EMDM: Efficient Motion Diffusion Model for Fast and High-Quality Motion Generation

Wenyang Zhou[1], Zhiyang Dou[2,3,4,†,⋆], Zeyu Cao[1], Zhouyingcheng Liao[2], Jingbo Wang[5], Wenjia Wang[2], Yuan Liu[2], Taku Komura[2], Wenping Wang[6], and Lingjie Liu[3]

[1] University of Cambridge
[2] The University of Hong Kong
[3] University of Pennsylvania
[4] TransGP
[5] Shanghai AI Laboratory
[6] Texas A&M University
[†] *Project Lead.*

Abstract. We introduce Efficient Motion Diffusion Model (EMDM) for fast and high-quality human motion generation. Current state-of-the-art generative diffusion models have produced impressive results but struggle to achieve fast generation without sacrificing quality. On the one hand, previous works, like motion latent diffusion, conduct diffusion within a latent space for efficiency, but learning such a latent space can be a non-trivial effort. On the other hand, accelerating generation by naively increasing the sampling step size, e.g., DDIM, often leads to quality degradation as it fails to approximate the complex denoising distribution. To address these issues, we propose EMDM, which captures the complex distribution during multiple sampling steps in the diffusion model, allowing for much fewer sampling steps and significant acceleration in generation. This is achieved by a conditional denoising diffusion GAN to capture multimodal data distributions among arbitrary (and potentially larger) step sizes conditioned on control signals, enabling fewer-step motion sampling with high fidelity and diversity. To minimize undesired motion artifacts, geometric losses are imposed during network learning. As a result, EMDM achieves real-time motion generation and significantly improves the efficiency of motion diffusion models compared to existing methods while achieving high-quality motion generation. Our code is available at https://github.com/Frank-ZY-Dou/EMDM.

Keywords: Text-to-motion · Motion generation · Diffusion model · GAN

1 Introduction

Tremendous efforts have been made for human motion generation with different modalities, including action labels [16, 21, 35, 53, 87], textual descriptions [1, 15,

[⋆] Collaborating with Tencent Games.

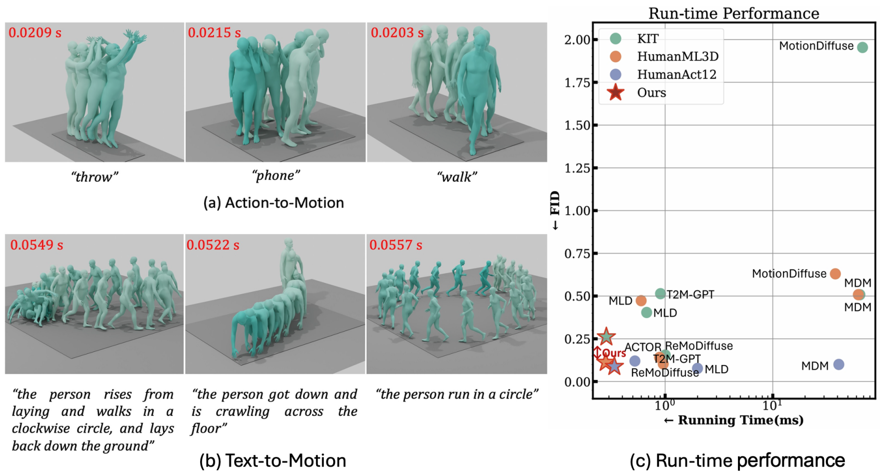

Fig. 1: EMDM produces high-quality human motion aligned with input conditions in a short runtime. The average run time of EMDM in (a) action-to-motion and (b) text-to-motion tasks is 0.02s and 0.05s per sequence, respectively. For reference, the corresponding times for MDM [76] are 2.5s and 12.3s. We deepen the color of the character with respect to the time step of the sequence. (c) Overall comparison of the inference time costs on the HumanML3D, KIT, and HumanAct12 datasets. For ease of illustration, the Running Time is plotted with a log scale. We compare the running time per frame vs. the FID of SOTA methods.

19, 20, 28, 31, 32, 54, 76, 95, 96, 99], and audio [2, 33, 36, 38], etc. The diffusion model [25, 58, 69] has been at the forefront of these advances [7, 30, 66, 76, 98], due to its promise of effectively capturing the target distribution of diverse body motions. However, these models struggle to achieve fast motion generation while maintaining high motion quality. For instance, MDM [76] takes around 12s to produce a motion sequence given a textual description. Such low efficiency limits their effectiveness in real-world applications, e.g., online motion synthesis.

Existing efforts to improve the generation efficiency of the motion diffusion model can be mainly categorized into two types: 1) motion latent diffusion proposed by MLD [7]. This involves first learning a latent space of body motion and then conducting latent diffusion. However, such a two-stage approach relies on effectively embedding the motion in the first stage—it is challenging to learn a good embedding space for the subsequent latent diffusion model. The expression of the latent space typically limits the performance of downstream motion generation, as evidenced by both quantitative (Sec. 4.3 and Sec. 4.4) and qualitative comparisons. 2) The DDIM sampling strategy [69] can be adopted to accelerate generation by reducing the number of denoising steps (using a larger step size), given that the standard number of denoising steps is typically 1000 [76, 98]. Additionally, the Gaussian assumption on the denoising distribution holds only for small step sizes. Therefore, directly using a larger step size during motion sampling skips numerous reverse steps and leads to much more complex data

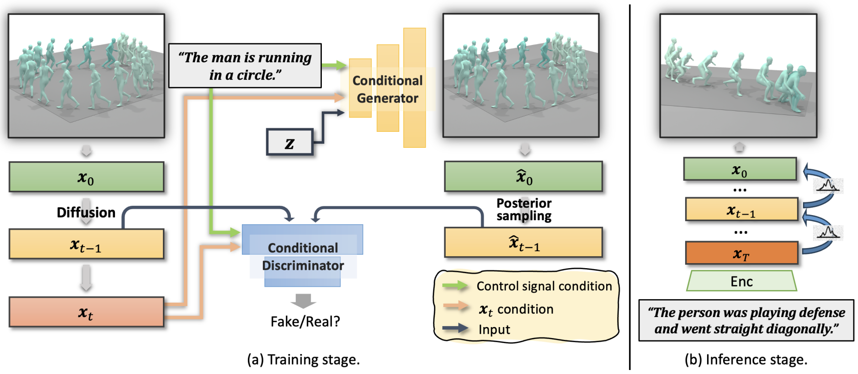

Fig. 2: Pipeline of EMDM. We develop condition denoising diffusion GAN to capture the complex denoising distribution of human body motion, allowing a larger sampling step size (Sec. 3.1). During inference, we use a larger sampling step for fast sampling of high-quality motion w.r.t. input condition. The detailed sampling algorithm is given in Alg. 1. Note that we ignore the time step t for simplicity.

distributions than Gaussians. As the complex data distributions cannot be approximated with fewer sampling steps, the performance of this approach drops, as shown in Tab. D2 in Appendix D. Thus, it is critical to capture the complex data distributions when a few-step sampling is involved; see Fig. 3.

In this paper, we present Efficient Motion Diffusion Model (EMDM) for fast and high-quality human motion generation. We seek to reduce the number of sampling steps while achieving fast motion generation. The key to allowing a larger sampling step size is to effectively capture the complex data distributions during a few-step sampling. Inspired by recent advances in efficient image synthesis [85], we develop a sampling strategy for fast motion generation while maintaining high motion quality. Specifically, we employ a conditional denoising diffusion GAN, incorporating a conditional generator and conditional discriminator that consider both the time step t and input control signals (e.g., text); See Fig. 2. The generator (denoiser) is trained to generate the motion $\hat{x}_0$ conditioning on input control signals, time step t, given the random variables. Then posterior sampling (Alg. 2) is applied to produce $\hat{x}_{t-1}$ at the $t-1$-th time step using $\hat{x}_0$. A discriminator is trained to distinguish whether a data sample $\hat{x}_{t-1}$ is a plausible denoised result of x_t. As t varies during diffusion model training, the generator learns to capture the complex denoising distribution introduced by an arbitrary (and potentially larger) sampling step size. As a result, during sampling, one could use a larger sampling

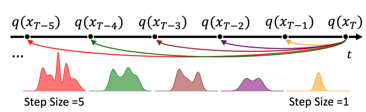

Fig. 3: Denoising distribution becomes complex (non-Gaussian) when increasing sampling step sizes for few-step sampling.

step size (fewer steps) to sample a motion given the conditions, significantly improving runtime performance. As a condition of the model, the control signals make the capture of the complex motion distribution more efficient by learning the conditional denoising distribution. Finally, to reduce unwanted artifacts, we further integrate geometric motion losses during model training to stabilize the training process and enhance motion quality. Our model is trained end-to-end, simplifying the training process and significantly reducing the overall training effort, which is a noteworthy advantage in practical applications.

As a result, EMDM effectively captures the complex motion distribution, enabling much fewer sampling steps during motion generation while maintaining high-quality motion; See Fig. 1 for some examples of generated motion and overall running time statistics. Our contribution is three-fold:

- We reveal the efficiency issues with existing motion diffusion models and the challenges in accelerating the models.
- We present EMDM for fast and high-quality motion generation by employing a conditional denoising diffusion GAN to effectively model complex denoising distributions for the few-step motion generation with high quality.
- We perform extensive experiments on EMDM to demonstrate its remarkable speed-up for diffusion-based approaches with competitive or even higher quality and diversity of the generated motions compared with SOTAs.

2 Related Work

Human Motion Generation Human motion generation is an important research problem in computer vision and computer animation. The ability to generate realistic and natural human motions has wide applications including virtual reality [34,84,91], game development [27,70–72,74], human behavior analysis [9, 14,22,43,90,100,101] and robotics [11,40,67,81,88]. The generated motion can condition on abundant, multi-modal inputs such as action labels [16,21,35,53,87], textual description [1,4,6,15,19,20,31,32,55,76,78,95,96,99,104], incomplete pose sequences [18,23,76,82], control signals [13,27,39,42,52,56,65,70,71,79,80,86,97], music or audio [3,33,36,38,48,106], and so on. For *Unconditional Motion Generation* [59, 62, 76, 89, 102, 105], the goal is to model the entire motion space based on motion data. For instance, VPoser [49] introduces a variational human pose prior primarily for image-based pose fitting, while ACTOR [53, 54] presents a class-agnostic transformer VAE as a baseline. Humor [62] employs a conditional VAE for learning motion prior in an auto-regressive manner. The recent study [65] learns phase-conditioned motion prior in the frequency domain. *Action-to-Motion* [16,21,35,53,87] can be viewed as the inverse task of the classical action recognition task, where the goal is to produce human motion given the input action labels. Specifically, ACTOR [53] introduces learnable biases within a transformer VAE to encapsulate action for motion generation. Nowadays, *Text-to-Motion* [1,4,19,31,32,54,76,95,98,99] has become popular, primarily because of the user-friendly and accessible nature of language descriptors. Specifically, T2M-GPT [95] proposes a classic framework based on VQ-VAE and GPT to synthesize human motion from textual descriptions. [76,98] employ diffusion models for high quality text-to-motion. Recently, [28,104] propose motion language pretraining using LLMs [12,61,77] for text-driven motion synthesis.

Motion Diffusion Models. Diffusion Generative Models [68] have shown impressive results in wide fields [10, 24, 44, 45, 47, 58, 63, 76, 92, 98] and Diffusion models have been employed for human motion generation [7,30,31,66,76,86,98]. Specifically, MotionDiffuse [98] stands as the first text-based motion diffusion model using fine-grained instructions for body part-level control. MDM [76] conducts motion diffusion that operates on raw motion data, learning the relationship between motion and input conditions. ReMoDiffuse [99] presents a retrieval-augmented motion diffusion model, where extra knowledge from the retrieved samples is used for motion synthesis. Recent efforts [30, 86] have concentrated on controllable human motion generation, leveraging either pelvis location [30] or specific body joints [86]. That being said, applying the diffusion model to the motion data [76, 98] as a sequential motion generation framework incurs high computational overheads and typically results in low inference speeds due to model size and their iterative sampling nature. To tackle the problem, MLD [7] introduces a motion latent-based diffusion model by first training a VAE for motion embedding, followed by the application of latent diffusion within the learned latent space. However, this is a two-stage method and requires non-end-to-end training: effectively capturing the motion distribution during motion embedding can be challenging, yet it is crucial for the success of the second stage. In contrast, our approach aims to boost efficiency by accelerating the sampling process and is end-to-end trainable. Retrieval-based method [99] could achieve relatively fast motion generation. As of yet, it relies on reference motion datasets and suffers from relatively low motion diversity. EMDM allows for much fewer sampling steps during the denoising process without the reliance on the reference motion for sequential motion generation with high quality.

3 Method

Our goal is to efficiently generate high-quality and diverse human motion given conditional inputs in real time. We propose an Efficient Motion Diffusion Model utilizing a conditional denoising diffusion GAN for fast motion generation, which will be elaborated in the following.

3.1 Efficient Motion Diffusion Model

In this task, the motion of humans, denoted as $\mathbf{x}^{1:N}$, is associated with a corresponding condition $\mathbf{c}$, e.g., action [21, 53, 76] or text [19, 75, 76, 98]. N is the number of frames in a motion sequence. Note that unconditioned motion generation is available by $\mathbf{c} = \emptyset$ similar to [7, 76]. We use probabilistic diffusion models [68] for motion generation. The forward process of the diffusion model is given by

$$q(\mathbf{x}_{1:T}^{1:N}|\mathbf{x}_0^{1:N}) = \prod_{t\geq 1} q(\mathbf{x}_t^{1:N}|\mathbf{x}_{t-1}^{1:N}), \quad q(\mathbf{x}_t^{1:N}|\mathbf{x}_{t-1}^{1:N}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}^{1:N}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. When α_t is small enough, we can approximate $\mathbf{x}_T^{1:N} \sim \mathcal{N}(0, \mathbf{I})$ [68]. The reverse process is given by

$$p_\theta(\mathbf{x}_{0:T}^{1:N}) = p(\mathbf{x}_T^{1:N}) \prod_{t\geq 1} p_\theta(\mathbf{x}_{t-1}^{1:N}|\mathbf{x}_t^{1:N}), p_\theta(\mathbf{x}_{t-1}^{1:N}|\mathbf{x}_t^{1:N}) = \mathcal{N}(\mathbf{x}_{t-1}^{1:N}; \boldsymbol{\mu}_\theta(\mathbf{x}_t^{1:N}, t), \sigma_t^2\mathbf{I}), \quad (2)$$

where θ is the learnable parameters of the diffusion model which gradually anneals the noise from a Gaussian distribution to the data distribution.

When training a motion diffusion model, a denoiser $\epsilon_\theta(\mathbf{x}_t, t)$ learns to iteratively anneal the random noise to the motion sequence $\{\hat{\mathbf{x}}_t^{1:N}\}_{t=1}^T$, where the human pose $\mathbf{x}^i \in \mathbb{R}^{J\times D}$ at the i-th frame is represented by either joint rotations or positions, where J is the number of joints and D is the dimension of the joint representation. When α_t is large, the denoising distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ can be both regarded as Gaussian. With this assumption, diffusion models often have thousands of steps with a large α_t, e.g., MDM [76] and MotionDiffuse [98] need 1000 steps for denoising, leading to a rather slow motion generation process. Obviously, when the denoising step size is naively increased (fewer sampling steps), i.e. in the case of DDIM sampling, the distribution is non-Gaussian; there is thus no guarantee that the Gaussian assumption on the denoising distribution holds (see Fig. 3). Consequently, the quality of generated motions drops.

Inspired by the recent progress [85] in image generation, we propose to model the expressive multimodal denoising distribution with a larger step size $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ by conditioning on the control signals and time step t. The training process is formulated by matching $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ when each diffusion step has smaller α_t, which allows T to be small ($T \leq 10$).

Conditional Generator. As $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = g_\theta(\mathbf{x}_\theta, t))$ [25], one can first predict $\mathbf{x}_0$ using the diffusion model $g_\theta(\mathbf{x}_\theta, t)$ and then sample $\mathbf{x}_{t-1}$ using the posterior distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ [85]. In this paper, we employ a conditional denoising diffusion GAN, which integrates a conditional generator and conditional discriminator, considering both the time step t and input control signals $\mathbf{c}$, for example text. To achieve motion denoising, the g_θ is modeled by a conditional generator $G_\theta(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)$ that outputs $\hat{\mathbf{x}}_0$, given $\mathbf{x}_t$, control signal $\mathbf{c}$ and an L-dimensional latent variable $\mathbf{z} \sim p(\mathbf{z}) := \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. Mathematically, with $G_\theta(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)$, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is obtained by

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \int p_\theta(\mathbf{x}_0|\mathbf{x}_t) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_0 \\ &= \int p(\mathbf{z}) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = G_\theta(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)) d\mathbf{z}. \end{aligned} \quad (3)$$

We further use posterior distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ to sample $\hat{\mathbf{x}}_{t-1}$ for discrimination based on the predicted $\hat{\mathbf{x}}_0$ in the following.

Conditional Discriminator. We employ a time step-dependent and control signal-conditioned discriminator as $D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t)$. The N-dimensional $\mathbf{x}_{t-1}$, $\mathbf{x}_t$ are two inputs at time step $t-1$ and t, and $\mathbf{c}$ is the control signal such as textual descriptions. It is trained to distinguish whether $\mathbf{x}_{t-1}$ is a plausible denoised result of $\mathbf{x}_t$. The discriminator is trained by

$$\begin{aligned}\min_\phi \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \; [\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)}[\mathrm{F}(-D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t))]+ \\ \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}[\mathrm{F}(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t))],\end{aligned} \tag{4}$$

where $\mathrm{F}(\cdot)$ denotes the softplus$(\cdot)$ function and fake samples from $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ are contrasted against real samples from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. By using the identity $q(\mathbf{x}_t, \mathbf{x}_{t-1}) = \int d\mathbf{x}_0 q(\mathbf{x}_0) q(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{x}_0) = \int d\mathbf{x}_0 q(\mathbf{x}_0) q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_{t-1})$, we have

$$\begin{aligned}\min_\phi \sum_{t\geq 1} (\mathbb{E}_{q(\mathbf{x}_0) q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_{t-1})}[\mathrm{F}(-D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t))] \\ + \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}[\mathrm{F}(D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t))]]).\end{aligned} \tag{5}$$

Given the training goal of the condition discriminator in Eq. 5, we can train the conditional generator $G_\theta(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)$ by $\max_\theta \mathcal{L}_{\mathrm{disc}}$, where $\mathcal{L}_{\mathrm{disc}}$ is defined by

$$\mathcal{L}_{\mathrm{disc}} = \mathbb{E}_{t\sim[1,T], q(\mathbf{x}_t)} \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}[\mathrm{F}(-D_\phi(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{c}, t))]. \tag{6}$$

The overall pipeline is shown in Fig. 2. Our method can be taken for conditional generation, where the condition of the control signal (text) provides a strong clue for capturing the complex data distribution, thus effectively enhancing the overall model's performance compared with naively applying model [85], as shown in Sec. 5. After training, the conditional generator is used to sample motion with a few denoising steps, which we discuss in Sec. 3.2.

Geometric Loss Functions. Moreover, during training, we found the training scheme of the conditional denoising diffusion GAN to be inefficient, resulting in low-quality human motion results (see comparisons in Sec. 5.2). We deem this is because motion generation requires more detailed constraints specifically tailored for the motion generation task, which cannot be effectively provided solely by the discrimination loss (Eq. 6). We thus employ geometric losses [76] in addition to discrimination loss during model training to enhance motion quality. Specifically, for generator (denoiser), we follow [7, 76] and predict the denoised motion itself, i.e., $\hat{\mathbf{x}}_0 = G(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)$ with the following losses on reconstruction, joint positions, foot contact, and joint velocities:

$$\mathcal{L}_{\text{recon}} = E_{\mathbf{x}_0 \sim q(\mathbf{x}_0|\mathbf{c}), t \sim [1,T]}[\|\mathbf{x}_0 - G_\theta(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t)\|_2^2], \tag{7}$$

$$\mathcal{L}_{\text{pos}} = \frac{1}{N} \sum_{i=1}^{N} \|FK(\mathbf{x}_0^i) - FK(\hat{\mathbf{x}}_0^i)\|_2^2, \tag{8}$$

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(FK(\hat{\mathbf{x}}_0^{i+1}) - FK(\hat{\mathbf{x}}_0^i)) \cdot f_i\|_2^2, \tag{9}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(\mathbf{x}_0^{i+1} - \mathbf{x}_0^i) - (\hat{\mathbf{x}}_0^{i+1} - \hat{\mathbf{x}}_0^i)\|_2^2 \tag{10}$$

The geometric loss is thus given by

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{recon}} + \lambda(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{foot}}). \tag{11}$$

Here, $FK(\cdot)$ denotes the forward kinematics converting joint rotations into joint positions. $f_i \in \{0,1\}^J$ is the binary foot contact mask for each frame i. Note that we use λ as a binary indicator variable; in this paper, we set λ to 1 and 0 for action-to-motion and text-to-motion tasks, respectively. We further investigate the effectiveness of the geometric loss functions in Sec. 5.2. Finally, we train the generator using the overall objective with a balancing term R:

$$\min_\theta(\mathcal{L}_{\text{disc}} + R \cdot \mathcal{L}_{\text{geo}}). \tag{12}$$

3.2 Motion Sampling

We adopt classifier-free guidance [26] in EMDM. Following [7], our generator G learns both the conditioned and the unconditioned motion generation task by randomly setting $\mathbf{c} = \emptyset$ for 10% of the samples, such that $G(\mathbf{x}_t, \mathbf{z}, t, \emptyset)$ approximates $p(\mathbf{x}_0)$. When sampling from G, we trade off diversity and fidelity by interpolating or even extrapolating the two variants using s:

$$G_s(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t) = G(\mathbf{x}_t, \mathbf{z}, \emptyset, t) + s \cdot (G(\mathbf{x}_t, \mathbf{z}, \mathbf{c}, t) - G(\mathbf{x}_t, \mathbf{z}, \emptyset, t)) \tag{13}$$

Given an input condition $\mathbf{c}$ which can be a sentence $\boldsymbol{w}^{1:N} = \{w^i\}_{i=1}^N$, a action label a from the predefined action categories set $a \in A$ [53] or even a empty condition $c = \varnothing$ [49,103], EMDM aims to generate a human motion $\hat{\mathbf{x}}^{1:N} = \{\hat{\mathbf{x}}^i\}_{i=1}^N$ in a non-deterministic way, where N denotes the motion length or frame number. Note that for the text-to-motion task, we employ the motion representation in [7,19,76,98]: a combination of 3D joint rotations, positions, velocities, and foot contact. The sampling algorithm is specified in Alg. 1.

4 Experiments

We conduct extensive experiments to evaluate our models on motion quality and model efficiency. We test our model on multiple datasets for different motion synthesis tasks, including text-to-motion (Sec. 4.3) and action-to-motion (Sec. 4.4),

Algorithm 1 Sample from Model

function SAMPLE($\mathbf{c}$)
  $\mathbf{x}_T \leftarrow$ random noise
  $T \leftarrow$ the number of time steps
  $G \leftarrow$ generator model
  $\mathbf{c} \leftarrow$ the label (text or action number)
  $\mathbf{x}_t \leftarrow \mathbf{x}_T$
  for $t \leftarrow T-1$ **to** 0 **do**
    $\mathbf{z} \leftarrow \text{RANDN}(\mathbf{z}_{dim})$
  $\triangleright$ dimension of $\mathbf{z}$ is 64 in our paper.
    $\mathbf{x}_0 \leftarrow \text{GENERATOR}(\mathbf{x}_t, t, \mathbf{z}, \mathbf{c})$
    $\mathbf{x}_t \leftarrow$ SAMPLE_POSTERIOR(ALG. 2)($\mathbf{x}_0, \mathbf{x}_t, t$)
  end for
  return $\mathbf{x}_t$
end function

Algorithm 2 Sample Posterior

function SAMPLE_POSTERIOR($\mathbf{x}_0$, $\mathbf{x}_t$, t)
  coef1 $\leftarrow$ posterior coefficient 1
  coef2 $\leftarrow$ posterior coefficient 2
  $mean \leftarrow \text{coef1}[t] \times \mathbf{x}_0 + \text{coef2}[t] \times \mathbf{x}_t$
  $log\_var \leftarrow$ posterior_log_variance[t]
  $noise \leftarrow \text{RANDN\_LIKE}(\mathbf{x}_t)$
  $m \leftarrow 0$ if $t = 0$ else 1
  return $mean + m \times \exp(0.5 \times log\_var) \times noise$
end function

qualitatively and quantitatively. The comparison with other few-step sampling methods for efficient motion generation is given in Appendix D1. We evaluate unconditional motion generation in Appendix B. More qualitative results can be found in Appendix A and the supplementary video. We also conduct comparisons with the original DDGAN [85] in Appendix D2.

Datasets. We use the following datasets for training and evaluating EMDM.
HumanML3D [19] has 14616 sequences from AMASS [46] with 44970 textual description.
KIT [57] collects 3911 motions with 6353 descriptions.
HumanAct12 [21] provides 1191 motion sequences and 12 action categories.
We use HumanML3D and KIT for the text-to-motion task while adopting HumanAct12 for the action-to-motion task. Refer to Appendix B for unconditional motion generation.

Metrics. We use the following metrics for evaluation.
Motion quality. We use Frechet Inception Distance (FID) as a principal metric to evaluate the feature distributions between the generated and real motions. The feature used is extracted following the previous approach in [19].
Motion diversity. Motion Diversity (DIV) calculates variance through features, and MultiModality (MM) measures the diversity using the same condition.
Condition matching. Following [19], we compute motion-retrieval precision (R Precision) and report the text and motion Top 1/2/3 matching accuracy, and multi-modal distance (MM Dist) is used to calculate the distance between motions and texts. For action-to-motion, we use the corresponding action recognition models [21, 53] to calculate Accuracy (ACC) for action categories.
Run-time Performance. We present the running time of various methods in milliseconds per frame as a metric to assess the inference efficiency of the models.

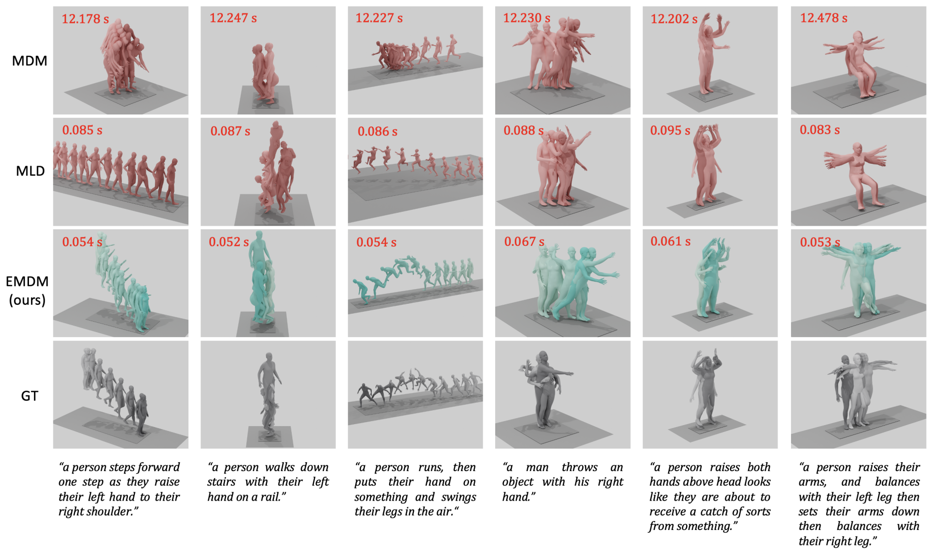

Fig. 4: Qualitative comparison on text-to-motion task. We visualize the generated motions and real references from six text prompts. EMDM achieves the fastest motion generation while delivering high-quality movements that align with the text input.

4.1 Implementation Details

We use a transformer-based denoiser ϵ_θ consisting of 12 layers and 32 heads with skip connections by default. The conditional discriminator is a 7-layer MLP network. The detailed architectures are given in Appendix C. We employ a frozen *CLIP-ViT-L-14* [60] model as the text encoder τ_θ^w for the text condition and a learnable embedding for action condition. All models are trained with the AdamW optimizer using a fixed learning rate of 3×10^{-5} and 2×10^{-5} for action-to-motion and text-to-motion, respectively. We use EMA decay on the optimizer during training. Our batch size is 64 during the training stage. The model is trained on an Nvidia RTX 4090 GPU with an AMD 16-core CPU. For inference, we use an RTX 4090 GPU with an Intel 8-core CPU to run all experiments under identical settings for ten passes.

4.2 Inference Time Costs

We first present the overall comparison of inference time cost on both text-to-motion and action-to-motion tasks. As demonstrated in Fig. 1 (c), on both tasks, EMDM demonstrates the best or second-best performance in FID, simultaneously achieving superior efficiency in motion generation. Notably, although MLD [7] and ReMoDiffuse [99] achieve competitive efficiency, these are two-stage methods that are non-end-to-end trainable. In contrast, EMDM exhibits a competitive or even better performance with reduced running time.

Table 1: Comparison of text-to-motion task on HumanML3D [19]. The right arrow $\rightarrow$ means the closer to real motion, the better.

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑	Running Time (per frame; ms)↓
	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$^{\pm.065}$	-	-
TEMOS [54]	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 [19]	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 [98]	0.491$^{\pm.001}$	0.681$^{\pm.001}$	0.782$^{\pm.001}$	0.630$^{\pm.001}$	3.113$^{\pm.001}$	**9.410$^{\pm.049}$**	1.553$^{\pm.042}$	38.235$^{\pm2.495}$
MDM [76]	0.418$^{\pm.005}$	0.605$^{\pm.005}$	0.708$^{\pm.005}$	0.508$^{\pm.034}$	3.630$^{\pm.023}$	9.373$^{\pm.094}$	2.880$^{\pm.088}$	62.505$^{\pm.071}$
MLD [7]†	0.481$^{\pm.003}$	0.673$^{\pm.003}$	0.772$^{\pm.002}$	0.473$^{\pm.013}$	3.196$^{\pm.010}$	9.724$^{\pm.082}$	2.413$^{\pm.079}$	**0.598$^{\pm.004}$**
T2M-GPT [95]†	0.492$^{\pm.003}$	0.679$^{\pm.002}$	0.775$^{\pm.002}$	0.141$^{\pm.005}$	3.121$^{\pm.009}$	9.722$^{\pm.082}$	1.831$^{\pm.048}$	0.886$^{\pm.007}$
MoFusion [15]	0.492$^{\pm.000}$	–	–	–	–	8.820$^{\pm.000}$	2.521$^{\pm.000}$	Not open source
M2DM [32]†	0.497$^{\pm.003}$	0.682$^{\pm.002}$	0.763$^{\pm.003}$	0.352$^{\pm.005}$	3.134$^{\pm.010}$	9.926$^{\pm.073}$	**3.587$^{\pm.072}$**	Not open source
ReMoDiffuse [99]†‡	**0.510$^{\pm.005}$**	**0.698$^{\pm.006}$**	**0.795$^{\pm.004}$**	**0.103$^{\pm.004}$**	**2.974$^{\pm.016}$**	9.018$^{\pm.075}$	1.795$^{\pm.043}$	0.959$^{\pm.002}$
EMDM (Ours)	**0.498$^{\pm.007}$**	**0.684$^{\pm.006}$**	**0.786$^{\pm.006}$**	**0.112$^{\pm.019}$**	**3.110$^{\pm.027}$**	**9.551$^{\pm.078}$**	1.641$^{\pm.078}$	**0.280$^{\pm.002}$**

Blue and orange indicate the best and the second best result.
† **Two-stage and non end-to-end approach.**
‡ **Reference dataset required at the inference stage.**

Table 2: Comparison of text-conditional motion generation on KIT [57].

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑	Running Time (per frame; ms)↓
	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$^{\pm.097}$	-	-
TEMOS	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	0.370$^{\pm.005}$	0.569$^{\pm.007}$	0.693$^{\pm.007}$	2.770$^{\pm.109}$	3.401$^{\pm.008}$	10.91$^{\pm.119}$	1.482$^{\pm.065}$	-
MotionDiffuse	0.417$^{\pm.004}$	0.621$^{\pm.004}$	0.739$^{\pm.004}$	1.954$^{\pm.062}$	2.958$^{\pm.005}$	**11.10$^{\pm.143}$**	0.730$^{\pm.013}$	68.403$^{\pm6.982}$
MDM	0.405$^{\pm.007}$	0.610$^{\pm.007}$	0.732$^{\pm.007}$	0.508$^{\pm.030}$	3.085$^{\pm.022}$	10.74$^{\pm.096}$	1.834$^{\pm.052}$	64.636$^{\pm.463}$
MLD [7]†	0.390$^{\pm.008}$	0.609$^{\pm.008}$	0.734$^{\pm.007}$	0.404$^{\pm.027}$	3.204$^{\pm.027}$	10.80$^{\pm.117}$	**2.192$^{\pm.071}$**	**0.673$^{\pm.008}$**
T2M-GPT [95]†	0.416$^{\pm.006}$	0.627$^{\pm.006}$	0.745$^{\pm.006}$	0.514$^{\pm.029}$	3.007$^{\pm.023}$	10.92$^{\pm.108}$	1.570$^{\pm.039}$	0.905$^{\pm.011}$
M2DM [32]†	0.416$^{\pm.004}$	0.628$^{\pm.004}$	0.743$^{\pm.004}$	0.515$^{\pm.029}$	3.015$^{\pm.017}$	11.42$^{\pm.970}$	**3.325$^{\pm.37}$**	Not open source
ReMoDiffuse [99]†‡	**0.427$^{\pm.014}$**	**0.641$^{\pm.004}$**	**0.765$^{\pm.055}$**	**0.155$^{\pm.006}$**	**2.814$^{\pm.012}$**	10.80$^{\pm.105}$	1.239$^{\pm.028}$	1.002$^{\pm.007}$
EMDM (Ours)	**0.443$^{\pm.006}$**	**0.660$^{\pm.006}$**	**0.780$^{\pm.005}$**	**0.261$^{\pm.014}$**	**2.874$^{\pm.015}$**	**10.96$^{\pm.093}$**	1.343$^{\pm.089}$	**0.284$^{\pm.002}$**

4.3 Comparisons on Text-to-motion

We evaluate EMDM on the text-to-motion task. We use the frozen CLIP [60] model as τ_θ^w to encode the text, giving $w_{clip}^1 \in \mathbb{R}^{1,024}$. The motion is then synthesized by conditioning on text input $\mathbf{c} = \{w^{1:N}\}$. We compare our model with SOTA methods on HumanML3D and KIT with the metrics proposed by [19]. Tab. 1 and Tab. 2 summarize the comparison results on HumanML3D [19] and KIT dataset [57], respectively. In Tab. 1, EMDM demonstrates the highest motion generation speed and highly competitive performance compared with the very recent reference-based [99] approach across all metrics, which validates its effectiveness and efficiency. In Tab. 2, EMDM consistently outperforms all existing methods in motion generation speed and Top1/2/3 matching accuracy. It produces competitive results on the remaining metrics. To summarize, EMDM demonstrates significant advantages compared to other methods, including multi-stage ones. Note that EMDM is single-stage and end-to-end trainable. Although ReMoDiffuse [99] attains competitive speeds in motion generation with relatively high quality, it employs a retrieval-based approach that relies on refer-

Table 3: Comparison of action-to-motion task on HumanAct12 [21]: $\text{FID}_{\text{train}}$ indicating the evaluated splits. Accuracy (ACC) for action recognition. Diversity (DIV) and MModality (MM) for generated motion diversity w.r.t each action label.

Methods	$\text{FID}_{\text{train}} \downarrow$	ACC $\uparrow$	DIV$\rightarrow$	MM$\rightarrow$	Running Time (per frame; ms)$\downarrow$
Real	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$	-
ACTOR [53]	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$\mathbf{6.840^{\pm.030}}$	$2.530^{\pm.020}$	$0.523^{\pm.009}$
INR [5]	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$	-
MDM [76]	$0.100^{\pm.000}$	$\mathbf{0.990^{\pm.000}}$	$\mathbf{6.860^{\pm.050}}$	$\mathbf{2.520^{\pm.010}}$	$41.154^{\pm.162}$
MLD [7]†	$\mathbf{0.077^{\pm0.004}}$	$0.964^{\pm.002}$	$6.831^{\pm0.050}$	$2.824^{\pm.038}$	$\mathbf{1.998^{\pm.001}}$
EMDM (Ours)	$\mathbf{0.084^{\pm.004}}$	$\mathbf{0.991^{\pm.003}}$	$\mathbf{6.876^{\pm.148}}$	$\mathbf{2.417^{\pm1.009}}$	$\mathbf{0.337^{\pm.005}}$

Blue and **orange** indicate the best and the second best result.
† **Two-stage and non end-to-end approach.**

ence motion datasets for generating motions at the inference stage. ReMoDiffuse is also a two-stage method that is non-end-to-end trainable. We provide qualitative results in Fig. 4, where EMDM achieves the fastest motion generation while maintaining competitive motion quality.

4.4 Comparisons on Action-to-motion

The action-conditioned task is to generate human motion given an action label. Following [7, 76], we report the FID, ACC, DIV, MM and Running Time of the aforementioned methods. The comparison on HumanAct12 [21] is shown in Tab. 3. EMDM achieves competitive results on HumanAct12 while achieving superior run-time performance. Notably, although MLD also uses less time for motion sampling, it is a two-stage method. The qualitative comparison of action-to-motion is visualized in Fig. 5, where EMDM achieves efficient motion generation performance while aligning with the semantics of the action label, while others have improper motion semantics, such as the "sit" motion of MDM and stiff "turn steering wheel" motion of MLD. See More results in the supplementary video.

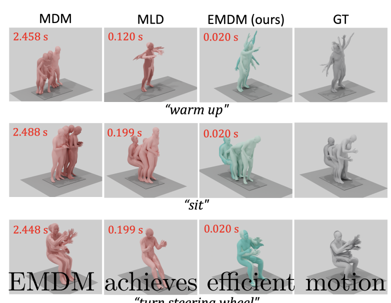

Fig. 5: Qualitative comparisons on the action-to-motion task.

5 Ablation Study

We validate the effectiveness of our key design choices in the following, with all experiments tested on HumanML3D [19] as text-to-motion is a more challenging task, compared to action-to-motion. The number of frames of the generated motion is 196. All models are trained with the same training settings. We also

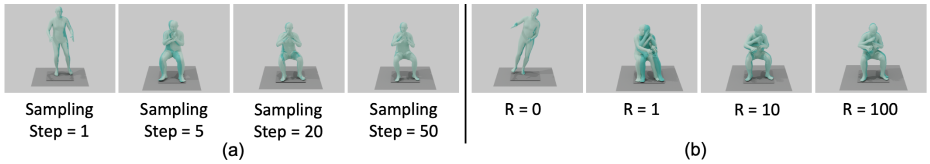

Fig. 6: Ablation studies on different sample steps (a) and weights of geometric loss (b) of generated motions. We use a classic textual description, "sit", as the input condition.

Table 4: Influence of sampling steps on motion generation using HumanML3D.

#Steps	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑	Running Time (per frame; ms)↓
	Top 1	Top 2	Top 3					
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-	-
1	$0.345^{\pm.005}$	$0.525^{\pm.007}$	$0.645^{\pm.007}$	$5.640^{\pm.127}$	$4.278^{\pm.021}$	$7.639^{\pm.071}$	$0.622^{\pm.016}$	$\mathbf{0.004^{\pm.000}}$
5	$0.368^{\pm.005}$	$0.547^{\pm.006}$	$0.655^{\pm.006}$	$1.306^{\pm.052}$	$4.047^{\pm.025}$	$9.168^{\pm.074}$	$\mathbf{2.285^{\pm.065}}$	$0.152^{\pm.000}$
10	$\mathbf{0.498^{\pm.007}}$	$\mathbf{0.684^{\pm.006}}$	$\mathbf{0.786^{\pm.006}}$	$\mathbf{0.112^{\pm.019}}$	$\mathbf{3.110^{\pm.027}}$	$9.551^{\pm.078}$	$1.641^{\pm.078}$	$0.280^{\pm.002}$
20	$0.490^{\pm.006}$	$0.679^{\pm.005}$	$0.780^{\pm.005}$	$0.191^{\pm.028}$	$3.142^{\pm.023}$	$9.531^{\pm.074}$	$1.688^{\pm.057}$	$0.555^{\pm.002}$
50	$0.479^{\pm.007}$	$0.671^{\pm.007}$	$0.770^{\pm.005}$	$0.216^{\pm.027}$	$3.168^{\pm.028}$	$\mathbf{9.482^{\pm.083}}$	$1.788^{\pm.046}$	$1.356^{\pm.000}$

study the performance of the model when trained without providing conditions to the discriminator with geometric loss in Appendix D3.

5.1 Influence of the Number of Sampling Steps

We investigate the influence of different sampling steps on the performance. We train and test our model with sampling step numbers $1, 5, 10, 20$ and 50. Notably, when the step number is set to 1, the whole model can be regarded as a GAN model. As shown in Tab. 4, when increasing the step size, the sampling speed is improved significantly. However, when the step size is too large, the motion quality indicated by FID, DIV, and MM drops. This is also witnessed by the qualitative results in Fig. 6 (a), where increasing sampling steps promote motion semantics, i.e., "sit". We consistently set the sampling step size to 10 in the experiments.

5.2 Influence of Geometric Loss

We study the influence of geometric loss during EMDM training. Recall the overall loss for our condition generator is denoted as $\mathcal{L} = \mathcal{L}_{\text{disc}} + R \cdot \mathcal{L}_{\text{geo}}$ (Eq. 12), where $\mathcal{L}_{\text{disc}}$ and $\mathcal{L}_{\text{geo}}$ represent the generator loss and geometric losses, respectively. Here, R serves as a balancing term. We evaluate the motion quality and running time by setting R to be $0.0, 1.0, 10.0, 100.0$ in Eq. 12. As shown in Tab. 5, when no geometric loss is applied, the motion quality significantly drops, e.g., FID $= 9.308$. Meanwhile, imposing geometric loss effectively improves the motion quality during the training process. We visualize the human motion under different weights R in Fig. 6 (b). In this paper, we empirically set the R to be 100.0 for text-to-motion tasks and 1.0 for action-to-motion tasks.

Table 5: Influence of geometric loss weights on motion generation using HumanML3D.

R value	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
0	$0.197^{\pm.005}$	$0.338^{\pm.006}$	$0.445^{\pm.006}$	$9.308^{\pm.190}$	$5.463^{\pm.027}$	$8.337^{\pm.086}$	$\mathbf{3.140^{\pm.079}}$
1	$0.468^{\pm.006}$	$0.656^{\pm.004}$	$0.761^{\pm.003}$	$0.449^{\pm.047}$	$3.272^{\pm.018}$	$9.445^{\pm.084}$	$1.978^{\pm.065}$
10	$0.486^{\pm.005}$	$0.672^{\pm.004}$	$0.768^{\pm.005}$	$0.232^{\pm.034}$	$3.169^{\pm.025}$	$9.347^{\pm.076}$	$1.706^{\pm.037}$
100	$\mathbf{0.498^{\pm.007}}$	$0.684^{\pm.006}$	$\mathbf{0.786^{\pm.006}}$	$\mathbf{0.112^{\pm.019}}$	$\mathbf{3.110^{\pm.027}}$	$\mathbf{9.551^{\pm.078}}$	$1.641^{\pm.078}$
1000	$0.494^{\pm.005}$	$\mathbf{0.685^{\pm.004}}$	$0.778^{\pm.005}$	$0.195^{\pm.026}$	$3.120^{\pm.022}$	$9.595^{\pm.084}$	$1.600^{\pm.045}$

6 Conclusion

In this paper, we reveal efficiency issues with the existing motion diffusion models and the challenges in accelerating the models. We introduce the Efficient Motion Diffusion Model (EMDM) to overcome the obstacles faced by existing generative diffusion models in achieving fast and high-quality motion generation. Different from previous approaches, we propose to sample motion from a diffusion model with much fewer sampling steps at the denoising stage. We utilize a conditional denoising diffusion Generative Adversarial Network to model the complex denoising distributions conditioning on the control signals. This enables the use of much larger step sizes, which in turn reduces the number of sampling steps while maintaining high motion quality and consistency in semantics with respect to the condition. We also incorporate a geometric loss to further elevate motion quality and enhance training efficiency. The whole model is end-to-end trainable. Consequently, EMDM achieves a remarkable speed-up without sacrificing motion quality when compared to current motion diffusion models, demonstrating its efficiency and effectiveness.
Limitations and Future Works. Although EMDM demonstrates encouraging performance in efficient human motion generation, its motion generation process lacks physical considerations, which may lead to issues like floating and ground penetration; See Fig. E5 in Appendix E. Efforts to integrate physics-based characters [16, 29, 50–52, 93] show promise for future improvements. In addition, currently EMDM accepts mainly textual inputs, but its potential extends to visual inputs [8, 17, 37, 41, 62, 64, 73, 83, 94] or music sources [2, 33, 36] for online motion synthesis, offering other exciting research directions.

Acknowledgements

The authors would like to thank Biao Jiang, Zhengyi Luo, Weilin Wan, and Chen Wang, as well as the AnySyn3D Team and the ECIG Team, for their engaging and insightful discussions. This work is partly supported by the Innovation and Technology Commission of the HKSAR Government under the ITSP-Platform grant (Ref: ITS/319/21FP) and the InnoHK initiative (TransGP project).

References

1. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019) 1, 4
2. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023) 2, 14
3. Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics (TOG) **41**(6), 1–19 (2022) 4
4. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. arXiv preprint arXiv:2303.14613 (2023) 4
5. Cervantes, P., Sekikawa, Y., Sato, I., Shinoda, K.: Implicit neural representations for variable length human motion generation. In: European Conference on Computer Vision. pp. 356–372. Springer (2022) 12
6. Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., Chen, X.: Taming diffusion probabilistic models for character control. arXiv preprint arXiv:2404.15121 (2024) 4
7. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023) 2, 5, 7, 8, 10, 11, 12
8. Chen, X., Su, Z., Yang, L., Cheng, P., Xu, L., Fu, B., Yu, G.: Learning variational motion prior for video-based motion capture. arXiv preprint arXiv:2210.15134 (2022) 14
9. Chong, E., Clark-Whitney, E., Southerland, A., Stubbs, E., Miller, C., Ajodan, E.L., Silverman, M.R., Lord, C., Rozga, A., Jones, R.M., et al.: Detection of eye contact with deep neural networks is as accurate as human experts. Nature communications **11**(1), 6386 (2020) 4
10. Chou, G., Bahat, Y., Heide, F.: Diffusion-sdf: Conditional generative modeling of signed distance functions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2262–2272 (2023) 5
11. Christen, S., Yang, W., Pérez-D'Arpino, C., Hilliges, O., Fox, D., Chao, Y.W.: Learning human-to-robot handovers from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9654–9664 (2023) 4
12. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 5
13. Cong, P., Dou, Z.W., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024) 4
14. Crawford, F.W., Jones, S.A., Cartter, M., Dean, S.G., Warren, J.L., Li, Z.R., Barbieri, J., Campbell, J., Kenney, P., Valleau, T., et al.: Impact of close interpersonal contact on covid-19 incidence: Evidence from 1 year of mobile device data. Science advances **8**(1), eabi5499 (2022) 4
15. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9760–9770 (2023) 1, 4, 11

16. Dou, Z., Chen, X., Fan, Q., Komura, T., Wang, W.: C·ase: Learning conditional adversarial skill embeddings for physics-based characters. arXiv preprint arXiv:2309.11351 (2023) 1, 4, 14
17. Dou, Z., Wu, Q., Lin, C., Cao, Z., Wu, Q., Wan, W., Komura, T., Wang, W.: Tore: Token reduction for efficient human mesh recovery with transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15143–15155 (2023) 14
18. Duan, Y., Shi, T., Zou, Z., Lin, Y., Qian, Z., Zhang, B., Yuan, Y.: Single-shot motion completion with transformer. arXiv preprint arXiv:2103.00776 (2021) 4
19. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022) 1, 4, 5, 8, 9, 11, 12
20. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022) 1, 4
21. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020) 1, 4, 5, 9, 12
22. Guo, Y., Dou, Z., Zhang, N., Liu, X., Su, B., Li, Y., Zhang, Y.: Student close contact behavior and covid-19 transmission in china's classrooms. PNAS nexus **2**(5), pgad142 (2023) 4
23. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion inbetweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020) 4
24. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 5
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 2, 6
26. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 8
27. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) 4
28. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36** (2024) 1, 5
29. Jiang, Y., Won, J., Ye, Y., Liu, C.K.: Drop: Dynamics responses from human motion prior and projective dynamics. arXiv preprint arXiv:2309.13742 (2023) 14
30. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023) 2, 5
31. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349 (2022) 1, 4, 5
32. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14806–14816 (2023) 1, 4, 11

33. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in neural information processing systems **32** (2019) 2, 4, 14
34. Lee, S., Starke, S., Ye, Y., Won, J., Winkler, A.: Questenvsim: Environment-aware simulated motion tracking from sparse sensors. arXiv preprint arXiv:2306.05666 (2023) 4
35. Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1231–1239 (2023) 1, 4
36. Li, B., Zhao, Y., Zhelun, S., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1272–1279 (2022) 2, 4, 14
37. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021) 14
38. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021) 2, 4
39. Li, T., Qiao, C., Ren, G., Yin, K., Ha, S.: Aamdm: Accelerated auto-regressive motion diffusion model. arXiv preprint arXiv:2401.06146 (2023) 4
40. Li, Z., Peng, X.B., Abbeel, P., Levine, S., Berseth, G., Sreenath, K.: Robust and versatile bipedal jumping control through reinforcement learning. Robotics: Science and Systems XIX, Daegu, Republic of Korea (2023) 4
41. Liao, Z., Golyanik, V., Habermann, M., Theobalt, C.: Vinecs: Video-based neural character skinning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1377–1387 (2024) 14
42. Liao, Z., Yang, J., Saito, J., Pons-Moll, G., Zhou, Y.: Skeleton-free pose transfer for stylized 3d characters. In: European Conference on Computer Vision. pp. 640–656. Springer (2022) 4
43. Liu, X., Dou, Z., Wang, L., Su, B., Jin, T., Guo, Y., Wei, J., Zhang, N.: Close contact behavior-based covid-19 transmission and interventions in a subway system. Journal of Hazardous Materials **436**, 129233 (2022) 4
44. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023) 5
45. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023) 5
46. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 9
47. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4328–4338 (2023) 5
48. Pang, K., Qin, D., Fan, Y., Habekost, J., Shiratori, T., Yamagishi, J., Komura, T.: Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. ACM Transactions on Graphics (TOG) **42**(4), 1–12 (2023) 4

49. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 4, 8
50. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG) **37**(4), 1–14 (2018) 14
51. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Transactions On Graphics (TOG) **41**(4), 1–17 (2022) 14
52. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. ACM Trans. Graph. **40**(4) (Jul 2021). https://doi.org/10.1145/3450626.3459670, http://doi.acm.org/10.1145/3450626.3459670 4, 14
53. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: International Conference on Computer Vision (ICCV) (2021) 1, 4, 5, 8, 9, 12
54. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022) 1, 4, 11
55. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022) 4
56. Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15061–15073 (2023) 4
57. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data **4**(4), 236–252 (dec 2016). https://doi.org/10.1089/big.2016.0028, http://dx.doi.org/10.1089/big.2016.0028 9, 11
58. Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J.T., Bermano, A.H., Chan, E.R., Dekel, T., Holynski, A., Kanazawa, A., et al.: State of the art on diffusion models for visual computing. arXiv preprint arXiv:2310.07204 (2023) 2, 5
59. Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.: Modi: Unconditional motion synthesis from diverse data. arXiv preprint arXiv:2206.08010 (2022) 4
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 10, 11
61. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020) 5
62. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: International Conference on Computer Vision (ICCV) (2021) 4, 14
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Con-

ference on Computer Vision and Pattern Recognition (CVPR) (2022), https://github.com/CompVis/latent-diffusionhttps://arxiv.org/abs/2112.10752 5
64. Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Transactions on Graphics (TOG) **40**(1), 1–15 (2020) 14
65. Shi, M., Starke, S., Ye, Y., Komura, T., Won, J.: Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14725–14737 (2023) 4
66. Shi, Y., Wang, J., Jiang, X., Dai, B.: Controllable motion diffusion model. arXiv preprint arXiv:2306.00416 (2023) 2, 5
67. Smith, L., Kew, J.C., Li, T., Luu, L., Peng, X.B., Ha, S., Tan, J., Levine, S.: Learning and adapting agile locomotion skills by transferring experience. arXiv preprint arXiv:2304.09834 (2023) 4
68. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015) 5, 6
69. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 2
70. Starke, S., Mason, I., Komura, T.: Deepphase: periodic autoencoders for learning motion phase manifolds. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022) 4
71. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6), 209–1 (2019) 4
72. Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021) 4
73. Sun, Q., Wang, Y., Zeng, A., Yin, W., Wei, C., Wang, W., Mei, H., Leung, C.S., Liu, Z., Yang, L., et al.: Aios: All-in-one-stage expressive human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1834–1843 (2024) 14
74. Tessler, C., Kasten, Y., Guo, Y., Mannor, S., Chechik, G., Peng, X.B.: Calm: Conditional adversarial latent models for directable virtual characters. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–9 (2023) 4
75. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. arXiv preprint arXiv:2203.08063 (2022) 5
76. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 1, 2, 4, 5, 6, 7, 8, 11, 12
77. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 5
78. Voas, J.: What is the best automated metric for text to motion generation? arxiv 2023. arXiv preprint arXiv:2309.10248 4
79. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. arXiv preprint arXiv:2311.17135 (2023) 4
80. Wan, W., Huang, Y., Wu, S., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Diffusionphase: Motion diffusion in frequency domain. arXiv preprint arXiv:2312.04036 (2023) 4

81. Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters **7**(2), 4702–4709 (2022) 4
82. Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters **7**(2), 4702–4709 (2022). https://doi.org/10.1109/LRA.2022.3151614 4
83. Wang, W., Ge, Y., Mei, H., Cai, Z., Sun, Q., Wang, Y., Shen, C., Yang, L., Komura, T.: Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. arXiv preprint arXiv:2303.13796 (2023) 14
84. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022) 4
85. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021) 3, 6, 7, 9
86. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023) 4, 5
87. Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., et al.: Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2228–2238 (2023) 1, 4
88. Yamane, K., Revfi, M., Asfour, T.: Synthesizing object receiving motions of humanoid robots with human motion database. In: 2013 IEEE International Conference on Robotics and Automation. pp. 1629–1636. IEEE (2013) 4
89. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4394–4402 (2019) 4
90. Yang, X., Dou, Z., Ding, Y., Su, B., Qian, H., Zhang, N.: Analysis of sars-cov-2 transmission in airports based on real human close contact behaviors. Journal of Building Engineering p. 108299 (2023) 4
91. Ye, Y., Liu, L., Hu, L., Xia, S.: Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. In: Computer Graphics Forum. vol. 41, pp. 183–194. Wiley Online Library (2022) 4
92. Yu, Z., Dou, Z., Long, X., Lin, C., Li, Z., Liu, Y., Müller, N., Komura, T., Habermann, M., Theobalt, C., et al.: Surf-d: High-quality surface generation for arbitrary topologies using diffusion models. arXiv preprint arXiv:2311.17050 (2023) 5
93. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16010–16021 (2023) 14
94. Zhang, H., Yuan, Y., Makoviychuk, V., Guo, Y., Fidler, S., Peng, X.B., Fatahalian, K.: Learning physically simulated tennis skills from broadcast videos. ACM Transactions On Graphics (TOG) **42**(4), 1–14 (2023) 14
95. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023) 1, 4, 5, 11
96. Zhang, J., Huang, S., Tu, Z., Chen, X., Zhan, X., Yu, G., Shan, Y.: Tapmo: Shape-aware motion generation of skeleton-free characters. arXiv preprint arXiv:2310.12678 (2023) 1, 4

97. Zhang, J., Weng, J., Kang, D., Zhao, F., Huang, S., Zhe, X., Bao, L., Shan, Y., Wang, J., Tu, Z.: Skinned motion retargeting with residual perception of motion semantics & geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13864–13872 (2023) 4
98. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 2, 4, 5, 6, 8, 11
99. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023) 1, 4, 5, 10, 11
100. Zhang, N., Liu, L., Dou, Z., Liu, X., Yang, X., Miao, D., Guo, Y., Gu, S., Li, Y., Qian, H., et al.: Close contact behaviors of university and school students in 10 indoor environments. Journal of Hazardous Materials **458**, 132069 (2023) 4
101. Zhang, N., Liu, X., Gao, S., Su, B., Dou, Z.: Popularization of high-speed railway reduces the infection risk via close contact route during journey. Sustainable Cities and Society **99**, 104979 (2023) 4
102. Zhang, Y., Black, M.J., Tang, S.: Perpetual motion: Generating unbounded human motion. arXiv preprint arXiv:2007.13886 (2020) 4
103. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3372–3382 (2021) 8
104. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023) 4, 5
105. Zhao, R., Su, H., Ji, Q.: Bayesian adversarial human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6225–6234 (2020) 4
106. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10553 (2023) 4