Executing your Commands via Motion Diffusion in Latent Space

Xin Chen[1*] Biao Jiang[2*] Wen Liu[1] Zilong Huang[1] Bin Fu[1]
Tao Chen[2] Gang Yu[1†]
[1]Tencent PCG [2]Fudan University
https://github.com/chenfengye/motion-latent-diffusion

Abstract

We study a challenging task, conditional human motion generation, which produces plausible human motion sequences according to various conditional inputs, such as action classes or textual descriptors. Since human motions are highly diverse and have a property of quite different distribution from conditional modalities, such as textual descriptors in natural languages, it is hard to learn a probabilistic mapping from the desired conditional modality to the human motion sequences. Besides, the raw motion data from the motion capture system might be redundant in sequences and contain noises; directly modeling the joint distribution over the raw motion sequences and conditional modalities would need a heavy computational overhead and might result in artifacts introduced by the captured noises. To learn a better representation of the various human motion sequences, we first design a powerful Variational AutoEncoder (VAE) and arrive at a representative and low-dimensional latent code for a human motion sequence. Then, instead of using a diffusion model to establish the connections between the raw motion sequences and the conditional inputs, we perform a diffusion process on the motion latent space. Our proposed Motion Latent-based Diffusion model (MLD) could produce vivid motion sequences conforming to the given conditional inputs and substantially reduce the computational overhead in both the training and inference stages. Extensive experiments on various human motion generation tasks demonstrate that our MLD achieves significant improvements over the state-of-the-art methods among extensive human motion generation tasks, with two orders of magnitude faster than previous diffusion models on raw motion sequences.

1. Introduction

Human motion synthesis has recently rapidly developed in a multi-modal generative fashion. Various condition inputs, such as music [34, 33, 32], control signals [45, 66, 65], action categories [46, 19], and natural language descriptions [16, 47, 69, 18, 2, 28], provide a more convenient and human-friendly way to animate virtual characters or even control humanoid robots. It will benefit numerous applications in the game industry, film production, VR/AR, and robotic assistance.

Figure 1. Our Motion Latent-based Diffusion (MLD) model can achieve high-quality and diverse motion generation given a text prompt. The darker colors indicate the later in time, and the colored words refer to the motions with same colored trajectory.

Among all conditional modalities, text-based conditional human motion synthesis has been driving and dominating research frontiers because the language descriptors provide a convenient and natural user interface for people to interact with computers [47, 2, 75, 69, 28]. However, since the distributions between the natural language descriptors and

[*]These authors contributed equally to this work.
[†]Corresponding author.

motion sequences are quite different, it is not easy to learn a probabilistic mapping function from the textural descriptors to the motion sequences, which is also mentioned in the previous work, MotionCLIP [68]. Two typical methods address this problem: 1) the cross-modal compatible latent space between motion and language [47, 2] and 2) the conditional diffusion model [75, 69, 28]. The formers, such as TEMOS [47], usually learn a motion Variational AutoEncoder (VAE) and a text Variational Encoding (without decoder) and then constrain the text encoder and the motion encoder into a compatible latent space via the Kullback-Leibler (KL) divergences loss, which pushes a foundational step forward on creating human motion sequences by natural language inputs. However, since the distributions of natural languages and motion sequences are highly different, forcibly aligning these two simple gaussian distributions, in terms of variational text encoding and variational motion encoding, into a compatible distribution might result in misalignments and thereby reduce the generative diversity inevitably. In light of the tremendous success of the diffusion-based generative models on other domains [53, 61, 56, 22, 79, 73], the latter category methods [75, 69, 28] propose a conditional diffusion model for human motion synthesis to learn a more powerful probabilistic mapping from the textual descriptors to human motion sequences and improve the synthesized quality and diversity. Nevertheless, the raw motion sequences are somewhat time-axis redundant, and diffusion models in raw sequential data [54, 22, 35] usually require exhausting computational overhead in both the training and inference phase, which is inefficient. Besides, since the raw motion data from the motion capture system might contain noises, the powerful diffusion models might learn the clues of a probabilistic mapping from the conditional inputs to the noise motion sequences and produce artifacts.

To efficiently synthesize plausible and diverse human motion sequences according to the conditional inputs, inspired by the success of the diffusion model on latent space in text-to-image synthesis [56], we combine the advantages of the latent space-based and the conditional diffusion-based methods and propose a motion latent-based diffusion model (MLD) for human motion generation. Specifically, we first design a transformer-based autoencoder [46] with the UNet-like long skip connections [59] to learn a representative and low-dimensional latent distribution of human motion sequences. Then, instead of using a diffusion model to establish the connections between the raw motion sequences and the conditional inputs, we propose a motion latent-based diffusion model (MLD) to learn a better probabilistic mapping from the conditions to the representative motion latent codes, which could not only produce the vivid motion sequences conforming to the given conditional inputs but also substantially reduce the computational overhead in both training and inference stage. In addition, high-quality human motion sequences with well-annotated action labels or textual descriptions are expensive and limited. In contrast, the large-scale non-annotated or weakly annotated motion sequences are publicly available, such as the AMASS dataset [41]. Our proposed MLD could individually train a motion latent autoencoder on these large-scale datasets, arriving at a representative and low-dimensional latent space for diverse human motion sequences. This low-dimensional latent space with higher information density could accelerate the model's convergence and significantly reduce computational consumption for the downstream conditional human motion generation tasks.

We summarize the contributions as follows: 1) we design and explore a more representative motion variational autoencoder (VAE), which provides state-of-the-art motion reconstruction and diverse generation, benefiting the training of the latent diffusion models; 2) we further demonstrate that motion generation tasks on latent spaces, such as text-to-motion and action-to-motion, are more efficient than the diffusion models on raw motion sequences; 3) our proposed MLD achieves competitive performance on multiple tasks (unconditional motion generation, action-to-motion, and text-to-motion), and codes are available.

2. Related Work

Human Motion Synthesis allows rich inputs of multimodal data, such as text [16, 47, 69, 18, 2, 28], action category [46, 19], incomplete pose sequences [12, 20, 69], control signals [65, 66, 45], musics [34, 33, 32] and image(s) [55, 9], here, we focus on some typical tasks. Firstly, **unconditional motion generation** [74, 78, 76, 51, 69] is a more universal task, which models the entire motion space, only needs motion data without any requirement of annotation, and benefits other generation tasks. VPoser [43] proposes a variational human pose prior mainly for image-based pose fitting. ACTOR [46, 47] recently proposes a class-agnostic transformer VAE as one baseline. After that, among all conditional tasks, **text-to-motion** [47, 2, 75, 69, 28, 16] has been driving and dominating research frontiers because the language descriptors are the most user-friendly and convenient. More recently, two categories of motion synthesis methods have emerged, joint-latent models [47, 2] and diffusion models [75, 69, 28]. The former category, like TEMOS [47], proposes a VAE architecture to learn a joint latent space of motion and text constrained on a Gaussian distribution. However, natural language and human motions are quite different with misaligned structure and distribution, thus it is difficult to forcibly align two simple Gaussian distributions [68]. Lastly, we introduce **action-to-motion** [46, 19], a reverse problem of the classical action recognition task. ACTOR [46] proposes learnable biases in transformer VAE to embed action for motion generation.

However, most above methods can only handle one task and hardly change condition inputs. We address this problem by separating models into a universal motion generative model and latent diffusion models to handle different motion generation tasks.

Motion data is critical in the development of motion synthesis tasks. Thanks to the marker-based and markless motion capture approaches [31, 72, 21, 8], they provide convenient and effective solutions for large raw motion data collection. KIT Motion-Language [48] annotates sequence-level description for motions from [42], and HumanML3d [17] provide more textual annotation for some motions of AMASS [41]. They are also our focus in the text-to-motion task. For the action-to-motion datasets, Babel [50] also collects motions from AMASS and provides action and behavior annotations. ACTOR [46] use [31] to process two action recognition datasets, HumanAct12 [19] and UESTC [26], for action-to-motion task.

Motion Representation. These datasets lead to the discussion about motion representation, such as the straightforward joint positions and the Master Motor Map (MMM) format [67]. For our setting, we employ two motion representations: 1) the classical SMPL-based [40, 31, 8] motion parameters and 2) the redundant hand-crafted motion feature [17, 66, 65] with a combination of joints features. The former is widely used in motion capture, and the latter is mainly used in character animation. As suggested by [17], we use the latter in most of our synthesis framework to avoid foot-sliding issues, and use the SMPL parameters for the action-based tasks for a fair comparison with other approaches. Besides, we also recognize the latent in Sec. 5 as one of motion representation.

Generative Models play an important role in motion synthesis tasks to generate high-quality human motion, Although motion generative models, like VAEs [46, 39, 19] and Generative Adversarial Networks (GAN) [38, 1], can enable effective human motion sampling, recent studies [3, 15, 19, 46] recommend VAEs rather than GANs since the latter are more difficult to train. We follow their suggestions and employ VAEs to compress and reconstruct human motion for the learning of diffusion models. We next introduce the diffusion models, especially in motion domain.

Diffusion Generative Models. Diffusion Generative Models [63] achieve significant success in the image synthesis domain, such as Imagen [61], DALL·E 2 [53] and Stable Diffusion [56]. Inspired by their works, most recent methods [69, 75, 28] leverage diffusion models for human motion synthesis. MotionDiffuse [75] is the first text-based motion diffusion model with fine-grained instructions on body parts. MDM [69], most recently, proposes a motion diffusion model on raw motion data to learn the relation between motion and input conditions. However, these diffusion models are not very applicable to raw motion data with potential noise and temporal consistency redundancy and thus are easily misdirected by outliers. In addition, directly applying the diffusion model [69, 75] to the raw motion data suffers from high computational overheads and low inference speed. Inspired by [56], we propose a motion latent-based diffusion model to reduce computational resources and improve the generative quality.

3. Method

To efficiently generate high-quality and diverse human motion sequences according to desired conditional inputs with fewer computational overheads, we propose to perform a diffusion process on a representative and low-dimensional motion latent space and consequently arrive at a motion latent-based diffusion model (MLD) for conditional human motion synthesis. It contains a motion Variational AutoEncoder (VAE) to learn a representative and low-dimensional latent space for diverse human motion sequences (details in Sec. 3.1) and a conditional diffusion model in this latent space (details in Sec. 3.2 and Sec. 3.3).

The conditions include action labels, textual descriptions, or even empty conditions. Specifically, given an input condition c, such as a sentence $\boldsymbol{w}^{1:N} = \{w^i\}_{i=1}^N$ describing a motion [47], a action label a from the predefined action categories set $a \in A$ [46] or even a empty condition $c = \varnothing$ [43, 77], our MLD aims to generate a human motion $\hat{x}^{1:L} = \{\hat{x}^i\}_{i=1}^L$ in a non-deterministic way, where L denotes the motion length or frame number. Here, we employ the motion representation in [17]: a combination of 3D joint rotations, positions, velocities, and foot contact. In addition, we propose the motion encoder $\mathcal{E}$ to encode the motion sequences, $x^{1:L} = \{x^i\}_{i=1}^L$, into a latent $z = \mathcal{E}(x^{1:L})$, and decode z into the motion sequences using a motion decoder $\mathcal{D}$, that is $\hat{x}^{1:L} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x^{1:L}))$.

3.1. Motion Representation in Latents

We build our motion Variational AutoEncoder, $\mathcal{V}$, based on a transformer-based architecture [46], which consists of a transformer encoder $\mathcal{E}$ and a transformer decoder $\mathcal{D}$. The motion VAE, $\mathcal{V} = \{\mathcal{E}, \mathcal{D}\}$, is trained by the motion $x^{1:L}$ reconstruction only with the Mean Squared Error (MSE) loss and the Kullback-Leibler (KL) loss. We further enhance two transformers [70] of $\mathcal{E}$ and $\mathcal{D}$ with long skip connections [59], and remove the action biases used in [46]. The encoder could produce a representative, low-dimensional latent space with high informative density, and the decoder could well reconstruct the latent into motion sequences.

More specifically, the motion encoder $\mathcal{E}$ takes as input learnable distribution tokens, and frame-wise motion features $x^{1:L}$ of arbitrary length L. We use the embedded distribution tokens as Gaussian distribution parameters μ and σ of the motion latent space $\mathcal{Z}$ to reparameterize [30] latent $z \in \mathbb{R}^{n \times d}$ whose dimension is similar to [46]. The motion decoder $\mathcal{D}$ relies on the architecture of the transformer decoder with cross attention mechanism, which takes the L number of zero motion tokes as queries, a latent $z \in \mathbb{R}^{n\times d}$ as memory, and finally, generates a human motion sequence $\hat{x}^{1:L}$ with L frames.

According to [47], both the latent space $\mathcal{Z}$ and variable durations help the model to produce more diverse motions. To further enhance the latent representation, we leverage a long skip-connection structure for the transformer-based encoder $\mathcal{E}$ and decoder $\mathcal{D}$. We also explore the effectiveness of the latent's dimensions on motion sequences representation in Tab. 4. Hence, our VAE models present a stronger motion reconstruction ability and richer diversity (cf. Tab. 5 and Tab. 6). We provide more details about the architecture and the training in the supplementary.

3.2. Motion Latent Diffusion Model

Diffusion probabilistic models [63] can gradually anneal the noise from a gaussian distribution to a data distribution $p(x)$ by learning the noise prediction from a T-length Markov noising process, giving $\{\boldsymbol{x_t}\}_{t=1}^{T}$. It leads to a significant influence in many research domains, such as the most famous image synthesis models [11, 23, 62, 56], the density estimation model [29] and the motion generation models [69, 75]. For motion generation, these works train the diffusion models with a transformer-based denoiser $\epsilon_\theta\left(x_t, t\right)$, which anneal the random noise to motion sequence $\{\boldsymbol{\hat{x}_t^{1:N}}\}_{t=1}^{T}$ iteratively.

However, diffusion on raw motion sequences is inefficient and requires exhausting computational resources. Besides, raw motion data from the markless or marker-based motion capture system usually remain high-frequency outliers, which might have a side effect on the diffusion model to learn the actual data distribution. To reduce the computational requirements of the diffusion models on raw motion sequences and improve the synthesized quality, we perform the diffusion process on a representative and low-dimensional motion latent space.

Here, we introduce our denoiser ϵ_θ. Different from the previous UNet-based architecture [59] on the 2D image latent z_I, we build a transformer-based denoising model with long skip connections [4] on the motion latent $z \in \mathbb{R}^{n\times d}$, which is more suitable for sequential data, like human motion sequences. The diffusion on latent space is modeled as a Markov nosing process using:

$$q\left(z_t \mid z_{t-1}\right) = \mathcal{N}\left(\sqrt{\alpha_t} z_{t-1}, \left(1-\alpha_t\right) I\right). \qquad (1)$$

where the constant $\alpha_t \in (0,1)$ is a hyper-parameters for sampling. We then use $\{z_t\}_{t=0}^{T}$ to denote the noising sequence, and $z_{t-1} = \epsilon_\theta\left(z_t, t\right)$ for the t-step denoising. We further focus on the unconditional generation with the simple objective [23]:

$$L_{\text{MLD}} := \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t\right)\right\|_2^2\right], \qquad (2)$$

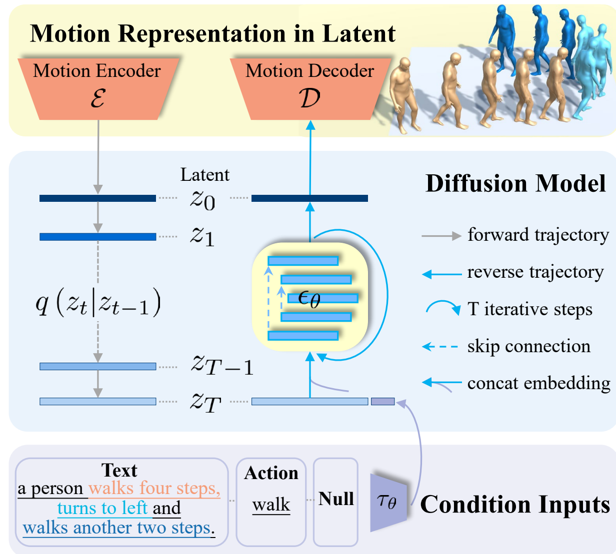

Figure 2. Method overview: MLD consists of a VAE model $\mathcal{V}$ (Sec. 3.1) and a latent diffusion model ϵ_θ (Sec. 3.2) conditioned on text or action embedding τ_θ (Sec. 3.3). We propose two-stage training: first learn $\mathcal{V}$ for *Motion Representations in Latents* and then learn a conditioned denoiser ϵ_θ from the diffusion process $q\left(z_t \mid z_{t-1}\right)$. During inference, In practice, the latent diffusion models predict the latent $\hat{z}_0$ from *condition inputs* and then $\mathcal{D}$ decode it to motions efficiently.

where $\epsilon \sim \mathcal{N}(0,1)$, $z_0 = \mathcal{E}(x^{1:L})$. During the training of ϵ_θ, the encoder $\mathcal{E}$ is frozen to compress motion into z_0. The samples of the diffusion forward process are from the latent distribution $p(z_0)$. During the diffusion reverse stage, ϵ_θ first predict $\hat{z}_0$ with T iterative denoising steps, then $\mathcal{D}$ decodes $\hat{z}_0$ to motion results with one forward.

3.3. Conditional Motion Latent Diffusion Model

Like many other diffusion models [69, 56, 61], our MLD model is also capable of conditional motion generation $\mathcal{G}(c)$ by applying the conditional distribution of $p(z|c)$, such as text [47, 16] and action [46, 19]. $\mathcal{G}(c)$ is implemented with conditional denoiser $\epsilon_\theta(z_t, t, c)$, which can share a common motion VAE model. Therefore, for different conditions, only the learning of $\epsilon_\theta(z_t, t, c)$ is necessary. To address various c, the domain encoder $\tau_\theta(c) \in \mathbb{R}^{m\times d}$ for condition embedding benefits the denoiser $\epsilon_\theta(z_t, t, \tau_\theta(c))$.

Here we introduce two specific generation tasks, text-to-motion $\mathcal{G}_w : w^{1:N} \mapsto x^{1:L}$ and action-to-motion $\mathcal{G}_a : a \mapsto x^{1:L}$. Through investigation, CLIP [52] text encoder $\tau_\theta^w(w^{1:N}) \in \mathbb{R}^{1\times d}$ is employed to map text prompt. On the other side, we build the learnable embedding for each action category, giving $\tau_\theta^a(a) \in \mathbb{R}^{1\times d}$. Injecting these embedded conditions into a transformer-based ϵ_θ, two effective ways are concatenation and cross-attention, and we figure out the former one seems to be more effective (cf. Sec. 5 and [69])

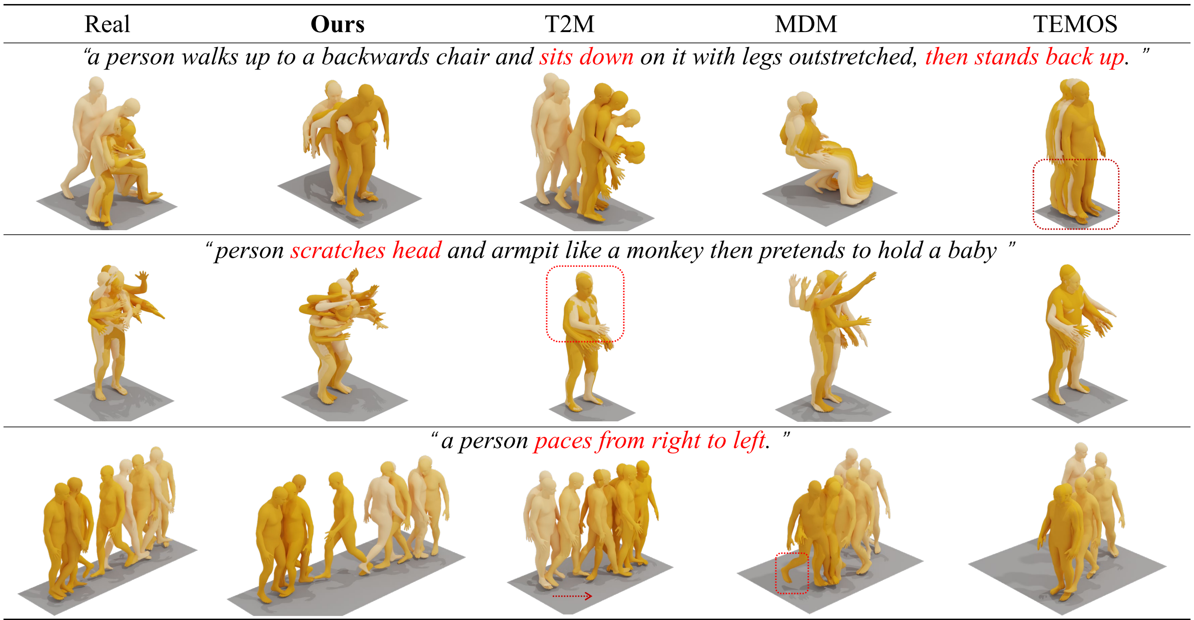

Figure 3. Qualitative comparison of the state-of-the-art methods. We provide the visualized motion results and real references from three text prompts. Under the same training and inference setting on HumanML3D [17], we find that our generations better match the descriptions, but others have downgraded motions or improper semantics.

for motion diffusion models. Thus the conditional objective follows:

$$L_{\text{MLD}} := \mathbb{E}_{\epsilon,t,c}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(c)\right)\|_2^2\right]. \quad (3)$$

We freeze τ_θ^w as suggested by [47] and joint optimize the τ_θ^a and ϵ_θ via this objective. In addition, our denoiser ϵ_θ is learned with classifier-free diffusion guidance [24], which is a trade-off to boost sample quality by reducing diversity in conditional diffusion models. Specifically, it learns both the conditioned and the unconditioned distribution with 10% dropout [61] of the samples, and we perform a linear combination to in as followed:

$$\epsilon_\theta^s(z_t, t, c) = s\epsilon_\theta(z_t, t, c) + (1 - s)\epsilon_\theta(z_t, t, \varnothing) \quad (4)$$

Here, s is the guidance scale and $s > 1$ can strengthen the effect of guidance. After the interactive reverse process of the conditional denoising, $\mathcal{D}$ reconstructs the motion from the predicted $\hat{z}_0$ efficiently.

4. Experiments

We provide extensive comparisons to evaluate our models on both quality and efficiency in the following. Firstly, we introduce the datasets settings, evaluation metrics and implementation details (Sec. 4.2). Importantly, we show the comparisons on multiple datasets for different motion generation tasks respectively, including text-to-motion (Sec. 4.3), action-to-motion (Sec. 4.4) and unconditional generation (Sec. 4.5). More qualitative results, user studies, and details are provided in supplements.

4.1. Datasets and Evaluation Metrics

Conditional motion synthesis can support rich inputs of multi-modal data, and thus multiple datasets are utilized to evaluate MLD. We briefly introduce these datasets. First is two text-to-motion datasets, HumanML3D and **KIT** [48], and the latter provides 6,353 textual descriptions for 3,911 motions. **HumanML3D** [17], a recent dataset, collects 14,616 motion sequences from **AMASS** [41] and annotates 44,970 sequence-level textual description. We use its motions, part of the AMASS, to evaluate unconditional task. As suggested by [17], we use the redundant motion representation in a combination of joint velocities, positions and rotations which is also used in [69, 75]. Lastly, action-to-motion task requires action-conditioned motions similar to action recognition datasets. Thanks to [46], after the processing, **HumanAct12** [19] provides 1,191 raw motion sequences and 12 action categories, and **UESTC** [26] provides 24K sequences and 40 action categories. We rely on these two datasets for action-to-motion evaluation.

Evaluation Metrics summarize in four parts. (a) Motion quality: Frechet Inception Distance (FID) is our principal metric to evaluate the feature distributions between the generated and real motions by feature extractor [16]. To evaluate reconstruction error of VAEs, we use popular metrics in motion capture [31, 8, 71], MPJPE and PAM-

PJPE [14] for global/local errors in millimeters, Acceleration Error (ACCL) for temporal quality. (2) Generation diversity: Diversity (DIV) calculates variance through features [16], while MultiModality (MM) measures the generation diversity within the same text or action input. (3) Condition matching: Under feature [16] space, motion-retrieval precision (R Precision) calculates the text and motion Top 1/2/3 matching accuracy, and Multi-modal Distance (MM Dist) calculates the distance between motions and texts. For action-to-motion, we use the corresponding action recognition model [19] [46] to calculate Accuracy (ACC) for action categories. (4) Time costs: we propose Average Inference Time per Sentence (AITS) measured in seconds to evaluate inference efficiency of diffusion models. ($cf.$ supplements)

4.2. Implementation Details

For the comparisons, motion transformer encoders $\mathcal{E}$ and decoders $\mathcal{D}$ of our VAE model $\mathcal{V}$ all consist of 9 layers and 4 heads with skip connection by default, as well as the transformer-based denoiser ϵ_θ in Sec. 3.2. The condition embedding $\tau_\theta(c) \in \mathbb{R}^{1,256}$ and the latent $z \in \mathbb{R}^{1,256}$ are concatenated for diffusion learning and inference. We employ a frozen *CLIP-ViT-L-14* model as the text encoder τ_θ^w for text condition, and a learnable embedding for action condition. We leave the ablation on the components in Sec. 5, like the shape of latent, the number of layers, injection of z_t through the cross-attention, and others. All our models are trained with the AdamW optimizer using a fixed learning rate of 10^{-4}. Our mini-batch size is set to 128 during the VAE training stage and 64 during the diffusion training stage separately. Each model was trained for 6K epochs during VAE stage and 3K epochs during diffusion stage. The number of diffusion steps is 1K during training while 50 during interfering, and the variances β_t are scaled linearly from 8.5×10^{-4} to 0.012. For runtime, training tasks 8 hours for VAEs $\mathcal{V}$ and 4 hours for denoiser ϵ_θ on 8 Tesla V100 GPUs, and we test MLD with a single V100 in Sec. 5, but it also can run inference on a general computer graphics card, such as RTX 2080/3060.

4.3. Comparisons on Text-to-motion

By introducing motion latent diffusion models based on text input $w^{1:N}$, we open up the exploration of conditional motion generation. We train a 25M parameter *MLD-1* conditioned on the language prompt and employ the frozen CLIP [52] model as τ_θ^w to encode the text to projected pooled output, giving $w_{clip}^1 \in \mathbb{R}^{1,256}$. We evaluate state-of-the-art methods on HumanML3D and KIT with suggested metrics [17] under the 95% confidence interval from 20 times running. Most results are borrowed from their own paper or the benchmark in [18], except TEMOS [47]. We train it with the proposed default model setting on two datasets to uniform the evaluation metrics. Besides, the deterministic methods [49, 5, 2] can not generate diverse results from one input and thus we leave their MModality metrics empty. Tab. 1 and Tab. 2 summarize the comparisons results. We achieve the best FID, R Precision and MM Dist on HumanML3D and KIT, outperforming previous cross-modal models as well as motion diffusion models. It indicates high-quality motion and high text prompt matching, as also shown in Fig. 3. Our generated results correctly match the text prompt while maintaining a rich diversity of generated motions.

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Seq2Seq [49]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
JL2P [2]	$0.246^{\pm.001}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2G [5]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
Hier [13]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
TEMOS [47]	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
T2M [16]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MotionDiffuse [75]	$\mathbf{0.491}^{\pm.001}$	$\mathbf{0.681}^{\pm.001}$	$\mathbf{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$\mathbf{3.113}^{\pm.001}$	$\underline{9.410}^{\pm.049}$	$1.553^{\pm.042}$
MDM [69]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$\underline{0.544}^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
MLD (Ours)	$\underline{0.481}^{\pm.003}$	$\underline{0.673}^{\pm.003}$	$\underline{0.772}^{\pm.002}$	$\mathbf{0.473}^{\pm.013}$	$\underline{3.196}^{\pm.010}$	$9.724^{\pm.082}$	$\underline{2.413}^{\pm.079}$

Table 1. Comparison of text-conditional motion synthesis on HumanML3D [17] dataset. These metrics are evaluated by the motion encoder from [16]. Empty MModality indicates the non-diverse generation methods. We employ real motion as a reference and sort all methods by descending FIDs. The right arrow → means the closer to real motion the better. **Bold** and underline indicate the best and the second best result.

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
Seq2Seq [49]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
T2G [5]	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
JL2P [2]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Hier [13]	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	$2.090^{\pm.083}$
TEMOS [47]	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$	$0.532^{\pm.034}$
T2M [16]	$0.370^{\pm.005}$	$0.569^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$3.401^{\pm.008}$	$\underline{10.91}^{\pm.119}$	$1.482^{\pm.065}$
MotionDiffuse [75]	$\mathbf{0.417}^{\pm.004}$	$\mathbf{0.621}^{\pm.004}$	$\mathbf{0.739}^{\pm.004}$	$1.954^{\pm.062}$	$\mathbf{2.958}^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
MDM [69]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$\underline{0.497}^{\pm.021}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$	$\underline{1.907}^{\pm.214}$
MLD (Ours)	$\underline{0.390}^{\pm.008}$	$\underline{0.609}^{\pm.008}$	$\underline{0.734}^{\pm.007}$	$\mathbf{0.404}^{\pm.027}$	$\underline{3.204}^{\pm.027}$	$10.80^{\pm.117}$	$\mathbf{2.192}^{\pm.071}$

Table 2. We involve KIT [48] dataset and evaluate the SOTA methods on the text-to-motion task. ($cf.$ Tab. 1 for metrics details)

Methods	UESTC					HumanAct12			
	$\text{FID}_{\text{train}}$ ↓	FID_{test} ↓	ACC↑	DIV→	MM→	$\text{FID}_{\text{train}}$ ↓	ACC ↑	DIV→	MM→
Real	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
ACTOR [46]	$20.5^{\pm2.3}$	$23.43^{\pm2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$6.840^{\pm.030}$	$2.530^{\pm.020}$
INR [7]	$\mathbf{9.55}^{\pm.06}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$
MDM [69]	$9.98^{\pm1.33}$	$\mathbf{12.81}^{\pm1.46}$	$0.950^{\pm.000}$	$33.02^{\pm.28}$	$\mathbf{14.26}^{\pm.12}$	$0.100^{\pm.000}$	$\mathbf{0.990}^{\pm.000}$	$6.680^{\pm.050}$	$\mathbf{2.520}^{\pm.010}$
MLD (Ours)	$12.89^{\pm.109}$	$15.79^{\pm.079}$	$\mathbf{0.954}^{\pm.001}$	$\mathbf{33.52}^{\pm.14}$	$13.57^{\pm.06}$	$\mathbf{0.077}^{\pm.004}$	$0.964^{\pm.002}$	$\mathbf{6.831}^{\pm.050}$	$2.824^{\pm.038}$

Table 3. Comparison of action-conditional motion synthesis on UESTC [26] and HumanAct12 [19] dataset: $\text{FID}_{\text{train}}$, $\text{FID}_{\text{train}}$ indicate the evaluated splits. Accuracy (ACC) for action recognition. Diversity (DIV), MModality (MM) for generated motion diversity within each action label.

4.4. Comparisons on Action-to-motion

The action-conditioned task is given an input action label to generate relevant motion sequences. We compare with ACTOR [46], INR [7] and MDM [69]. ACTOR and INR are transformer-based VAE models and focus on the action-conditioned task, and MDM is a diffusion model using the same learnable action embedding module as ours. We still

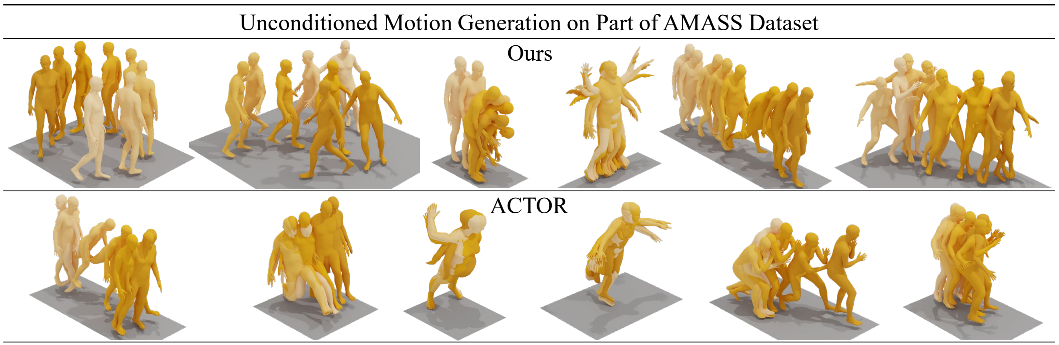

Figure 4. Qualitative comparison of unconditioned motion generation. Samples for our MLD and ACTOR [46] trained on a split of AMASS [41] dataset, the motion part of [17]. More samples in the supplements.

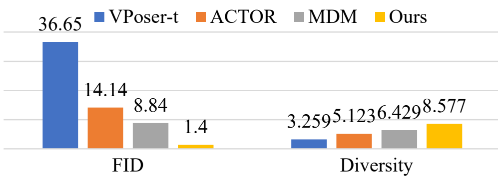

Figure 5. Comparison of unconditional motion generation on part of AMASS [41] dataset with the state-of-the-art methods. We provide both FID and Diversity to evaluate generated motions.

provide 20 evaluations as introduced and report FID scores on the training set and test set like [46] for comparison. Tab. 3 shows the comparison on two datasets, UESTC [26] and HumanAct12 [19]. MLD achieves state-of-the-art accuracy and diversity on UESTC and competitive results on HumanAct12, indicating that diffusion models in motion latent can also benefit action-conditioned generation task.

4.5. Comparisons on Unconditional Generation

We then evaluate the generation effect of MLD by introducing unconditional task on motions of HumanML3D [17], actually part of AMASS [41]. MLD supports two manners for unconditional generation, latent sampling (cf. Sec. 5) and diffusion sampling. Here we focus on the evaluation of the latter and employ FID and Diversity for motion quality and diversity. With the same process on training and evaluations on the part of AMASS [41] data, we provide real motion, ACTOR [46], and VPoser-t [43] and MDM [69] as our comparison baselines, We employ the transformer VAE from ACTOR, then follow TEMOS [47] to make it class-agnostic and set 6 heads/layers for transformers, 10^{-4} as learning rate. To perform the temporal-based task, the input of VPoser-t is modified as a motion of fixed length. MDM also supports this task, thus we finetune and evaluate their provided model. Fig. 5 reports that MLD has the best motion generation quality and diversity.

Method	Reconstruction			Generation	
	MPJPE ↓	PAMPJPE↓	ACCL↓	FID↓	DIV→
Real	-	-	-	0.002	9.503
VPoser-t [43]	75.6	48.6	9.3	1.430	8.336
ACTOR [46]	65.3	41.0	7.0	0.341	9.569
Ours-7 ($\mathcal{V}$,skip,9 layers)	**14.7**	**8.9**	**5.1**	**0.017**	**9.554**
Ours-1 (z,$\mathbb{R}^{1\times256}$)	54.4	41.6	8.3	0.247	9.630
Ours-2 (z,$\mathbb{R}^{2\times256}$)	51.8	37.8	8.3	0.166	9.626
Ours-5 (z,$\mathbb{R}^{5\times256}$)	24.3	14.7	5.8	0.043	9.593
Ours-7 (z,$\mathbb{R}^{7\times256}$)	14.7	8.9	5.1	0.017	9.554
Ours-10 (z,$\mathbb{R}^{10\times256}$)	17.3	11.5	5.8	0.025	9.589
Ours-7 ($\mathcal{V}$,w/ skip)	14.7	8.9	5.1	0.017	9.554
Ours-7 ($\mathcal{V}$, w/o skip)	18.5	10.4	5.6	0.027	9.528
Ours-7 ($\mathcal{V}$, 7 layers)	16.0	10.2	5.3	0.022	9.593
Ours-7 ($\mathcal{V}$, 9 layers)	14.7	8.9	5.1	0.017	9.554
Ours-7 ($\mathcal{V}$, 11 layers)	17.2	11.2	5.4	0.021	9.533

Table 4. Evaluation of our VAE models $\mathcal{V}$ on the motion part of HumanML3D [17] dataset: MPJPE and PAMPJPE are measured in millimeter. ACCL indicates acceleration error. We evaluate FID and DIV the same as Tab. 1. From top to down, we propose real reference, VPoser-t [43] and ACTOR [46] as baselines, the evaluation on latent $z \in \mathbb{R}^{i\times256}$, with (w/) or without (w/o) skip connection, $\mathcal{V}$ with different number of transformer layers.

5. Ablation Studies

MLD comprises a motion VAE model $\mathcal{V}$ and latent diffusion models ϵ_θ, and both influence its effect. We first focus on $\mathcal{V}$ to evaluate its components with generation and reconstruction metrics. Based on these $\mathcal{V}$, we evaluate MLDs in diffusion learning aiming at text-to-motion and unconditional synthesis, and then report time costs on inference.

Effectiveness of Latents in Motion Sequences Representation. We first ablate several components of our VAE models $\mathcal{V}$ in a controlled setup, studying the shape of latent z, skip connection, and the number of transformer layers, as shown in Tab. 4. The most important variable of MLD, the latent vector z, is a bridge between $\mathcal{V}$ and diffusion models τ_θ. We lock the pose x^i (one frame of motion) embedding dimensionality to 256, which is the same as [47] , and explore $z \in \mathbb{R}^{i\times256}$, giving MLD-$i$. We then evaluate the skip connection and transformer layers on the best MLD-7. All comparison baselines, including ACTOR [46] and

Models	R Precision Top 3↑	FID↓	MM Dist.↓	Diversity→	MModality↑
Real	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
MLD-1 (z, $\mathbb{R}^{1\times256}$)	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
MLD-2 (z, $\mathbb{R}^{2\times256}$)	$0.727^{\pm.003}$	$0.585^{\pm.015}$	$3.448^{\pm.011}$	$9.084^{\pm.081}$	$\mathbf{2.725}^{\pm.093}$
MLD-5 (z, $\mathbb{R}^{5\times256}$)	$0.722^{\pm.003}$	$1.554^{\pm.019}$	$3.511^{\pm.008}$	$8.424^{\pm.081}$	$2.542^{\pm.080}$
MLD-7 (z, $\mathbb{R}^{7\times256}$)	$0.731^{\pm.002}$	$1.011^{\pm.019}$	$3.415^{\pm.008}$	$8.736^{\pm.064}$	$2.463^{\pm.089}$
MLD-10 (z, $\mathbb{R}^{10\times256}$)	$0.703^{\pm.003}$	$1.716^{\pm.027}$	$3.616^{\pm.012}$	$8.606^{\pm.067}$	$2.604^{\pm.087}$
MLD-1 (ϵ_θ, cross-att)	$0.592^{\pm.004}$	$1.922^{\pm.041}$	$4.480^{\pm.015}$	$8.598^{\pm.088}$	$3.768^{\pm.126}$
MLD-1 (ϵ_θ, concat)	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
MLD-1 (ϵ_θ, w/o skip)	$0.749^{\pm.003}$	$0.784^{\pm.015}$	$3.363^{\pm.010}$	$9.568^{\pm.093}$	$2.597^{\pm.098}$
MLD-1 (ϵ_θ, w/ skip)	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
MLD-1 (ϵ_θ, 5 layers)	$0.760^{\pm.002}$	$\mathbf{0.314}^{\pm.010}$	$3.259^{\pm.009}$	$9.706^{\pm.072}$	$2.635^{\pm.085}$
MLD-1 (ϵ_θ, 7 layers)	$0.771^{\pm.003}$	$0.349^{\pm.012}$	$3.199^{\pm.012}$	$\mathbf{9.624}^{\pm.062}$	$2.504^{\pm.088}$
MLD-1 (ϵ_θ, 9 layers)	$\mathbf{0.772}^{\pm.002}$	$0.473^{\pm.013}$	$\mathbf{3.196}^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
MLD-1 (ϵ_θ, 11 layers)	$0.771^{\pm.003}$	$0.402^{\pm.011}$	$3.203^{\pm.013}$	$9.876^{\pm.088}$	$2.478^{\pm.076}$

Table 5. Evaluation of text-based motion synthesis on HumanML3D [17]: we use metrics in Tab. 1 and provides real reference, the evaluation on latent z ($cf.$ $\mathcal{V}$ in Tab. 4), cross-attention or concatenation with conditions τ_θ, with (w/) or without (w/o) skip connection, ϵ_θ with different number of transformer layers.

Methods	FID↓	Diversity→	Methods	FID↓	Diversity→
Real	-	9.503	Real	-	9.503
VPoser-t [43]	36.65	3.259	MLD-1 (z,$\mathbb{R}^{1\times256}$)	**1.055**	**8.577**
ACTOR [46]	14.14	5.123	MLD-2 (z,$\mathbb{R}^{1\times256}$)	4.408	7.420
MDM [69]	8.848	6.429	MLD-5 (z,$\mathbb{R}^{5\times256}$)	7.829	6.247
MLD-1 (ϵ_θ, w/o skip)	2.575	7.566	MLD-7 (z,$\mathbb{R}^{7\times256}$)	7.614	6.233
MLD-1 (ϵ_θ, w/ skip)	**1.055**	**8.577**	MLD-10 (z,$\mathbb{R}^{10\times256}$)	9.624	6.194

Table 6. Evaluation of unconditional motion generation. From left to right, we evaluate the denoiser ϵ_θ with (w/) or without (w/o) skip connection and the latent z ($cf.\mathcal{V}$ in Tab. 4)

VPoser-t [43], follow the same training and evaluation with our proposed MLD. Since the original VPoser can only handle single frame pose, we modified it to a sequential manner with a fixed length. The results in Tab. 4 demonstrate the effectiveness of our proposed VAEs over others in the motion sequences representation.

Effectiveness of Latents in Motion Latent-based Diffusion Models. In Tab. 5, we select the text-to-motion task as our focus and evaluate latent diffusion models ϵ_θ, using the similar metrics in Tab. 1. MLD-i denotes the shape of latent $z \in \mathbb{R}^{i\times256}$. Importantly, MLD-1, using the smallest latent, wins the best performance in most metrics. After that, the evaluation on the components of ϵ_θ is provided, *cross-att* and *concate* represent the cross-attention or concatenation for condition embedding $\epsilon_\theta(c)$. Interestingly, MDM [69] also reports the encoder design by concatenating embedding is better. We find that skip connection, which is important for images [4, 56], also provided significant improvement in motion latent diffusion models, but MLDs using different numbers of layers in ϵ_θ achieve similar effects on this dataset. We then evaluate the generation of MLD by diffusion sampling, different from the generation in $\mathcal{V}$ by latent sampling ($cf.$ Tab. 4). As shown in Tab. 6, the MLD using the smallest latent and skip connection outperforms others. The evaluation of how different language models influence MLDs and the details of latent/diffusion sampling are provided in supplements.

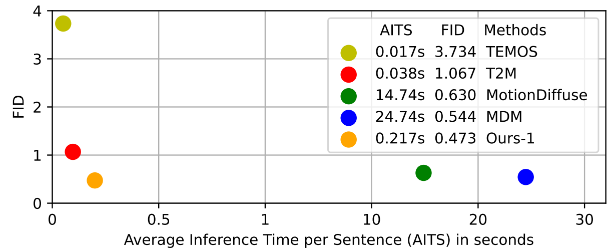

Figure 6. Comparison of the inference time costs on text-to-motion. We calculate AITS on the test set of HumanML3D [17] without model or data loading parts. All tests are performed on the same Tesla V100. The closer the model is to the origin the better.

Inference time. While diffusion models lead to significant improvements, one notable limitation of motion diffusion models [69, 75] is the long inference time. In Sec. C, we adopted Denoising diffusion implicit models (DDIM) [64] to provide a detailed evaluation of the inference time, floating-point operations (FLOPs), and FID. As shown in Fig. 6, MDM [69] requires 24.74 seconds for average inference and up to a minute for maximum inference on a single V100. Compared to them, our MLD needs less computational overhead and achieves higher performance with two orders of magnitude faster speed.

6. Disscusion

As the trial to explore conditional motion generation with motion latent diffusion models, the proposed MLD still owns limitations as follows. First, same as most motion generation methods, our method can generate arbitrary length results but still under the max-length in the dataset. It's interesting to model a non-stop human motion in temporal consistency. Besides, MLD focuses on articulated human bodies, while there is also other work on faces [27, 6], hands [58, 37, 36] and even animal [60, 80] motion.

We propose a motion latent-based diffusion model to generate plausible human motion sequences conforming to the action classes or natural language descriptions. Compared to the compatible cross-modal latent space-based method, our MLD could produce more diverse and plausible human motion sequences; Compared to the previous diffusion-based methods on raw motion sequences, our MLD needs less computational overhead, with two orders of magnitude faster. Extensive experiments on various human motion generation tasks demonstrate the effectiveness and efficiency of our proposed MLD.

7. Acknowledgements

This work is supported by Zhejiang Lab Project (No. 2021KH0AB05) and Shanghai Natural Science Foundation (No. 23ZR1402900).

References

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

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

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. *arXiv preprint arXiv:2209.12152*, 2022.

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

[6] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, and Jingyi Yu. Sparse photometric 3d face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4635–4644, 2018.

[7] 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.

[8] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129(10):2846–2864, 2021.

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

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

[13] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1396–1406, 2021.

[14] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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

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

[18] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022.

[19] 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.

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

[21] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021.

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

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

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

[25] 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.

[26] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *Proceedings of the 26th ACM international Conference on Multimedia*, pages 1510–1518, 2018.

[27] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

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

[29] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

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

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

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

[34] 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.

[35] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. 2022.

[36] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. Piano: A parametric hand bone model from magnetic resonance imaging. *arXiv preprint arXiv:2106.10893*, 2021.

[37] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

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

[39] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020.

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015.

[41] 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 (ICCV)*, October 2019.

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

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, June 2019.

[45] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), July 2021.

[46] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

[47] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

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

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

[50] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[51] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data. *arXiv preprint arXiv:2206.08010*, 2022.

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

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[54] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021.

[55] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[57] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.

[58] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[60] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3876–3884, June 2022.

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[62] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

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

[65] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[66] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019.

[67] Ömer Terlemez, Stefan Ulbrich, Christian Mandery, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Master motor map (mmm)—framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 894–901. IEEE, 2014.

[68] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022.

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

[70] 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.

[71] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[72] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021.

[73] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

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

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

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

[77] 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.

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

[79] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5826–5835, October 2021.

[80] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3963. IEEE Computer Society, 2018.