A Survey on Diffusion-based Motion Generation

Juncheng Ma

Abstract—Motion generation is a foundational task in artificial intelligence with broad practical applications. Compared to traditional methods such as VAE, GAN, and Normalizing Flows, diffusion models have become increasingly important in recent years and have emerged as a mainstream approach in motion generation. To harness the full potential of diffusion models, this survey comprehensively reviews literature on motion generation based on diffusion. Beginning with the fundamental principles of diffusion models, the survey focuses particularly on their evolution and conditional generation. After that, relative works on diffusion-based motion generation are categorized into three main types: motion diffusion models, controllability enhencement, and data availability, providing detailed discussions on each type. This survey aims to serve as a comprehensive guide, offering insights into the current landscape of diffusion-based motion generation and highlighting promising avenues for future research. Relevant representative articles on diffusion-based motion generation are compiled and organized, accessible at here.

Index Terms—Human motion generation, diffusion model, literature survey.

1 Introduction

Synthesizing realistic human motion sequences has been a cornerstone in various real-world applications, such as animation production, game development, virtual and augmented reality[1], [2], and robotic control[3], [4], [5]. Despite numerous technological breakthroughs over the decades, traditional methods to design and synthesize human motion sequences using professional software remain labor-intensive, and motion capture equipment is expensive. The high demand for specialized equipment and professional skills hinders the customization and democratization of motion generation for widespread content creation. To eliminate technical barriers and extend accessibility to a broader audience, developing human motion generation models that automatically produce diverse and controllable motion sequences has become a significant research direction.

Over the past few decades, human motion generation has been a key task in artificial intelligence and has garnered increasing attention in recent years[6], [7], [8]. Some early works focused on unconditional motion generation[9], [10], [11], [12]. With the rapid development of deep generative models[13], [14], [15] and the emergence of large-scale motion datasets[16], [17], [18], [19], [20], [21], recent years have witnessed breakthroughs in the effectiveness of motion generation and a surge in related works integrating extensive multimodal conditional signals, such as action[22], [23], [24], [25], audio[26], [27], [28], [29], [30], [31], text[32], [33], [34], [35], [36], scene[37], [38], [39] and incomplete pose sequences[40], [41], [42], [43].

Significant progress has been made in conditional motion generation over the past few years, leading to diverse and distinctive methodological approaches. Early works[44], [45], [46] primarily focused on deterministically **aligning** motion sequences with language descriptions to obtain a joint embedding space. To enhance the diversity of motion sequences, **Generative Adversarial Net (GAN)** modeling was employed to introduce randomness[47], [48], [49]. Additionally, several studies[50], [51], [52], [53], [54], [23], [55] incorporated variational mechanisms to adopt **Variational Autoencode (VAE)**[15]. In contrast to GAN and VAE, **Normalizing Flows (NF)**[56] explicitly learn the data distribution to generate conditional motions[57], [58], [59], [60]. Inspired by successes in language[61] and image generation[62], **autoregressive models**[63] have recently been extensively explored in the motion generation domain to further improve motion generation quality[64], [65], [66], [67], [68], [54]. Motions are transformed into discrete tokens through vector quantization (VQ)[69], which are then fed into autoregressive models to generate motion token sequences unidirectionally.

Diffusion models[13], [70], [71], which have achieved remarkable success in visual generation tasks[72], [73], [74], have become one of the most popular approaches for motion generation in recent years[6], [7], [32], [8], [75], [76]. These models effectively overcome the challenges of aligning posterior distributions within VAEs, mitigate the instability inherent in the adversarial objectives of GANs, address the computational burdens of Markov Chain Monte Carlo methods in Energy-Based Models[77], [78], and enforce network constraints similar to those in NFs[79]. Unlike the previously dominant VAE-based pipelines, motion diffusion models enhance generative capabilities through a stochastic diffusion process, which models complex distributions, resulting in diverse and high-fidelity motion sequences. Furthermore, diffusion models preserve the format of the original motion sequences during generation, allowing for more convenient constraints during the denoising process to ensure physical plausibility and realism. Despite significant progress in motion diffusion models, the research community remains highly active, frequently producing novel insights focusing on specific application scenarios such as controllability[1], [80], [81], [36], data availability[82], [83], [35], [84], generalization[85], and physical-plausibility[76], [6], [32].

Juncheng Ma is with University of Chinese Academy of Sciences, Beijing, China. E-mail: majuncheng21@mails.ucas.ac.cn.

1.1 Scope

This survey primarily focuses on diffusion-based motion generation methods, covering the most recent literature and representing the cutting edge of this domain, with a particular emphasis on conditional generation driven by text descriptions. We provide a comprehensive review on various aspects of these methods and discuss several challenges and potential future directions for diffusion-based motion generation.

The survey by [86] offered a detailed and comprehensive review of human motion generation. However, the field has progressed rapidly since its publication in 2022, rendering the works included less representative of current development trends. For example, among the 75 representative works selected in [86], the mainstream methods were VAE (31/75) and GAN (17/75), with only 10 based on diffusion models. Currently, the most advanced motion generation methods are based on diffusion models and newly emerging autoregressive models. Therefore, this survey serves as a supplement to [86], summarizing representative works and outlining several noteworthy and valuable research directions in diffusion-based motion generation. We aim to provide valuable insights and inspire researchers to tackle the challenges in this evolving field.

1.2 Organization

The rest of the survey is organized as follows. In Section 2, we give a brief introduction of the principles of diffusion models and human motion representations. We then categorize the majority of diffusion-based motion generation methods into three main types. In Section 3, we discuss works that build effective motion diffusion models, which apply diffusion models to the motion domain. In Section 4, we explore methods that enhance controllability through text control or spatial constraint enhancements. In Section 5, we review approaches addressing motion generation in data-limited scenarios. In Section 6, we highlight several other notable research directions. Finally, we draw conclusions and suggest worthwhile future research directions in Section 7.

2 PRELIMINARIES

2.1 Diffusion Model

First, we introduce the basic principles of diffusion models to provide foundational knowledge. Diffusion models[13] consist of two interconnected processes: a predefined forward process that maps the data distribution to a simpler prior distribution, often Gaussian, and a corresponding reverse process that employs a trained neural network to gradually reverse the effects of the forward process. In Section 2.1.1, we formalize the problem, starting with the classical Denoised Diffusion Probabilistic Models (DDPM) [13]. Next, in Section 2.1.2, we explain the extension of DDPM from a discrete-time process to a continuous-time framework based on stochastic differential equations [71]. Finally, we discuss conditional generation by presenting the principles of the two most popular forms of guidance: Classifier Guidance [87] and Classifier-Free Guidance [88].

In composing this subsection, we referenced the survey [79]. For comprehensive information on the principles and recent advancements in diffusion models, we suggest referring to the survey [79], as well as two informative tutorials [89], [90].

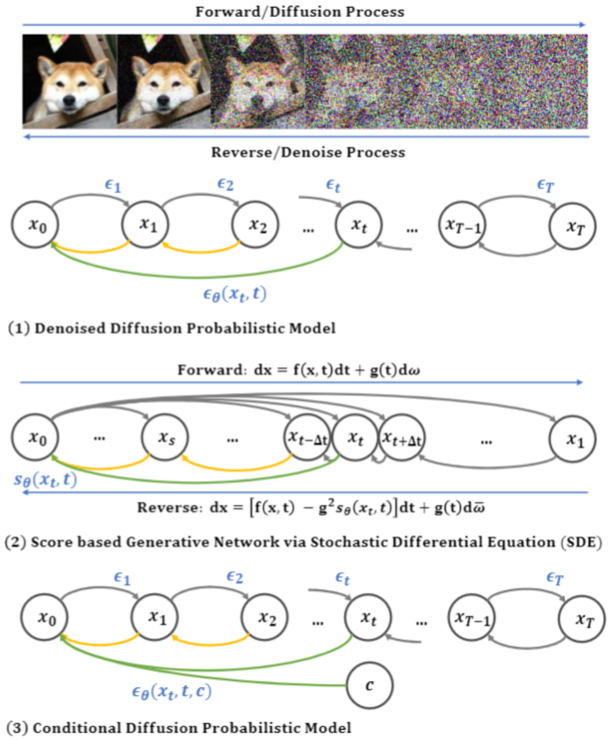
Forward/Diffusion Process
Reverse/Denoise Process
ϵ_1
ϵ_2
ϵ_t
ϵ_T
x_0
x_1
x_2
x_t
x_{T-1}
x_T
$\epsilon_\theta(x_t, t)$
(1) Denoised Diffusion Probabilistic Model
Forward: $dx = f(x,t)dt + g(t)d\omega$
x_0
x_s
$x_{t-\Delta t}$
x_t
$x_{t+\Delta t}$
x_1
$s_\theta(x_t, t)$
Reverse: $dx = [f(x,t) - g^2 s_\theta(x_t,t)]dt + g(t)d\bar{\omega}$
(2) Score based Generative Network via Stochastic Differential Equation (SDE)
ϵ_1
ϵ_2
ϵ_t
ϵ_T
x_0
x_1
x_2
x_t
x_{T-1}
x_T
c
$\epsilon_\theta(x_t, t, c)$
(3) Conditional Diffusion Probabilistic Model

Fig. 1. **Overview of Diffusion Models.**[79]

2.1.1 Denoised Diffusion Probabilistic Models

Problem Formalization. Diffusion Models can be formalized as a Markov chain $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$, where $\mathbf{x}_1, ..., \mathbf{x}_T$ are the noised sequences distorted from the real data $\mathbf{x}_0 \sim p_0(\mathbf{x}_0)$. All variables x_t share the same dimensionality.

Forward Process In the DDPM framework, to approximate posterior $q(\mathbf{x}_{1:t}|\mathbf{x}_0)$, the forward process follows a Markov chain that gradually adds Gaussian noises to the data until its distribution is close to the latent distribution $\mathcal{N}(\mathbf{0}, \mathcal{I})$ according to variance schedules given by $\beta_1, ..., \beta_t$:

$$\begin{aligned} q(\mathbf{x}_{1:T}|\mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}). \end{aligned} \tag{1}$$

Reverse Process. The reverse process $p_\theta(\mathbf{x}_{0:T})$ is also a Markov chain that predicts and eliminates the noise with learned Gaussian transitions starting at $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$:

$$\begin{aligned} p_\theta(\mathbf{x}_{0:T}) &:= p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &:= \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_t). \end{aligned} \tag{2}$$

$\mu_\theta(\mathbf{x}_t, t)$ represent learnable mean of reverse Gaussian kernels determined by reverse-step distribution p_θ. Σ_t denotes a fixed variance corresponding to the predefined hyperparameter β_t in DDPM.

Training. To approximate the data distribution p_0 by the joint probability distribution $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$, DDPM should maximize its log-likelihood of $\mathbf{x}_0$. However, direct optimization of the likelihood function $log\ p_\theta(\mathbf{x}_0)$ is challenging. Therefore, DDPM reformulates the optimization objective to maximize the Evidence Lower Bound (ELBO) through a series of derivations. Based on this equivalent objective, DDPM constructs and optimizes a neural network to model $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and obtains training objective based on the posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, x_0)$.

Reparameterization offers three equivalent implementation methods for modeling $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, as predicting noise ϵ, origin samples $\mathbf{x}_0$, or score functions $\triangledown log\ p(\mathbf{x})$ (as detailed in Section 2.1.2). Taking ϵ as an example, the noise prediction model $\epsilon_\theta(\mathbf{x}_t, t)$ are optimized to minimize a simplified training objective as

$$\mathcal{L} = \mathrm{E}_{t\in[1,T],\mathbf{x}_0\sim p_0(\mathbf{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathcal{I})}[\|\ \epsilon - \epsilon_\theta(\mathbf{x}_t, t)\ \|]. \tag{3}$$

To efficiently acquire $\mathbf{x}_t$ from $\mathbf{x}_0$, [13] formulate the diffusion process as

$$q(\mathbf{x}_t|\mathbf{x}_0) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Hence, we can simply sample a noise ϵ and t to directly generate $\mathbf{x}_t$ by this formulation.

Sampling. To generate diverse results, we denoise the sequence from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathcal{I})$. During the sampling process, we can sample $\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_t)$ step by step and finally get a clean result $\hat{\mathbf{x}}_0$. The mean $\mu_\theta(\mathbf{x}_t, t)$ can be acquired from $\mathbf{x}_t$ and $\epsilon_\theta(\mathbf{x}_t, t)$ by the following equation

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)). \tag{5}$$

Advancements. Following DDPM, several notable advancements have contributed to its improvement. DDIM[70] extended DDPM's training from Markov to non-Markovian processes, introducing an accelerated sampling method. iDDPM[91] introduced learnable variance to boost the log-likelihood of DDPM. Analytic-DPM[92] demonstrated the existence of an analytic solution for variance in the reverse process, significantly enhancing sampling speed and quality by correcting variance terms during sampling. LDM[93] applies diffusion processes on latent space, notably improving efficiency and enabling synthesis of higher-resolution images. DiT[94] introduced a transformer-based LDM model with superior scalability.

2.1.2 *Score-based Diffusion Models*

DDPM operates as a discrete process in both the forward and reverse phases. However, by introducing Stochastic Differential Equations (SDE), ScoreSDE[71] extends the discrete-time scheme to a continuous-time framework, allowing DDPM to be viewed as a discrete approximation at various levels (T). This enables the use of existing techniques in the ODE/SDE community to diffusion models for theoretical analysis while employing appropriate discretization in practice. Furthermore, [71] introduces deterministic samplers based on the probability flow Ordinary Differential Equation (ODE), which facilitates fast adaptive sampling via black-box ODE solvers, flexible data manipulation through latent codes, uniquely identifiable encoding, and, notably, exact likelihood computation.

Forward SDE. Instead of perturbing data with a finite number of noise distributions, ScoreSDE considers a continuum of distributions that evolve over time according to a diffusion process. This process progressively diffuses a data point into random noise, and is given by a prescribed SDE that does not depend on the data and has no trainable parameters. This can be modeled as the solution to an Itô SDE:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{6}$$

where $\mathbf{w}$ is the standard Wiener process (a.k.a., Brownian motion), $\mathbf{f}(\cdot, t)$ is a vector-valued function called the *drift* coefficient of $\mathbf{x}(t)$, and $g(\cdot)$ is a scalar function known as the *diffusion* coefficient of $\mathbf{x}(t)$. The SDE has a unique strong solution as long as the coefficients are globally Lipschitz in both state and time.

Reverse SDE. By reversing the forward process, ScoreSDE can smoothly mold random noise $\mathbf{x}_T \sim p(x_T)$ into clean data for sample generation. Crucially, this reverse process satisfies a reverse-time SDE[95], which can be derived from the forward SDE.

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}. \tag{7}$$

where $\bar{\mathbf{w}}$ is a standard Wiener process when time flows backwards from T to 0, and $\mathrm{d}t$ is an infinitesimal negative timestep. Once the score of each marginal distribution, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, is known for all t, we can derive the reverse diffusion process from eq. (7) and simulate it to sample from p_0. p_t and p_0 are the marginal and prior distributions.

Training Objective. To approximate the reverse-time SDE, a time-dependent score-based model s is trained to estimate the scores $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. However, due to the difficulty in obtaining the ground truth for $\nabla_{\mathbf{x}_t} \log p_t(x)$, ScoreSDE adopts the score matching objective, enabling $\mathbf{s}_\theta(\mathbf{x}_t, t)$ to predict $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)$ at any $\mathbf{x}_t$, and then takes the expectation over $\mathbf{x}_t$:

$$L = E_t\{\lambda(t)E_{\mathbf{x}_0}E_{q(\mathbf{x}_t|\mathbf{x}_0)}[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)\|_2^2]\}. \tag{8}$$

Here $\lambda(t)$ is a positive weighting function, t is uniformly sampled over $[0, T]$, $q(\mathbf{x}_t|\mathbf{x}_0)$ is the Gaussian transition kernel associated with the forward process in eq. (6). If $\mathbf{f}(\mathbf{x}_t)$ is an affine transformation, then $p(\mathbf{x}_t|\mathbf{x}_0)$ is also a tractable Gaussian distribution. With sufficient data and model capacity, score matching ensures that the optimal solution to eq. (8), denoted by $\mathbf{s}_\theta(\mathbf{x}, t)$, equals $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ for almost all $\mathbf{x}$ and t. After obtaining the score-based model $\mathbf{s}_\theta(\mathbf{x}_t, t)$, ScoreSDE can generate diverse samples using numerical SDE solvers.

Probability Flow ODE. [71] introduced another numerical method probability flow ODE for solving the reverse-time SDE. For all diffusion processes, there exists a corresponding deterministic process which shares the same marginal probability density $\{p_t(\mathbf{x})\}_{t=0}^{T}\}$ with SDE in eq. (7), which satisfies an ODE:

$$\mathrm{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g(t)^2\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\Big]\mathrm{d}t. \tag{9}$$

Similar to how SDEs extend DDPM into continuous time, ODEs represent the continuous-space counterpart of DDIM. When the variance is set to zero, SDEs degenerate into ODEs.Unlike SDEs, probability flow ODEs can be solved with larger step sizes due to the absence of randomness. Consequently, several studies[96], [97], [98], [99] have achieved faster sampling speeds using advanced ODE solvers.

2.1.3 Conditional Generation

In the previous sections, we focused on modeling the original data distribution $p(\mathbf{x})$. However, in practice, modeling conditional distributions $p(\mathbf{x} \mid c)$ is often more significant, particularly in motion generation. This allows us to explicitly control the samples generated through an input conditional signal c.

Conditional generation primarily comprises two methods: Classifier Guidance[87] and Classifier-Free Guidance[88]. The former involves training a classifier on top of a pre-trained unconditional diffusion model and applying gradient guidance only during sampling, significantly reducing training costs. In contrast, the latter entails training the model with conditional inputs from scratch, allowing for more fine-grained control.

Classifier Guidance. Using the score-based formulation as an example (as described in Section 2.1.2), we can derive the following equivalent form[89] by Bayes rule, for learning $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c)$

$$\begin{aligned}\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|c) &= \nabla_{\mathbf{x}_t} \log\left(\frac{p(\mathbf{x}_t)p(c|\mathbf{x}_t)}{p(c)}\right)\\ &= \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}_{\text{unconditional score}} + \underbrace{\nabla_{\mathbf{x}_t} \log p(c|\mathbf{x}_t)}_{\text{adversarial gradient}}\end{aligned} \tag{10}$$

where we utilize the fact that the gradient of $\log p(c)$ with respect to $\mathbf{x}_t$ is zero.

Our final derived result in eq. (10) can be interpreted as learning an unconditional score function combined with the adversarial gradient from a classifier $p(\mathbf{x}|c)$. Initially, a classifier is trained to predict the conditional signal y from $\mathbf{x}_t$ at arbitrary noise levels t. During sampling, the conditional scoring function is computed as a weighted sum of the unconditional score and the adversarial gradient from the noise classifier. To control the consideration of conditional signal, classifier guidance employs gradient scaling through γ. The scoring function learned under classifier guidance can be summarized as follows:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \gamma \nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) \tag{11}$$

Based on the above derivation, regarding DDPM, the modification required primarily involves adjusting the sampling process to

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t) + \gamma \Sigma_t \nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t), \Sigma_t). \tag{12}$$

Classifier-Free Guidance. In contrast to Classifier Guidance, Classifier-Free Guidance directly defines $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, c, t), \Sigma_t)$ during training, thereby transforming the training objective eq. (3) to

$$\mathcal{L} = \mathrm{E}_{t\in[1,T],\mathbf{x}_0\sim p_0(\mathbf{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathcal{I})}[\| \ \epsilon - \epsilon_\theta(\mathbf{x}_t, c, t) \ \|]. \tag{13}$$

Specifically, Classifier-Free Guidance also incorporates γ for gradient scaling, aiming to balance relevance and diversity. Similar to eq. (12), it introduces $\omega = \gamma - 1$, which modifies the sampling process to

$$\widetilde{\epsilon}_\theta(\mathbf{x}_t, c, t) = (1+\omega)\epsilon_\theta(\mathbf{x}_t, c, t) - \omega\epsilon_\theta(\mathbf{x}_t, t). \tag{14}$$

To obtain unconditional output $\epsilon_\theta(\mathbf{x}_t, t)$, a specific conditional input ϕ, corresponding to all images, is introduced. This input appears with a certain probability during training, thereby acquiring unconditional output $\epsilon_\theta(\mathbf{x}_t, t) = \epsilon_\theta(\mathbf{x}_t, \phi, t)$.

2.2 Motion Data Representation

Motion data is typically collected using methods such as motion capture[100], [101], [102], [17], pseudo-labeling[103], [104], [105], [106], and manual annotation[107], [108], [109], [110]. The motion sequence is commonly represented as $m_{1:N} = \{m_i\}_{i=1}^N$, where $m_i \in \mathbb{R}^D$ denotes the pose state in the i-th frame, and N is the total number of frames. In this subsection, we briefly introduce motion data representation, which is categorized into keypoint-based and rotation-based approaches[86].

Keypoint-based representation identifies specific joints or other critical locations on the body as keypoints, depicting the entire body pose through these points. Each keypoint is defined by its 2D/3D coordinates in either pixel or world coordinates. Motion data is then represented as a sequence of keypoint configurations over time. Alternatively, motion sequences can be represented using **rotation-based representation**. This method utilizes joint angles to depict the human body's posture, leveraging the hierarchical structure of the body to record rotations of each joint relative to its parent. These rotations are parameterized using formats such as Euler angles and quaternions. Moreover, statistical mesh models like Skinned Multi-Person Linear (SMPL) [111] can model the body with rotation-based representation, effectively capturing shape variations and deformations during body movements.

Each of these two methods has its own advantages. Keypoint-based representation can be directly obtained from motion capture systems, making it more accessible and interpretable. Rotation-based representation can utilize statistical mesh models for human body modeling, facilitating applications in animation and robotics. Additionally, conversion between these two types of representations is possible using forward kinematics (FK) and inverse kinematics (IK).

3 Motion Diffusion Model

Given the impressive results achieved by diffusion models in various generative domains[72], [112], [113], [114], early studies[8], [6], [115] have applied these models to motion generation. **Motion Diffusion Model (MDM)**[6] introduces a classifier-free, diffusion-based generative model specifically designed for human motion. MDM utilizes the DDPM framework, following the methodology of [73] by directly predicting the original motion sequences during denoising. This approach enables the application of geometric losses, validated in the motion domain, directly during the generation process. These losses constrain positions, foot contacts, and velocities, enhancing physical properties, preventing artifacts, and promoting natural and coherent motion. As

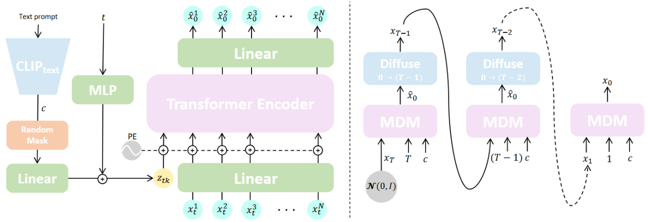

Fig. 2. An overview of Motion Diffusion Model (MDM).[6]

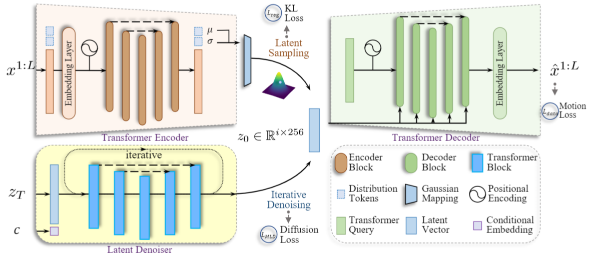

Fig. 3. Motion Latent Diffusion Model(MLD) overview.[7]

illustrated in Fig. 2, MDM employs a denoising model based on the transformer[116] architecture. It first uses the text encoder of CLIP to extract features from the text prompt, which are then combined with the timestamp encoding to form one of the input tokens z_{tk}. This token is then added to the position embedding and input into the transformer encoder along with other noised motion tokens. Additionally, the paper proposes an inpainting-based[117], [118] motion editing method, where the part to be edited is noised while the other parts remain unchanged, followed by iterative denoising to obtain the modified motion sequence. Due to the lack of relevant data, MDM faces challenges in generating long multi-segment motions, multi-person motions, and fine-grained control, such as trajectory and end-effector tracking. To address these issues, **PriorMDM**[119] introduces three combinations based on diffusion priors (pre-trained MDM) aimed at expanding MDM's application scenarios through few-shot fine-tuning or zero-shot correction during sampling. For generating arbitrarily long motions, PriorMDM proposes the DoubleTake method, which can concatenate short motions (10 seconds) generated by MDM into long sequences, with transitions added between segments. This allows each segment of the motion to be controlled by different text prompts and vary in sequence length. To generate two-person actions, the paper proposes a few-shot method and designs the ComMDM module (a single-layer transformer) to coordinate between two frozen instances of pre-trained MDM. For achieving fine-tuned motion control, the paper employs an inpainting method to fine-tune specific control conditions and uses model composition to combine different single control conditions.

MotionDiffuse[8], a contemporary work of MDM, designed a transformer-based model for noise prediction using the DDPM architecture, integrating text features into the generation process via cross-attention. To enhance efficiency, linear attention was employed in both the self-attention and cross-attention layers, thereby reducing computational complexity. To incorporate the diffusion time condition, time t was encoded and added to the text features before inputting them into the Stylization Block. Notably, MotionDiffuse implemented two fine-grained controllable generation methods: Body Part-independent Controlling and Time-varied Controlling. **FLAME**[115] is also one of the earliest works to apply diffusion models to motion generation, using learnable variance following iDDPM[91] and additionally set an ML token to indicate the length of the motion sequence.

To optimize efficiency and reduce computational overhead, **Motion Latent Diffusion (MLD)** [7], building upon LDM, operates within a lower-dimensional latent space. As illustrated in Fig. 3, MLD first trains a motion variational autoencoder (VAE). This VAE utilizes transformer-based encoder to process original motion tokens and distribution tokens, predicting μ and σ, and is trained with Kullback–Leibler divergence loss. In the decoding phase, L zero motion tokens function as queries, while latent representations serves as both keys and values in the decoder's cross-attention layers. This process generates L tokens supervised by MSE loss relative to the original motion tokens. Once the motion VAE is trained, it achieves a compact, efficient latent space, effectively filtering out high-frequency and imperceptible details. Subsequently, MLD trains a diffusion model based on transformers in the latent space. In contrast to LDM, which conducts diffusion in continuous feature spaces, **priority-centric motion discrete diffusion model (M2DM)**[75] explores diffusion within discrete spaces. This approach begins by training a Vector Quantized Variational Autoencoder (VQ-VAE)[69] to encode a rich codebook for motion sequences, followed by applying diffusion to the discrete latent representations. The paper also introduces specific noise injection and denoising procedures tailored for discrete diffusion.

Since motions generated dirrectly by DDPM cannot guarantee physical and anatomical plausibility, they may exhibit artifacts such as motion jitter, illegal skeletal structures, and foot sliding. To enhance the realism of the motions, **Mofusion**[32] introduces kinematic loss to constrain skeletal motion, ensuring that bone lengths remain consistent over time and symmetrical between the left and right sides. Through reparameterization, Mofusion transforms the predicted noise to the motion domain, making it easier and more intuitive to apply kinematic constraints. As illustrated in Fig. 4, **Physdiff**[76] proposes a simple yet effective physics-based projection strategy to improve the physical plausibility of the pretrained T2M model during sampling. Specifically, Physdiff employs a reinforcement learning approach to train a motion simulator. The agent generates physically plausible next-frame motions based on the current state and the predicted next-frame motion from the denoising model. This Markov process produces a sequence of physically plausible motions. After obtaining this physics-based projection, Physdiff applies it at each sampling time step to ensure the generated motion sequence is physically plausible. Several noteworthy works further explore the potential of Motion Diffusion Model. **InterGen**[120] focuses on generating motion involving multiple individuals and introduces a multimodal dataset, InterHuman, for two-person interactions. To achieve dual-person motion generation, InterGen

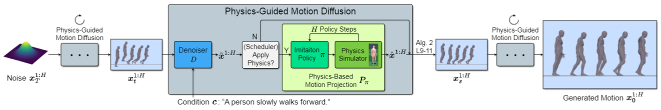
Physics-Guided Motion Diffusion
Noise $x_T^{1:H}$
Physics-Guided Motion Diffusion
$x_t^{1:H}$
Denoiser D
$\tilde{x}^{1:H}$
(Scheduler) Apply Physics?
N
Y
H Policy Steps
Imitaiton Policy π
Physics Simulator
Physics-Based Motion Projection P_π
$\hat{x}^{1:H}$
Alg. 2 L9-11
$x_s^{1:H}$
Physics-Guided Motion Diffusion
Generated Motion $x_0^{1:H}$
Condition c: "A person slowly walks forward."

Fig. 4. An overview of Physdiff.[76]

employs two collaborative transformers for denoising. These transformers share weights and process each individual separately while connecting features at different levels through a novel mutual attention mechanism. Additionally, InterGen proposes an effective motion representation and two auxiliary regularization losses to model the complex spatial relationships in human interactions. **Ude**[29] introduces a Modality-Agnostic Transformer Encoder (MATH) to achieve multimodal motion consistency, unifying text and audio conditions. It also presents a generative method combining VQ-VAE and diffusion models. VQ-VAE encodes discrete code tokens to ensure the quality and semantic consistency of long sequences, while diffusion models act as the decoder to enhance diversity. Once the conditional tokens are obtained, they are input as queries into a transformer to autoregressively generate a latent representation sequence of motion, which is then decoded by the diffusion model to obtain the generated motion sequence. The **Accelerated Autoregressive Motion Diffusion Model (AAMDM)**[121] also combines autoregressive and diffusion models, integrating denoising diffusion GANs for fast generation and using autoregressive diffusion models as a polishing module, significantly improving generation efficiency.

4 Controllability Enhancement

Despite the rapid progress in conditional motion generation, such as text-driven and audio-driven approaches that generate diverse and realistic human motions, users still face challenges. Text descriptions often lack detailed specificity and sufficient information density, making it difficult to control specific posture details at precise moments in synthesized motion sequences. Moreover, current methods struggle to achieve precise control. Spatial control signals, such as trajectories and keyframes, are crucial for controllable motion generation. Therefore, in this section, we will discuss diffusion-based controllable motion generation, focusing on enhancing text control and spatial constraints.

4.1 Text Control Enhancement

As shown in Fig. 5, **FineMoGen**[122] aims to achieve fine-grained spatio-temporal motion generation by enriching text descriptions to provide detailed spatio-temporal information, enhancing control at any stage or for any body part. However, past datasets lacked such fine-grained descriptions. To address this, this paper expands the HuMMan[18] dataset and introduces a new dataset that divides each sequence into multiple temporal stages, providing both overall text descriptions and fine-grained descriptions for seven body parts at each stage. Additionally, FineMoGen introduces a novel attention module, Spatio-Temporal Mixture Attention, to model temporal-spatial constraints. The increased focus on finer-grained regions significantly raises the complexity of model learning, prompting FineMoGen to apply a Sparsely-activated Mixture of Experts (MoE)[123] to enhance the model's representation learning capability. To extend application scenarios, FineMoGen also leverages large language models to enable users to achieve fine-grained control and editing using natural language.

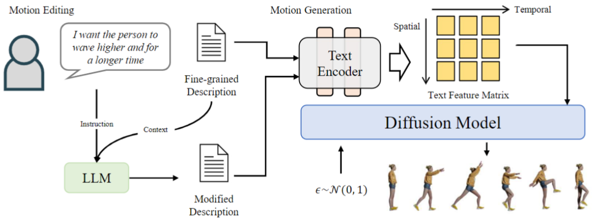

Fig. 5. An overview of FineMoGen.[122]

Unlike FineMoGen, several works[124], [125], [36] focus on enhancing text control for either temporal or body-part aspects. To achieve long, complex text control and variable-length motion generation, **Autoregressive Motion Diffusion (AMD)**[124] introduces a new text-motion dataset where the text descriptions consist of short sentences, each corresponding to a single motion segment. AMD combines autoregressive models with diffusion models to predict each motion segment autoregressively, ultimately generating the entire sequence. Simultaneously, **FlowMDM**[125] addresses the long text motion generation task, which generates long, continuous sequences guided by long texts describing each time clip separately. FlowMDM, based on a diffusion model, fully controls the sequence and duration of the target motion while achieving seamless and realistic transitions between motions, eliminating the need for post-processing or redundant denoising steps. For fast sampling, Blended Positional Encodings are introduced, incorporating both absolute and relative position encodings. Additionally, a Pose-Centric Cross-Attention mechanism ensures that each pose is denoised based on its own conditions and neighboring poses. Beyond long texts, **LGTM**[36] employs a large language model to decompose the global motion description into part-specific narratives, enhancing control over body parts through enriched text descriptions. These descriptions are processed by independent body-part motion encoders to ensure precise local semantic alignment. Subsequently, an attention-based full-body optimizer refines the motion generation results and ensures overall coherence.

The previously introduced works enhance text control by *explicitly* augmenting the text prompts. In contrast, **Fg-t2m**[126] *implicitly* enhances text control by thoroughly analysis of text prompts and is the first to introduce NLP methods into the Text-to-Motion (T2M) field. Previous T2M approaches primarily used text encoders to extract textual features directly, often failing to enable the model to accurately understand the text. Fg-t2m proposes a novel method for parsing and integrating textual features to support precise text descriptions. Specifically, Fg-t2m designs the Linguistics-Structure Assisted Module, which employs dependency parsing trees and graph networks to extract textual features, thereby preserving the linguistic structure of the text while

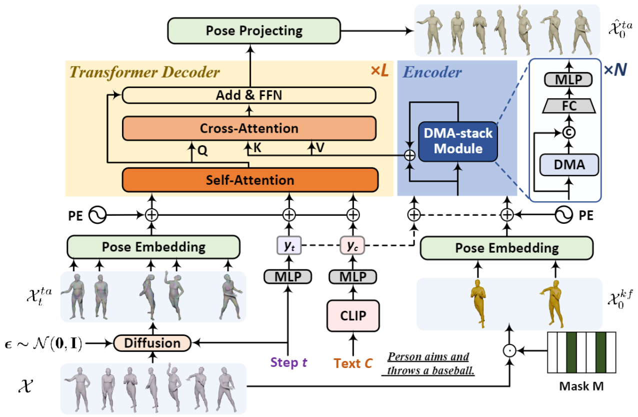

Fig. 6. An overview of DiffKFC.[42]

aggregating semantic content. Additionally, the Context-Aware Progressive Reasoning Module is introduced, which mimics human reasoning by implementing a multi-step strategy that progresses from global (semantic) to local (word-level) understanding.

4.2 Spatial Constraints Enhencement

Addressing the challenge of accurately generating motion from textual descriptions alone, recent research has prioritized improving motion controllability through the integration of spatial constraints[42], [43], [80], [127]. **DiffKFC**[42] focuses on text-to-motion generation specifically under sparse keyframe constraints. Previous inpainting methods struggled with sparse keyframes, often treating them as noise during denoising and failing to obtain effective guidance. DiffKFC addresses this issue by incorporating keyframe awareness into its training process, enabling KeyFrames Collaboration during training. The model architecture, illustrated in Fig, 6, involves encoding the motion sequence into tokens with added noise, which are then inputted into a transformer decoder along with timestamps and text tokens. Sparse keyframes are expanded to match the sequence length and encoded to extract features, which are subsequently conditionally fused with the noisy motion sequence in the cross-attention layers of the transformer decoder. The paper also introduces the Dilated Mask Attention (DMA) Module to effectively utilize keyframes, progressively expanding sparse keyframe features and iteratively aggregating them to obtain dense features. Furthermore, **CondMDI**[43] addresses human motion generation guided by keyframes, allowing for flexible placement of dense or sparse keyframes to generate diverse, high-quality motions aligned with the provided keyframes. The **Guided Diffusion Model (GMD)**[80] tackles motion generation under spatial constraints such as trajectories and obstacles, proposing an effective feature projection scheme to enhance spatial information and ensure motion consistency. For sparse keyframe conditions, GMD offers a method to efficiently convert sparse signals into dense guidance.

In addition to trajectories and keyframes, other research[1], [81], [128] has explored broader spatial constraints. **AGRoL**[1] focuses on scenarios involving AR/VR head-mounted devices, predicting full-body poses from sparse upper-body signals using a diffusion model based on MLP. **OmniControl**[128] integrates flexible spatial control signals across different joints and times, providing spatial and realism guidance. The objective functions are meticulously crafted for gradient-guidance during generation processes. Furthermore, **programmable motion generation**[81] decomposes arbitrarily complex spatial control signals into combinations of atomic constraints such as trajectories, keyframes, and interactions. It quantifies each atomic constraint error and aggregates errors for a flexible and fully customizable set of motion controls. Following the error computation, it utilizes a pretrained MDM and optimizes its latent code to minimize the error function.

5 Data Availability

Given the challenges in acquiring the text2motion dataset, most motion generation models rely heavily on meticulously annotated motion datasets. To mitigate this dependency, several approaches have emerged to improve data availability, enabling adaptation of motion generation models in scenarios with limited data.

An intuitive idea to enhance data availability involves expanding the dataset. **Make-An-Animation (MAA)**[82] utilizes the text2image dataset for pre-training to broaden the domain coverage of generated results from Text2Motion models. Initially, MAA filters out human body pose-related data from the text2image dataset and extracts key points using a pre-trained model to derive 3D pose features, organizing them into a Text Pseudo-Pose (TPP) dataset. The training process consists of two stages. In the first stage, a text2pose diffusion model is trained on the TPP dataset, depicted in Fig. 7. The second stage involves fine-tuning on the text2motion dataset, integrating 1D temporal convolution and temporal attention layers into ResNet and attention blocks as adapters. Additionally, MAA utilizes motion data without captions for classifier-free training. Similarly, **Multi-view Ancestral Sampling (MAS)**[84] leverages in-the-wild video data, extracting 2D poses for training. During training, multiple 2D motion sequences representing different views of the same 3D motion are simultaneously denoised. At each step, these sequences are constrained to a unified 3D sequence based on their poses and then projected back to the original view, ensuring consistency across all views in each diffusion step. **MotionMix**[83] proposes a straightforward and effective weakly supervised diffusion model trained using noisy and unlabeled data. Through a two-stage training process, the model initially generates coarse motions based on conditions and subsequently refines these motions into high-quality outputs, effectively leveraging noisy and clean motion sequences.

In addition to expanding the dataset, several intriguing approaches[35], [129], [130] address the challenge of data scarcity in motion generation. **Open-vocabulary Motion Generation (OMG)**[35] adopts a pretrain-then-finetune paradigm to generate motions under zero-shot open-vocabulary text prompts. Initially, OMG scales up DiT and pretrains on a large-scale unlabeled motion dataset to learn diverse out-of-domain motion characteristics. During fine-tuning, it introduces Motion ControlNet[131] and incorporates the proposed Mixture-of-Controllers (MoC)

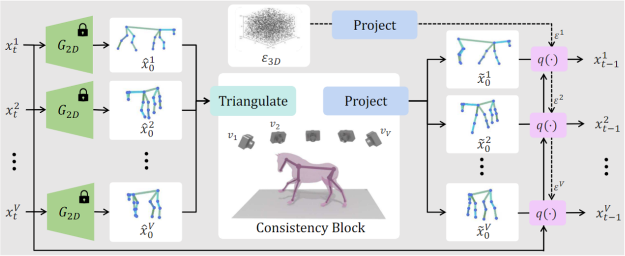

Fig. 7. An overview of MAA.[82]

block. Utilizing cross-attention mechanisms between text and motion, OMG adaptively identifies sub-motions and processes them with different experts, effectively enhancing alignment between text descriptions and generated motions. To mitigate domain dependencies, **Retrieval-Augmented Motion Diffusion Model (ReMoDiffuse)**[129] employs a retrieval-augmented approach to generate diverse motions by leveraging knowledge from retrieval samples. It calculates text similarity using CLIP during retrieval and incorporates a coefficient related to motion length to retrieve the most similar samples. Features from these retrieval samples serve as conditional features during the generation process. [130], diverging from diffusion models, explores preference learning in T2M by learning from preference pairs rather than annotated motion data. Preference learning proves highly effective due to simplified data collection, enabling cost-effective training of T2M models without expert annotation.

Recent attention has focused on motion generation challenges for more generalizable subjects. Due to significant differences in posture attributes, such as skeleton and geometric structures, across various subjects, motion generation models trained on human data cannot be directly applied to non-human subjects like dogs or dragons. This limitation severely restricts the potential applications of motion generation models. For non-human skeletal structures, such as animals (e.g., dragons), motion data is often scarce. Therefore, **SinMDM**[85] introduces a one-shot method capable of modeling and synthesizing motions for any skeletal topology, faithfully generating motions of arbitrary length even when only one animation sequence is available for training. To capture local motion sequences, SinMDM employs a combination of local attention and a shallow U-Net[132] for denoising, aiming to reduce the receptive field size to enhance diversity and mitigate overfitting. **OmniMotionGPT**[133] introduces a VAE-based method that released an animal motion generation dataset. It proposes a technique to generate diverse animal motions based on limited animal data, leveraging prior knowledge acquired from human data and facilitating knowledge transfer.

6 Other Notable Directions

The preceding sections summarized research primarily focused on text-driven action generation. In recent years, there have also been advancements in audio-driven dance generation leveraging diffusion models. Early work [31] specifically addressed audio-driven dance and speech-driven gesture generation. The model architecture derived from DiffWave[112], which is designed originally for audio generation, and explored methods using Mixture of Experts (MoE). **EDGE**[28] applied MDM to dance generation, utilizing Jukebox[134], a music-pretrained model, for extracting musical features. **FineDance**[30] introduced a comprehensive 3D dance generation dataset covering various genres and proposed a model tailored for choreography. The core model structure, based on MDM, facilitated fine-grained generation of hands and limbs through expert networks, integrating genre and coherence matching via a Genre & Coherent Aware Retrieval Module.

Similarly, diffusion-based methods have emerged for scene-driven motion generation[37], [38], [39]. Wang et al. [38] focuses on language-guided human motion generation in 3D scenes, using features such as beds and walls represented by RGB point clouds as input conditions. This approach effectively achieve motion generation despite data constraints involving language-scene-motion pairs. Since scene-aware human motion generation struggles to generalize interactions with new objects outside the training distribution, **ROAM** [37] ensures robustness and generalization to new scene objects in 3D object-aware character synthesis by training the motion model with as few as one reference object. Given a set of sparse joint locations and a seed motion sequence in a 3D scene, the method by Mir et al. [39] can generate continuous motion that adapts to diverse scenes.

Recently, new generative models have emerged in the field of motion generation, demonstrating capabilities comparable to diffusion models. Autoregressive motion models, such as MotionGPT[66], HumanTOMATO[68], and Motiongpt[67], have been developed to further enhance motion generation quality. These models first train a VQ-VAE to encode motion into codes, which serve as tokens and are input into a transformer model along with conditional information for autoregressive motion generation. However, because the unidirectional decoding of autoregressive models may limit their expressive capacity, generative masked transformers have been introduced, such as MoMask[33] and MMM[34]. These models use an auto-encoding training method similar to BERT[135], where some tokens are randomly masked and input into a transformer encoder along with text descriptions to predict the complete tokens. During sampling, a fully masked token sequence is input, and the complete motion sequence is generated iteratively.

7 Conclusion and Future Work

Diffusion models are becoming increasingly crucial in the field of motion generation. To harness their potential, this paper provides a comprehensive and up-to-date review of diffusion-based motion generation. We begin with the fundamental principles of diffusion models, focusing on their development and conditional generation. Subsequently, we categorize diffusion-based motion generation works according to their research contributions and discuss each category in detail. We hope this survey serves as a comprehensive guide for readers, clarifying the progress in diffusion-based motion generation and highlighting several promising research directions.

Despite rapid progress, significant challenges remain that warrant further exploration. Therefore, we outline several promising future directions from different perspectives, aiming to inspire new breakthroughs in human motion generation research.

7.1 Advanced Foundational Models

While motion diffusion models have been extensively studied, further developing foundational generative models remains critical for advancing motion generation. Just as recent innovations such as autoregressive models and masked generative models have emerged, future research could explore integrating the strengths and core concepts of different advanced generative models. Moreover, looking ahead, although iterative denoising represents the current state-of-the-art, it may not constitute the ultimate solution. After all, humans do not create images solely from noise. Therefore, future research should continue to investigate more advanced foundational generative models to enhance modeling capabilities and efficiency in capturing data distributions.

7.2 Controllable Generation

Controllable motion generation has seen extensive exploration in recent years. However, improving the controllability of models under various input conditions remains a critical challenge. Moreover, current control signals (*e.g.*, long texts, trajectories, keyframes) are less user-friendly for the general public. Exploring ways to enhance the controllability of models, supporting more flexible and interactive motion control methods, merits further consideration in the future.

7.3 Data Availability

As discussed in section 5, diffusion models often face challenges in identifying patterns and regularities from low-quality data, which hinders their ability to generalize to new scenarios or datasets. In practical motion generation scenarios, text2motion data is often limited, which significantly affects the implementation of Motion Generation methods. Approaches explored by SinMDM[85] are valuable for transferring knowledge learned from data-rich human skeleton motion data to generate motion in other species. We anticipate that data availability will continue to pose a long-term challenge for data-driven motion diffusion models.

7.4 Effiencincy Improvement

Despite their impressive ability to produce high-quality and diverse results, diffusion models suffer from low efficiency for a long time. Their slow generation speed restricts their practical applications, largely due to iterative denoising sampling strategies. Future research should focus on developing methods to improve both training and sampling efficiency of motion diffusion model.

7.5 Evaluation Metrics

While numerous evaluation metrics exist in the field of motion generation, they all possess inherent limitations. The challenge persists in proposing evaluation metrics that closely align with human standards. Future research should prioritize the development of more principled objective metrics that are closely linked to human perception and maintain interpretability.

References

[1] Y. Du, R. Kips, A. Pumarola, S. Starke, A. Thabet, and A. Sanakoyeu, "Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 481–490.

[2] A. Castillo, M. Escobar, G. Jeanneret, A. Pumarola, P. Arbeláez, A. Thabet, and A. Sanakoyeu, "Bodiffusion: Diffusing sparse observations for full-body human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4221–4231.

[3] F. Xia, C. Li, R. Martín-Martín, O. Litany, A. Toshev, and S. Savarese, "Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation," in *ICRA*. IEEE, 2021, pp. 4583–4590.

[4] S. Jauhri, S. Lueth, and G. Chalvatzaki, "Active-perceptive motion generation for mobile manipulation," *arXiv preprint arXiv:2310.00433*, 2023.

[5] Y. Nishimura, Y. Nakamura, and H. Ishiguro, "Long-term motion generation for interactive humanoid robots using gan with convolutional network," in *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 375–377.

[6] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *ICLR*, 2023.

[7] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[8] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[9] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4394–4402.

[10] R. Zhao, H. Su, and Q. Ji, "Bayesian adversarial human motion synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6225–6234.

[11] Y. Zhang, M. J. Black, and S. Tang, "Perpetual motion: Generating unbounded human motion," *arXiv preprint arXiv:2007.13886*, 2020.

[12] S. Raab, I. Leibovitch, P. Li, K. Aberman, O. Sorkine-Hornung, and D. Cohen-Or, "Modi: Unconditional motion synthesis from diverse data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 873–13 883.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[14] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014.

[16] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[17] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.

[18] Z. Cai, D. Ren, A. Zeng, Z. Lin, T. Yu, W. Wang, X. Fan, Y. Gao, Y. Yu, L. Pan *et al.*, "Humman: Multi-modal 4d human dataset for versatile sensing and modeling," in *European Conference on Computer Vision*. Springer, 2022, pp. 557–577.

[19] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[20] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[22] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2021–2029.

[23] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 985–10 995.

[24] P. Cervantes, Y. Sekikawa, I. Sato, and K. Shinoda, "Implicit neural representations for variable length human motion generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 356–372.

[25] T. Lucas, F. Baradel, P. Weinzaepfel, and G. Rogez, "Posegpt: Quantization-based 3d human motion generation and forecasting," in *European Conference on Computer Vision*. Springer, 2022, pp. 417–435.

[26] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, and X. Wang, "Tm2d: Bimodality driven 3d dance generation via music-text integration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9942–9952.

[27] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, "Bailando: 3d dance generation by actor-critic gpt with choreographic memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 050–11 059.

[28] J. Tseng, R. Castellon, and K. Liu, "Edge: Editable dance generation from music," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 448–458.

[29] Z. Zhou and B. Wang, "Ude: A unified driving engine for human motion generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5632–5641.

[30] R. Li, J. Zhao, Y. Zhang, M. Su, Z. Ren, H. Zhang, Y. Tang, and X. Li, "Finedance: A fine-grained choreography dataset for 3d full body dance generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 234–10 243.

[31] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.

[32] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, "Mofusion: A framework for denoising-diffusion-based motion synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9760–9770.

[33] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1900–1910.

[34] E. Pinyoanuntapong, P. Wang, M. Lee, and C. Chen, "Mmm: Generative masked motion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1546–1555.

[35] H. Liang, J. Bao, R. Zhang, S. Ren, Y. Xu, S. Yang, X. Chen, J. Yu, and L. Xu, "Omg: Towards open-vocabulary motion generation via mixture of controllers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 482–493.

[36] H. Sun, R. Zheng, H. Huang, C. Ma, H. Huang, and R. Hu, "Lgtm: Local-to-global text-driven human motion diffusion model," *SIGGRAPH*, 2024.

[37] W. Zhang, R. Dabral, T. Leimkühler, V. Golyanik, M. Habermann, and C. Theobalt, "Roam: Robust and object-aware motion generation using neural pose descriptors," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 1392–1402.

[38] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Move as you say interact as you can: Language-guided human motion generation with scene affordance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 433–444.

[39] A. Mir, X. Puig, A. Kanazawa, and G. Pons-Moll, "Generating continual human motion in diverse 3d scenes," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 903–913.

[40] F. G. Harvey, M. Yurick, D. Nowrouzezahrai, and C. Pal, "Robust motion in-betweening," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 60–1, 2020.

[41] Y. Duan, T. Shi, Z. Zou, Y. Lin, Z. Qian, B. Zhang, and Y. Yuan, "Single-shot motion completion with transformer," *arXiv preprint arXiv:2103.00776*, 2021.

[42] D. Wei, X. Sun, H. Sun, S. Hu, B. Li, W. Li, and J. Lu, "Enhanced fine-grained motion diffusion for text-driven human motion synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5876–5884.

[43] S. Cohan, G. Tevet, D. Reda, X. B. Peng, and M. van de Panne, "Flexible motion in-betweening with diffusion models," *SIGGRAPH*, 2024.

[44] C. Ahuja and L.-P. Morency, "Language2pose: Natural language grounded pose forecasting," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 719–728.

[45] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek, "Synthesis of compositional animations from textual descriptions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1396–1406.

[46] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision*. Springer, 2022, pp. 358–374.

[47] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 366–382.

[48] Z. Wang, P. Yu, Y. Zhao, R. Zhang, Y. Zhou, J. Yuan, and C. Chen, "Learning diverse stochastic human-action generators by learning smooth latent transitions," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 281–12 288.

[49] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1418–1427.

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

[51] S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould, "A stochastic conditioning scheme for diverse human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5223–5232.

[52] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action composition for 3d humans," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 414–423.

[53] M. Petrovich, M. J. Black, and G. Varol, "Temos: Generating diverse human motions from textual descriptions," in *European Conference on Computer Vision*. Springer, 2022, pp. 480–497.

[54] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2m-gpt: Generating human motion from textual descriptions with discrete representations," *arXiv preprint arXiv:2301.06052*, 2023.

[55] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.

[56] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[57] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[58] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 93–98.

[59] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 401–13 412.

[60] G. Valle-Pérez, G. E. Henter, J. Beskow, A. Holzapfel, P.-Y. Oudeyer, and S. Alexanderson, "Transflower: probabilistic autoregressive dance generation with multimodal attention," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–14, 2021.

[61] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[62] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[63] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[64] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, and X. Wang, "Tm2d: Bimodality driven 3d dance generation via music-text integration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9942–9952.

[65] C. Guo, X. Zuo, S. Wang, and L. Cheng, "Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts," in *European Conference on Computer Vision*. Springer, 2022, pp. 580–597.

[66] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[67] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang, "Motiongpt: Finetuned llms are general-purpose motion generators," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7368–7376.

[68] S. Lu, L.-H. Chen, A. Zeng, J. Lin, R. Zhang, L. Zhang, and H.-Y. Shum, "Humantomato: Text-aligned whole-body motion generation," *arXiv preprint arXiv:2310.12978*, 2023.

[69] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[70] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *ICLR*, 2021.

[71] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021.

[72] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[73] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[74] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[75] H. Kong, K. Gong, D. Lian, M. B. Mi, and X. Wang, "Priority-centric human motion generation in discrete latent space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 806–14 816.

[76] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, "Physdiff: Physics-guided human motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 010–16 021.

[77] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[78] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng, "Learning deep energy models," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 1105–1112.

[79] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[80] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Gmd: Controllable human motion synthesis via guided diffusion models," *arXiv preprint arXiv:2305.12577*, 2023.

[81] H. Liu, X. Zhan, S. Huang, T.-J. Mu, and Y. Shan, "Programmable motion generation for open-set motion control tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1399–1408.

[82] S. Azadi, A. Shah, T. Hayes, D. Parikh, and S. Gupta, "Make-an-animation: Large-scale text-conditional 3d human motion generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 039–15 048.

[83] N. M. Hoang, K. Gong, C. Guo, and M. B. Mi, "Motionmix: Weakly-supervised diffusion for controllable motion generation," *arXiv preprint arXiv:2401.11115*, 2024.

[84] R. Kapon, G. Tevet, D. Cohen-Or, and A. H. Bermano, "Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1965–1974.

[85] S. Raab, I. Leibovitch, G. Tevet, M. Arar, A. H. Bermano, and D. Cohen-Or, "Single motion diffusion," *ICLR*, 2024.

[86] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[87] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[88] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *NIPSw*, 2021.

[89] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022.

[90] S. H. Chan, "Tutorial on diffusion models for imaging and vision," *arXiv preprint arXiv:2403.18103*, 2024.

[91] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[92] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," *arXiv preprint arXiv:2201.06503*, 2022.

[93] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[94] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[95] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[96] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[97] ——, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[98] K. Zheng, C. Lu, J. Chen, and J. Zhu, "Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[99] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," *arXiv preprint arXiv:2202.09778*, 2022.

[100] J. Liao, C. Luo, Y. Du, Y. Wang, X. Yin, M. Zhang, Z. Zhang, and J. Peng, "Hardmo: A large-scale hardcase dataset for motion capture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1629–1638.

[101] J. P. Araújo, J. Li, K. Vetrivel, R. Agarwal, J. Wu, D. Gopinath, A. W. Clegg, and K. Liu, "Circle: Capture in rich contextual environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 211–21 221.

[102] S. Ghorbani, Y. Ferstl, D. Holden, N. F. Troje, and M.-A. Carbonneau, "Zeroeggs: Zero-shot example-based gesture generation from speech," in *Computer Graphics Forum*, vol. 42, no. 1. Wiley Online Library, 2023, pp. 206–216.

[103] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, and M. J. Black, "Generating holistic 3d human motion from speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 469–480.

[104] N. Le, T. Pham, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen, "Music-driven group choreography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8673–8682.

[105] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt, "Learning speech-driven 3d conversational gestures from video," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021, pp. 101–108.

[106] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3497–3506.

[107] K. Chen, Z. Tan, J. Lei, S.-H. Zhang, Y.-C. Guo, W. Zhang, and S.-M. Hu, "Choreomaster: choreography-oriented music-driven dance synthesis," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[108] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "Danceformer: Music conditioned 3d dance generation with parametric motion transformer,"

in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1272–1279.

[109] J. Gao, J. Pu, H. Zhang, Y. Shan, and W.-S. Zheng, "Pc-dance: Posture-controllable music-driven dance synthesis," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1261–1269.

[110] Z. Wang, J. Jia, H. Wu, J. Xing, J. Cai, F. Meng, G. Chen, and Y. Wang, "Groupdancer: Music to multi-people dance synthesis with style collaboration," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1138–1146.

[111] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[112] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[113] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.

[114] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[115] J. Kim, J. Kim, and S. Choi, "Flame: Free-form language-based motion synthesis & editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8255–8263.

[116] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[117] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.

[118] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[119] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," *ICLR*, 2024.

[120] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "Intergen: Diffusion-based multi-human motion generation under complex interactions," *International Journal of Computer Vision*, pp. 1–21, 2024.

[121] T. Li, C. Qiao, G. Ren, K. Yin, and S. Ha, "Aamdm: Accelerated auto-regressive motion diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1813–1823.

[122] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, and Z. Liu, "Finemogen: Fine-grained spatio-temporal motion generation and editing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[123] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.

[124] B. Han, H. Peng, M. Dong, Y. Ren, Y. Shen, and C. Xu, "Amd: Autoregressive motion diffusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2022–2030.

[125] G. Barquero, S. Escalera, and C. Palmero, "Seamless human motion composition with blended positional encodings," *arXiv preprint arXiv:2402.15509*, 2024.

[126] Y. Wang, Z. Leng, F. W. Li, S.-C. Wu, and X. Liang, "Fg-t2m: Fine-grained text-driven human motion generation via diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 035–22 044.

[127] M. Diomataris, N. Athanasiou, O. Taheri, X. Wang, O. Hilliges, and M. J. Black, "Wandr: Intention-guided human motion generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 927–936.

[128] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "Omnicontrol: Control any joint at any time for human motion generation," *ICLR*, 2024.

[129] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu, "Remodiffuse: Retrieval-augmented motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 364–373.

[130] J. Sheng, M. Lin, A. Zhao, K. Pruvost, Y.-H. Wen, Y. Li, G. Huang, and Y.-J. Liu, "Exploring text-to-motion generation with human preference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1888–1899.

[131] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[132] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[133] Z. Yang, M. Zhou, M. Shan, B. Wen, Z. Xuan, M. Hill, J. Bai, G.-J. Qi, and Y. Wang, "Omnimotiongpt: Animal motion generation with limited data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1249–1259.

[134] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[135] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.