diverse motion generation module in this section.

4.2.1. Deterministic Feature Mapping Procedure

To start with, we first introduce the deterministic feature mapping procedure. The deterministic feature mapping procedure is designed to model the relationship between Gaussian distribution $p(\boldsymbol{Z}_{t=1})$ and the latent distribution for human motion $p(\boldsymbol{Z}_{t=0})$ efficiently. Specifically, we propose Deterministic Ordinary Equation (DerODE) operation in the deterministic feature mapping procedure by depicting the transformation with Proposition 1 to achieve the corresponding goal.

Proposition 1. *Given the ordinary equation $d\boldsymbol{z}_t = u(\boldsymbol{z}_t, t)dt$, where $u(\boldsymbol{z}_t, t)$ denotes the drift function, and suppose the probability of data distribution $z(t)$ is set to be $p(z(t)) = \mathcal{N}(\mu(t), \sigma^2(t))$ at the time step t, where $\mu(t)$ and $\sigma(t)$ denote the mean and variance of the Gaussian distribution respectively, the drift function $u(\boldsymbol{z}_t, t)$ can be shown as:*

$$u(z_t, t) = \sigma'(t) \cdot \frac{z(t) - \mu(t)}{\sigma(t)} + \mu'(t). \tag{4}$$

The illustrations of Proposition 1 can be found in [19]. We can utilize Proposition 1 to transform the data across different distributions. Specifically, we need to establish the connections among the latent motion feature space $p(\boldsymbol{Z}_{t=0})$ and the standard Normal distribution $p(\boldsymbol{Z}_{t=1}) = \mathcal{N}(0, \boldsymbol{I})$ by carefully designing $\mu(t)$ and $\sigma(t)$ for the downstream generation task. However, previous methods (e.g., Flow Matching [19]) just randomly sample data across different distributions, leading to less efficient model training and inference. To get straighter paths for the training process, we can introduce the optimal transport (OT) theory into this task. As discussed in [21–26, 34], the OT problem aims to minimize the displacement cost between two distributions. Thus, we can leverage the transport plan $\boldsymbol{\pi}$ to build connections between two different distributions. The calculation of the optimal transport $\boldsymbol{\pi}$ can be formulated as:

$$\min_{\boldsymbol{\pi} \in \Delta} J_{\mathrm{OT}} = \langle \boldsymbol{\pi}, \boldsymbol{C} \rangle$$
$$s.t.\ \Delta = \left\{ \sum_{j=1}^{N} \pi_{ij} = a_i, \quad \sum_{i=1}^{N} \pi_{ij} = b_j, \quad \pi_{ij} \geq 0 \right\}, \tag{5}$$

where Δ denotes the constraints on $\boldsymbol{\pi}$. $\boldsymbol{C}$ denotes the cost distance matrix which can be calculated as $C_{ij} = ||\boldsymbol{z}_{0,i} - \boldsymbol{z}_{1,j}||_2^2$ accordingly, where $\boldsymbol{z}_{0,i} \sim p(\boldsymbol{Z}_{t=0})$ and $\boldsymbol{z}_{1,j} \sim p(\boldsymbol{Z}_{t=1})$. The optimization process for solving $\boldsymbol{\pi}$ has been provided in the Appendix A. Then we can obtain the matched data samples $(\boldsymbol{z}_{0,i}, \boldsymbol{z}_{1,j}) \sim \boldsymbol{\pi}$ via the coupling

matrix. Hence we can utilize the dynamic process $p(\boldsymbol{z}, t)$ on $\boldsymbol{\mu}(t) = t\boldsymbol{z}_{0,i} + (1-t)\boldsymbol{z}_{1,j}$ and $\boldsymbol{\sigma}(t) = \mathbf{0}$ where $t \in [0, 1]$ as:

$$p(\boldsymbol{z}_t, t) = \mathcal{N}(t\boldsymbol{z}_{1,j} + (1-t)\boldsymbol{z}_{0,i}, \mathbf{0}). \tag{6}$$

Meanwhile we can obtain the drift function $\boldsymbol{u}(\boldsymbol{z}, t)$ via using the Proposition 1 as below:

$$\boldsymbol{u}(\boldsymbol{z}_t, t) = \boldsymbol{\mu}'(t) + \frac{\boldsymbol{z}(t) - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}\boldsymbol{\sigma}'(t) = \boldsymbol{z}_{1,j} - \boldsymbol{z}_{0,i}. \tag{7}$$

Specifically, we can employ a neural network $v_\theta(\cdot)$ with matching samples $(\boldsymbol{z}_0, \boldsymbol{z}_1) \sim \boldsymbol{\pi}$ to predict the deterministic drift $u(x, t)$ using the drift-estimate loss function:

$$\min_{\boldsymbol{\theta}} J_{\text{drift}} = \mathbb{E}_{(\boldsymbol{z}_0, \boldsymbol{z}_1)\sim\boldsymbol{\pi}}\left[||\boldsymbol{v}_\theta(\boldsymbol{z}_t, t) - (\boldsymbol{z}_1 - \boldsymbol{z}_0)||_2^2\right]. \tag{8}$$

Moreover, we intend to figure out more consistent results [40] for achieving better performance. That is, the coupling data samples with different time interpolation should have the same drift output as expected. Therefore, we propose drift-consistent loss function:

$$\min_{\boldsymbol{\theta}} J_{\text{CL}} = \mathbb{E}_{t,t'\in U[0,1]}\left[||\boldsymbol{v}_\theta(\boldsymbol{z}_t, t) - \boldsymbol{v}_\theta(\boldsymbol{z}_{t'}, t')||_2^2\right],$$
$$\text{where } \boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, \quad \boldsymbol{z}_{t'} = (1-t')\boldsymbol{z}_0 + t'\boldsymbol{z}_1 \tag{9}$$

Finally, we combine the drift-estimate and drift-consistent loss functions for training our proposed DerODE:

$$\min_{\boldsymbol{\theta}} J_{\text{DerODE}} = J_{\text{drift}} + \lambda_{cl} J_{\text{CL}}, \tag{10}$$

where λ_{cl} denotes the balanced parameter. It is noticeable that DerODE will not involve complex denoising or score estimation procedures during the training stage. Therefore, it could be much easier to train compared with other diffusion approaches. Once we obtain the optimal solution on $\boldsymbol{v}^*(\cdot)$, we can generate new motion features in the latent space via randomly sample noise in the standard Gaussian distribution via:

$$\widetilde{\boldsymbol{z}}_{0,i} = \widetilde{\boldsymbol{z}}_{1,i} - \boldsymbol{v}_\theta(\widetilde{\boldsymbol{z}}_{1,i}, t = 1) = \text{DerODE}(\widetilde{\boldsymbol{z}}_{1,i}), \tag{11}$$

where $\widetilde{\boldsymbol{z}}_{1,i} \sim \mathcal{N}(0, \boldsymbol{I})$ and it can obtain the deterministic output result $\widetilde{\boldsymbol{z}}_{0,i}$. Finally, we can utilize the decoder $\text{Dec}(\cdot)$ to generate human motion as $\widetilde{\boldsymbol{E}} = \text{Dec}(\widetilde{\boldsymbol{z}}_{0,i})$ accordingly.

4.2.2. Stochastic Diverse Output Generation Procedure

Although we have obtained the deterministic ordinary differential equations (DerODE) between the latent space distribution of human motions and the standard Gaussian distribution, it remains challenging to generate highly diverse motion patterns. This difficulty arises from the deterministic nature of the ODEs, as identical initial conditions result in the same output paths, thereby reducing the diversity of the generated samples. To provide more diverse while accurate human motions, we tend to involve the stochastic differential equations based on the ordinary differential equations in the stochastic diverse output generation procedure.

Proposition 2. *Given the stochastic differential equations $d\boldsymbol{z}_t = f(\boldsymbol{z}_t, t)dt + g(t)d\boldsymbol{w}_t$ with the drift and diffusion terms, the mean $\boldsymbol{\mu}(t)$ and covariance $\boldsymbol{\Sigma}(t)$ can be formulated as:*

$$\begin{cases} \frac{d\boldsymbol{\mu}(t)}{dt} = \mathbb{E}[f(\boldsymbol{z}, t)] \\ \frac{d\boldsymbol{\Sigma}(t)}{dt} = \mathbb{E}\left[f(\boldsymbol{z}, t)(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \\ \qquad + \mathbb{E}\left[(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))f^\top(\boldsymbol{z}, t)\right] + \mathbb{E}[g^2(t)]. \end{cases} \tag{12}$$

The proof of the Proposition 2 can be found in Appendix B. We can observe that the stochastic differential equations can transform the distributions according to the specific settings of drift and diffusion terms, which leads to diverse output results based on Proposition 2. Therefore, it is intuitive to consider a proper stochastic differential equations with carefully designed $f(\boldsymbol{z}_t, t)$ and $g(t)$ respectively in the stochastic diverse output generation procedure.

Proposition 3. *Given the Diverse Stochastic Differential Equations (DivSDE) as $d\boldsymbol{z}_t = \left(-\frac{1}{1-t}\right)\boldsymbol{z}_t dt + \eta\sqrt{\frac{2t}{1-t}}d\boldsymbol{w}_t$ with the initial data sample $\boldsymbol{x}_0$ and the noise level η, the probability of data distribution $\boldsymbol{x}_t$ is $p(\boldsymbol{z}_t) = \mathcal{N}((1-t)\boldsymbol{z}_i, \eta^2 t^2 \boldsymbol{I})$ at the time step t when $p(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_0, \mathbf{0})$.*

The proof of the Proposition 3 can be found in Appendix C. It is obvious that the diffusion term $\eta\sqrt{\frac{2t}{1-t}}d\boldsymbol{w}_t$ which involves noise can enhance the diversity of the model output and thus DivSDE is different than DerODE. Note that the stochastic differential equations in Proposition 2 have the backward process as $d\boldsymbol{x}_t = \left[-\frac{1}{1-t}\boldsymbol{z}_t - \frac{2t}{1-t}\nabla\log p(\boldsymbol{z}_t)\right]dt + \eta\sqrt{\frac{2t}{1-t}}d\boldsymbol{w}_t$, where $\nabla\log p(\boldsymbol{z}_t)$ denotes the score function of the data probability. Specifically, $\nabla\log p(\boldsymbol{z}_t)$ can be calculated via $\nabla\log p(\boldsymbol{z}_t) = \frac{(1-t)\boldsymbol{z}_i - \boldsymbol{z}_t}{t^2}$. Previous score-based approaches [39] may involve a new neural network to estimate $\nabla\log p(\boldsymbol{z}_t)$ even if it is rather time-consuming and hard to train in real practice. However, it is important to note that we can already obtain $\widetilde{\boldsymbol{z}}_{0,i}$ by utilizing DerODE via $\widetilde{\boldsymbol{z}}_{0,i} = \text{DerODE}(\widetilde{\boldsymbol{z}}_{1,i})$ and it can be further utilized for DivSDE. Therefore, we can rewrite the discrete form of the backward process on DivSDE as follows:

$$\begin{aligned} \boldsymbol{z}_{i,t} = \boldsymbol{z}_{t+\Delta t,i} &+ \frac{\Delta t}{1-t}\boldsymbol{z}_{t+\Delta t,i} \\ &+ \frac{2t\Delta t}{1-t}\frac{(1-t)\widetilde{\boldsymbol{z}}_{0,i} - \boldsymbol{z}_{t,i}}{t^2} + \eta\varepsilon\sqrt{\frac{2t}{1-t}}\sqrt{\Delta t}, \end{aligned} \tag{13}$$

where $\varepsilon \sim \mathcal{N}(0, \boldsymbol{I})$ denotes the randomly sample noise. Meanwhile, η denotes the strengths of diversity. That is, larger value of η will provide more diverse output human motions. Moreover, DivSDE can directly borrow the previously calculated results from DerODE for secondary computations without the need for re-introducing other training processes.

4.3. Model Summary

In summary, the proposed DSDFM can synthesize diverse and accurate human motion sequences through the designed two stages, i.e., *human motion reconstruction* and *diverse motion generation*. In the human motion reconstruction

Algorithm 1 The process for generating diverse human motions.

Require: time interval: T, time steps: $\Delta t = \frac{1}{T}$, noise for diversity: η

Ensure: Generated new samples $\hat{E}$.

Initialize $\widetilde{\boldsymbol{z}}_{1,i}$ from Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$.

Adopting DerODE to obtain $\widetilde{\boldsymbol{x}}_{i,0}$.

$\widetilde{\boldsymbol{z}}_{0,i} = \mathrm{DerODE}(\widetilde{\boldsymbol{z}}_{1,i})$

Adopting DivSDE to obtain diverse human motions.

for $t \in range(T - \Delta t, 0)$ **do**

  Obtain the score function as: $\nabla \log p(\boldsymbol{z}_{t,i}) = \frac{(1-t)\widetilde{\boldsymbol{z}}_{0,i} - \boldsymbol{z}_{t,i}}{(t/T)^2}$.

  Obtain the diffusion term as: $D_{\mathrm{Diffu}} = \frac{2(t/T)}{1-(t/T)} \nabla \log p(\boldsymbol{z}_{t,i}) \Delta t$

  Obtain the drift term as: $D_{\mathrm{Drift}} = \frac{\Delta t}{1-(t/T)} z_{t+\Delta t,i}$

  Obtain the noise term as: $\epsilon_{\mathrm{noise}} = \eta \cdot \varepsilon \sqrt{\frac{2(t/T)}{1-(t/T)}} \sqrt{\Delta t}$

  Obtain $\boldsymbol{z}_{t,i} = \boldsymbol{z}_{t+\Delta t,i} + D_{\mathrm{Diffu}} + D_{\mathrm{Drift}} + \epsilon_{\mathrm{noise}}$

end for

Generate the human motion: $\hat{E} = \mathrm{Dec}(\boldsymbol{z}_{0,i})$

stage, we first adopt the human motion reconstruction network to learn a well-structured latent space of human motions through VQVAE network. In the diverse motion generation stage, we tend to build the connections between the Gaussian distribution and latent space of human motions, thereby enhancing the diversity while guaranteeing the accuracy of the generated human motions through the designed deterministic feature mapping procedure with DerODE and stochastic diverse output generation procedure with DivSDE. Specifically, DerODE can provide deterministic output results in an efficient way. Meanwhile, DivSDE can obtain more diverse human motions without introducing additional training process. The pseudo algorithm of the DSDFM is provided in Algorithm 1.

5. Experiment

In this section, we provide extensive experiments to evaluate the performance of our proposed DSDFM across widely used human motion datasets. We first describe the utilized human motion datasets and implementation details (Section. 5.1). Subsequently, we present a comparative results analysis of our method with other state-of-the-art approaches on conditional and unconditional human motion synthesis. Additionally, we provide ablation studies to assess the effectiveness of the modules in our method (Section 5.2). Finally, we visually showcase the generated diverse human motion sequences to provide a qualitative performance (Section 5.3).

5.1. Datasets and Implementation Details

Datasets. The experiments are conducted on two widely used motion capture datasets, i.e., HumanAct12 [9], and HumanML3D [10]. **HumanAct12** provides 1,191 raw motion sequences, and contains 12 subjects in which 12 categories of actions with per-sequence annotation are provided. **HumanML3D** dataset is a recent dataset that contains 14,616 motion sequences annotated by 44970 textual descriptions obtained from AMASS [29].

Table 1. The comparison results of unconditional human motion synthesis between our method and state-of-the-art methods on HumanAct12 dataset. **Bold** and underline indicate the best and the second best result.

Method	FID ↓	KID ↓	Precision ↑	Recall ↑	Diversity ↑	#params
VPoser (CVPR'19)	48.65	0.72	0.68	0.72	12.75	29M
Action2Motion (MM'21)	49.76	0.68	0.70	0.71	13.80	21M
ACTOR (CVPR'21)	48.80	0.53	0.72	0.74	14.10	20M
MDM (ICLR'23)	31.92	0.96	0.66	0.62	17.00	24M
MLD (CVPR'23)	14.25	0.55	0.70	0.79	16.85	27M
Modi (CVPR'23)	13.03	0.12	0.71	0.81	17.57	23M
DSDFM (Ours)	**12.86**	**0.10**	**0.75**	**0.85**	**18.41**	**15M**
Improvement	1.31%	1.67%	4.17%	4.93%	4.78%	2.50%

Evaluation metrics. For a fair comparison, our method employs the following evaluation metrics: Frechet Inception Distance (FID), Kernel Inception Distance (KID), Precision, Recall, Accuracy, Diversity, and Multimodality. FID is the distance between the feature distribution of generated motions and that of the real motions, namely the difference in mean and variance. KID compares skewness as well as the values compared in FID, namely mean and variance. Precision measures the probability that a randomly generated motion falls within the support of the distribution of real data. Recall measures the probability that a real motion falls within the support of the distribution of generated data. Accuracy is measured by the corresponding action recognition model. Diversity measures the variance of the whole motion sequences across the dataset. Multimodality measures the diversity of human motion generated from the same text description. A lower value implies better for FID and KID. Higher Precision, Recall, Accuracy, Diversity, and Multi-Modality values imply better results. FID, KID, Precision, Recall, and Accuracy are utilized to evaluate the generated human motion accuracy. Diversity and MultiModality are utilized for the generation diversity.

Implementation Details. For the human motion reconstruction process, the VQVAE consists of 4 Transformer layers with 8 heads, and the codebook size is set to 512 $\times$ 512. The batch size is set to 128, learning rate is initially set to 10^{-2} with a 0.98 decay every 10 epochs. The proposed method is trained for 500 epochs. For the diverse motion generation process, the time interval Δt is set to 0.01, and the strength of diversity η is set to 0.1. The diffusion step is set to 100. The balanced parameter λ_{cl} for J_{CL} loss is set to 0.3.

5.2. Experimental Results

Comparisons on Unconditional Human Motion Synthesis. We compare our method DSDFM with other state-of-the-art methods under the unconditional generation settings on the

Table 2. The comparison results of Action-to-Motion task on HumanAct12 dataset. $\pm$ indicates 95% confidence interval, $\rightarrow$ indicates that closer to real is better. The best results are in bold.

Method	FID ↓	Accuracy ↑	Diversity →	Multimodality ↑	#params
Real	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$	-
Action2Motion (MM'21)	$0.338^{\pm.015}$	$0.917^{\pm.003}$	$6.879^{\pm.066}$	$2.511^{\pm.023}$	21M
ACTOR (CVPR'21)	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$6.840^{\pm.030}$	$2.530^{\pm.020}$	20M
INR (ECCV'22)	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$	25M
MLD (CVPR'23)	$0.077^{\pm.004}$	$0.964^{\pm.002}$	$6.831^{\pm.050}$	$2.824^{\pm.0.38}$	27M
MDM (ICLR'23)	$0.100^{\pm.000}$	$0.990^{\pm.000}$	$6.860^{\pm.050}$	$2.520^{\pm.010}$	24M
MotionDiffuse (TPAMI'24)	$0.070^{\pm.000}$	$0.992^{\pm.013}$	$\mathbf{6.850}^{\pm.020}$	$2.460^{\pm.020}$	25M
DSDFM (Ours)	$\mathbf{0.068}^{\pm.010}$	$\mathbf{0.994}^{\pm.001}$	$6.851^{\pm.008}$	$\mathbf{2.455}^{\pm.025}$	**15M**
Improvement	2.85%	0.21 %	-0.01%	0.21%	2.50%

Table 3. Ablation study on the comparison results of training and inference time on the HumanAct12 dataset. m denotes minute, s denotes second.

Method	Epoch	Training Time (m)	Inference Time (s)/FID		
			100 steps/FID	500 steps/FID	1000 steps/FID
VPSDE	500	42.93	2.54/16.74	9.93/15.63	18.09/14.31
VESDE	500	40.57	2.68/16.49	9.48/14.92	16.12/14.17
DSDFM(Ours)	500	25.33	1.60/13.61	5.03/12.86	10.33/12.24

HumanAct12 dataset, the results are shown in Table 1. The input of the baseline methods is modified to the same length as our method. From the comparison results, we can observe that the baseline methods report poor performance in terms of accuracy and diversity metrics due to the mode collapse or unstable training processes. DSDFM outperforms these methods owing to the designed diverse motion generation procedure. In addition, to assess the training efficiency of our method, we also investigate the number of training parameters. The comparison results show that our method utilizes fewer parameters than baseline methods while achieving superior performance, which demonstrates the effectiveness of the proposed method. This suggests that our method is more computationally efficient and achieves the balance between the accuracy and diversity of generated samples.

Comparisons on Conditional Human Motion Synthesis. Our method can also be extended to conditional generation, i.e., Action-to-Motion task. This task involves generating relevant human motion sequences given an input action label. The comparison results on the HumanAct12 dataset are presented in Table 2. From the comparison results, we can observe that DSDFM also achieves comparable performance under the accuracy and diversity metrics. Our method performs slightly worse than MotionDiffuse method in terms of the diversity metric by 0.01%, but our method reduces the confidence interval, demonstrating that our method is more stable and reliable. Additionally, it significantly decreases the number of training parameters. These results further report the effectiveness of our method for conditional human motion generation.

5.3. Ablation studies

To report the effectiveness of each component of our method, we compare the baseline methods with DSDFM under different settings on the HumanML3D and HumanAct12 datasets, including the training time, inference time for different diffusion steps, and the corresponding FID. The results are shown in Table 3 and Table 4. Table 3 shows the comparison results for unconditional motion generation on the HumanAct12 dataset. From the results, we can observe that, compared to VPSDE and VESDE given the same number of epochs, DSDFM significantly reduces the training time while achieving a comparable performance under the FID metric, which demonstrates that our method is easier to train than the baseline methods and guaranteeing the quality of generated human motions. We also test these methods under different diffusion steps, and the performance of our method is improved in inference time. In addition, Table 4 shows the comparison results for Action-to-Motion task on HumanML3D dataset. The results under the same metrics are consistent with the results on the HumanAct12 dataset, which further demonstrates the effectiveness of our method.

Table 4. Ablation study on the comparison results of training and inference time on the HumanML3D dataset.

Method	Epoch	Training Time (m)	Inference Time (s)/FID		
			100 steps/FID	500 steps/FID	1000 steps/FID
VPSDE	500	12.54	2.47/ $0.092^{\pm.003}$	4.95/ $0.088^{\pm.002}$	6.51/ $0.080^{\pm.024}$
VESDE	500	12.57	2.35/ $0.094^{\pm.005}$	4.48/ $0.089^{\pm.001}$	6.62/ $0.078^{\pm.013}$
DSDFM(Ours)	500	7.02	1.01/$0.073^{\pm.005}$	2.15/$0.068^{\pm.008}$	4.82/$0.054^{\pm.010}$

Table 5. Ablation studies of the proposed method. We compare our method with other score-based methods and provide the comparison results under the accuracy and diversity metrics, as well as the number of training parameters.

Method	FID ↓	KID ↓	Precision ↑	Recall ↑	Diversity ↑	#param
VESDE	14.92	0.36	0.59	0.65	16.21	28M
VPSDE	15.63	0.29	0.64	0.68	17.00	24M
SDE (DDPM++)	13.25	0.21	0.68	0.75	17.46	22M
SDE (NCSN++)	13.01	0.19	0.72	0.79	17.54	21M
DSDFM (Ours)	**12.86**	**0.10**	**0.75**	**0.85**	**18.41**	**15M**

The ablation studies also test the performance of the designed stochastic diverse output generation procedure in DSDFM under the diversity and accuracy metrics, the results are shown in Table 5. Specifically, we employ other score-based methods to enhance the diversity of generated human motion sequences, i.e., variance preserving SDE (VPSDE), variance exploding SDE (VESDE), DDPM++, and NCSN++. From the comparison results, we can observe that our method exhibits comparable performance in terms of accuracy compared to the baseline methods, while showing a slight improvement in diversity. Notably, we have achieved a significant reduction in the number of training parameters, which report the effectiveness of our method.

5.4. Visualization

In this section, we show the visualization results of our method on the unconditional human motion synthesis and Action-to-Motion tasks. As depicted in Figure 3, the top is the unconditional human motion synthesis, all human motion sequences are unconditionally generated from random noise sampled from Gaussian distribution on the HumanAct12 dataset. The figure shows that our method can generate diverse and high fidelity human motion sequences. The bottom is the sequences generated by the Action-to-Motion task, the generated sequences are under the action labels on

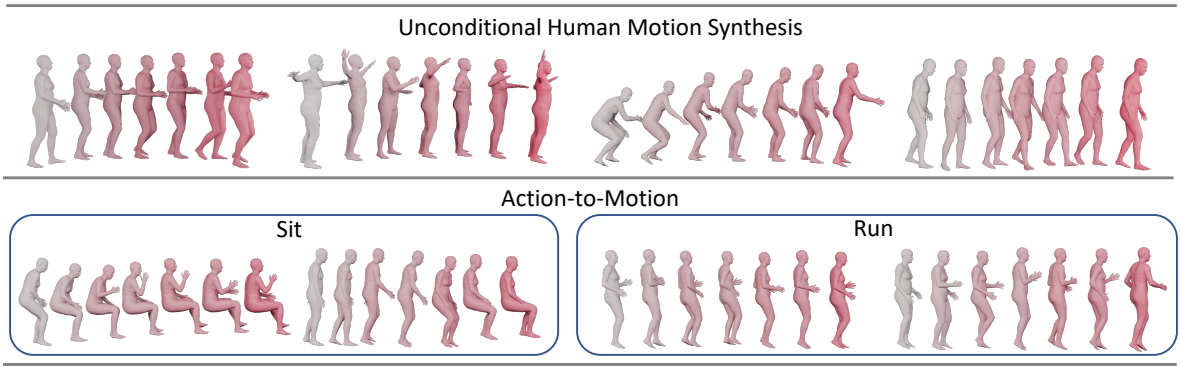

Figure 3. Qualitative results of DSDFM. We present the generated human motion sequences under different settings. The unconditional human motion sequences (top) are generated from the HumanAct12 dataset. The Action-to-Motion results (bottom) show the generated diverse motion sequences under the Sit and Run action labels, which are sampled from the HumanML3D dataset.

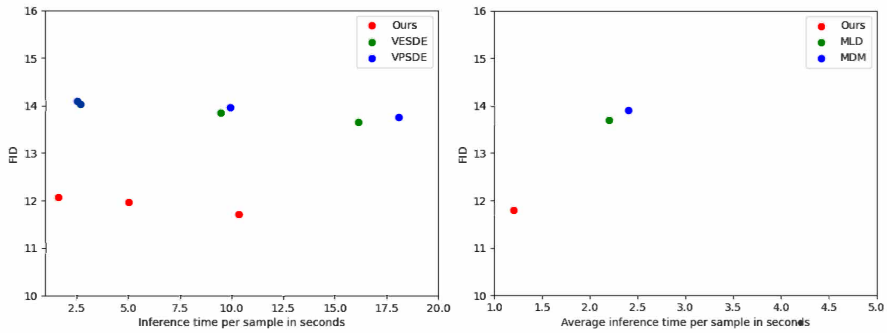

Figure 4. Comparison of the inference time costs of our method on the HumanAct12 dataset. We calculate the ablation studies (left) and average inference time comparison with baselines (right). All tests are performed on the NVIDIA A100.

the HumanML3D dataset. We can observe that the generated diverse motion sequences match the descriptions well. These qualitative results demonstrate that DSDFM can generate diverse and coherent human motion sequences.

We compare and visualize the comparison results of inference time in Figure 4. The left of this Figure is the ablation studies of our method with different diffusion steps. This figure shows that using VPSDE and VESDE as our backbone has long inference time and relatively low accuracy. The right of this Figure is the average inference time comparison with baselines, which shows that our method can speed up the inference time when generating new samples.

6. Conclusion

In this paper, we propose a Deterministic-to-Stochastic Diverse Latent Feature Mapping (DSDFM) for human motion synthesis. DSDFM is easy to train compared with the recent SGMs-based method, while facilitating the diversity and accuracy of generated human motions. DSDFM includes two stages, human motion reconstruction and diverse motion generation. Human motion reconstruction aims to learn a well-structured latent space of human motions. Diverse motion generation aims to enhance the diversity of the generated human motion sequences through the designed deterministic feature mapping procedure with DerODE and stochastic diverse output generation procedure with DivSDE. Extensive experimental results demonstrate the efficacy of the proposed DSDFM method for human motion synthesis.

7. Acknowledgment

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-031), the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No.: AISG2-GC-2023-010, "Design Beyond What You Know": Material-Informed Differential Generative AI (MIDGAI) for Light-Weight High-Entropy Alloys and Multi-functional Composites (Stage 1a)", the National Natural Science Foundation of China under Grant 62372081, the Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001, the Liaoning Provincial Natural Science Foundation Program under Grant 2024010785-JH3107, the Dalian Science and Technology Innovation Fund under Grant 2024JJ12GX020, the Dalian Major Projects of Basic Research under Grant 2024JJ12GX020 2023JJ11CG002 and the 111 Project under Grant D23006.

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 1

[2] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *European Conference on Computer Vision*, pages 356–372. Springer, 2022. 2

[3] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[5] Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23263–23274, 2023. 2, 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[7] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 927–936, 2024. 1

[8] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020. 3

[9] 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. 6, 13

[10] 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. 6, 13

[11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[13] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[14] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Wei Yang, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[15] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021. 1

[16] 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

[17] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2(3), 2021. 1

[18] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, pages 25268–25280. Curran Associates, Inc., 2023. 1

[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[20] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 4

[21] Weiming Liu, Xiaolin Zheng, Jiajie Su, Mengling Hu, Yanchao Tan, and Chaochao Chen. Exploiting variational domain-invariant user embedding for partially overlapped cross domain recommendation. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pages 312–321, 2022. 4

[22] Weiming Liu, Xiaolin Zheng, Chaochao Chen, Jiajie Su, Xinting Liao, Mengling Hu, and Yanchao Tan. Joint internal multi-interest exploration and external domain alignment for cross domain sequential recommendation. In *Proceedings of the ACM web conference 2023*, pages 383–394, 2023.

[23] Weiming Liu, Xiaolin Zheng, Jiajie Su, Longfei Zheng, Chaochao Chen, and Mengling Hu. Contrastive proxy kernel stein path alignment for cross-domain cold-start recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11216–11230, 2023.

[24] Weiming Liu, Chaochao Chen, Xinting Liao, Mengling Hu, Jiajie Su, Yanchao Tan, and Fan Wang. User distribution mapping modelling with collaborative filtering for cross domain recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 334–343, 2024.

[25] Weiming Liu, Chaochao Chen, Xinting Liao, Mengling Hu, Yanchao Tan, Fan Wang, Xiaolin Zheng, and Yew Soon Ong. Learning accurate and bidirectional transformation via dynamic embedding transportation for cross-domain recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8815–8823, 2024.

[26] Weiming Liu, Xiaolin Zheng, Chaochao Chen, Jiahe Xu, Xinting Liao, Fan Wang, Yanchao Tan, and Yew-Soon Ong.

Reducing item discrepancy via differentially private robust embedding alignment for privacy-preserving cross domain recommendation. In *Forty-first International Conference on Machine Learning*, 2024. 4

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2, 3

[28] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 1

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6

[30] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021. 3

[31] Odena et al. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017. 1

[32] 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

[33] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 1

[34] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 4

[35] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13873–13883, 2023. 1, 2

[36] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 13

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[38] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 3

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS. 2021. 1, 2, 3, 4, 5

[40] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 5

[41] Jiangxin Sun, Zihang Lin, Xintong Han, Jian-Fang Hu, Jia Xu, and Wei-Shi Zheng. Action-guided 3d human motion prediction. *Advances in Neural Information Processing Systems*, 34:30169–30180, 2021. 2

[42] Zitang Sun, Yen-Ju Chen, Yung-Hao Yang, and Shin' ya Nishida. Modeling human visual motion processing with trainable motion energy sensing and a self-attention network. In *Advances in Neural Information Processing Systems*, pages 24335–24348. Curran Associates, Inc., 2023. 2

[43] Vanessa Tan, Junghyun Nam, Juhan Nam, and Junyong Noh. Motion to dance music generation using latent diffusion model. In *SIGGRAPH Asia 2023 Technical Communications*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[44] Jianwei Tang, Jiangxin Sun, Xiaotong Lin, Wei-Shi Zheng, Jian-Fang Hu, et al. Temporal continual learning with prior compensation for human motion prediction. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model, 2022. arXiv:2209.14916 [cs]. 2

[46] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021. 1, 3

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 4

[48] 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. 4

[49] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[50] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems*, pages 14959–14971. Curran Associates, Inc., 2022. 2

[51] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6110–6118, 2023. 1

[52] Lemeng Wu, Dilin Wang, Chengyue Gong, Xingchao Liu, Yunyang Xiong, Rakesh Ranjan, Raghuraman Krishnamoorthi, Vikas Chandra, and Qiang Liu. Fast point cloud generation with straight flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9445–9454, 2023. 3

[53] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, pages 265–281, 2018. 1

[54] Hua Yu, Xuanzhe Fan, Yaqing Hou, Wenbin Pei, Hongwei Ge, Xin Yang, Dongsheng Zhou, Qiang Zhang, and Mengjie

Zhang. Toward realistic 3d human motion prediction with a spatio-temporal cross- transformer approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 5707–5720, 2023. 2

[55] Hua Yu, Yaqing Hou, Wenbin Pei, Yew-Soon Ong, and Qiang Zhang. Divdiff: A conditional diffusion model for diverse human motion prediction. *IEEE Transactions on Multimedia*, pages 1–12, 2024. 2

[56] Hua Yu, Weiming Liu, Jiapeng Bai, Xu Gui, Yaqing Hou, YewSoon Ong, and Qiang Zhang. Towards efficient and diverse generative model for unconditional human motion synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 2535–2544, New York, NY, USA, 2024. Association for Computing Machinery. 2

[57] Yuan et al. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 1

[58] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023. 1

[59] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023. 2

[60] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 1

[61] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 1