基于扩散模型的动作生成综述

马俊程

摘要：动作生成作为人工智能领域的一项基础任务，在现实世界中拥有广泛的应用前景。近年来，扩散模型在动作生成领域的重要性日益凸显，逐渐成为该领域的主导技术。为了深入挖掘扩散模型的潜力，本研究对基于扩散模型的动作生成方法进行了全面的文献综述。本综述首先介绍了扩散模型的基本原理，并特别强调了其发展轨迹和条件生成方法。随后，我们将相关研究根据科研贡献分为三个主要类别——运动扩散模型、可控性增强和数据可用性，并详细讨论了每一类任务。我们期望本综述能够为读者提供一份全面的指导，明确基于扩散模型的动作生成研究的现状，并指出未来研究的方向，以期促进运动生成研究的进一步发展。我们对基于扩散模型的动作生成相关文献进行了整理和统计，详细信息可参见GitHub 仓库。

关键词：动作生成，扩散模型，生成模型，文献综述

1 引言

合成逼真的人类运动序列一直是许多现实应用的基石，例如动画制作、游戏开发、虚拟现实和增强现实 [1, 2]，以及机器人控制 [3, 4, 5] 等。尽管数十年来动作生成技术在工业界取得了显著进步，但使用专业软件设计和合成人类运动序列的传统方法仍然需要大量的劳动力，且昂贵的动作捕捉设备大大提高了成本和技术门槛。对专业设备和专业技能的高要求限制了运动生成的个性化和普及化，影响了内容创作的广泛性。为了消除这些技术障碍并提高运动生成的可及性，开发能够自动生成多样化和可控制运动序列的人类运动生成模型，已成为该领域的一个重要方向。

在过去的几十年中，人类运动生成一直是人工智能领域的一个关键任务，并在近几年得到了越来越多的关注 [6, 7, 8]。一些早期的工作集中在无条件运动生成上 [9, 10, 11, 12]。随着深度生成模型的快速发展 [13, 14, 15] 和大规模运动数据集的出现 [16, 17, 18, 19, 20, 21]，运动生成模型的效果在过去几年里得到了巨大突破，并涌现了大量整合了多种不同模态条件信号的相关研究，例如动作 [22, 23, 24, 25]、声音 [26, 27, 28, 29, 30, 31]、文本 [32, 33, 34, 35, 36]、场景 [37, 38, 39] 和不完整姿势序列 [40, 41, 42, 43]。

条件运动生成领域在取得了显著进展的同时，涌现出多样化和具有特色的研究方法。早期研究 [44, 45, 46] 主要集中于确定性地将运动序列与语言描述**对齐**，以获得联合嵌入空间。为了增强运动序列的多样性，采用了**生成对抗网络（GAN）**建模引入随机性 [47, 48, 49]。此外，一些研究 [50, 51, 52, 53, 54, 23, 55] 结合变分机制，采用了**变分自编码器（VAE）**[15]。与 GAN 和 VAE 不同，**标准化流（NFs）**[56] 明确学习数据分布以生成条件运动 [57, 58, 59, 60]。受到语言 [61] 和图像生成 [62] 领域成功应用的启发，**自回归模型** [63] 最近在运动生成领域被广泛探索，以进一步提高运动生成质量 [64, 65, 66, 67, 68, 54]。通过向量量化（VQ）[69] 将运动转化为离散标记序列，然后输入自回归模型以迭代生成多样化的运动标记序列。

扩散模型 [13, 70, 71] 在视觉生成任务中取得了显著成功 [72, 73, 74]，近年来已成为运动生成领域最流行的方法之一 [6, 7, 32, 8, 75, 76]。这些模型有效克服了 VAEs 中后验分布对齐的挑战，缓解了 GANs 对抗目标中固有的不稳定性，解决了基于能量模型的马尔可夫链蒙特卡洛方法中的计算负担 [77, 78]，并实施了类似于 NFs 的网络约束 [79]。与之前占主导地位的基于 VAE 的方案不同，基于扩散的运动生成模型通过随机扩散过程增强了生成能力，该过程建模了复杂分布，从而产生多样化和高保真的运动序列。此外，扩散模型在生成过程中保留了原始运动序列的形式，允许在去噪过程中更容易地施加约束，以确保物理合理性和真实性。尽管运动扩散模型取得了显著进展，但研究社区仍然非常活跃，涌现出针对各种特定应用场景（如可控性 [1, 80, 81, 36]、数据可用性 [82, 83, 35, 84]、泛化能力 [85] 和物理合理性 [76, 6, 32]）的新研究方向。

1.1 范围

本综述主要聚焦于基于扩散模型的运动生成方法，涵盖了该领域的代表性工作，主要围绕由文本描述驱动的条件生成。我们对这些方法进行了全面的回顾和归类，并讨论了基于扩散模型的运动生成所面临的几个关键挑战和潜在的未来发展方向。

先前的综述 [86] 提供了对人体运动生成的详尽和全面的回顾。然而，自 2022 年发表以来，该领域发展迅速，使得该综述所包含的文献不再充分代表当前的发展动向。例如，在 [86] 选出的 75 个代表性工作中，主流方法为 VAE（31/75）和 GAN（17/75），仅有 10 个基于扩散模型。然而，目前最先进的运动生成方法大多基于扩散模型和新近出现的自回归模型。因此，本综述作为对 [86] 的补充，总结了最新的代表性工作，并概述了基于扩散模型的运动生成领域中几个值得关注和具有价值的研究方向。我们旨在提供有价值的见解，并激励研究人员应对这一不断演变领域的挑战。

1.2 大纲

本综述的其余部分安排如下。在第2节中，我们将简要介绍扩散模型的基本原理和运动的表示方法。然后，我们将基于扩散的运动生成方法分为三种主要类型。在第3节中，我们讨论了构建有效运动扩散模型的工作，这些工作将扩散模型应用于运动领域。在第4节中，我们探索通过文本控制或空间约束增强来提高生成可控性的方法。在第5节中，我们回顾了在数据有限的场景中解决运动生成的方法。在 6节中，我们重点介绍了其他几个值得注意的研究方向。最后，在第7节中，我们得出结论并指出值得关注的未来研究方向。

2 预备知识

2.1 扩散模型

首先，我们将介绍扩散模型的基本原理，以提供必要的前置知识。扩散模型 [13] 由两个相互联系的过程组成：一个预定义的正向过程，它将数据分布映射到一个更简单的先验分布，通常是高斯分布；以及一个相应的逆向过程，该过程利用训练有素的神经网络逐步逆转正向过程的影响。在第2.1.1节中，我们从经典的去噪扩散概率模型（DDPM）[13] 讲起，并对问题进行形式化。接下来，在第2.1.2节中，我们解释了将 DDPM 从离散时间过程扩展到基于随机微分方程 [71] 的连续时间框架。最后，我们通过介绍两种最流行的引导方法的原理来讨论条件生成：分类器引导 [87] 和无分类器引导 [88]。如果希望进一步了解扩散模型的原理和最新进展的全面信息，我们建议参考综述文献 [79]，以及两篇内容丰富的教程 [89, 90]。

2.1.1 去噪扩散模型 (DDPM)

问题形式化。扩散模型可以被形式化为一个马尔可夫链 $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}$，其中 $\mathbf{x}_1, ..., \mathbf{x}_T$ 是从真实数据 $\mathbf{x}_0 \sim p_0(\mathbf{x}_0)$。经过添加噪声得到的序列。所有变量 x_t 具有相同的维度。

前向过程在 DDPM 框架中，为了近似后验分布 $q(\mathbf{x}_{1:t}|\mathbf{x}_0)$，正向过程遵循一个马尔可夫链，逐步向数据添加高斯噪声，直到其分布接近潜在分布 $\mathcal{N}(\mathbf{0}, \mathcal{I})$，这一过程是根据预定义方差系数 $\beta_1, ..., \beta_t$ 来进行的：

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

反向过程。逆向过程 $p_\theta(\mathbf{x}_{0:T})$ 也是一个马尔可夫链，它从 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ 开始，通过学习到的高斯转换来

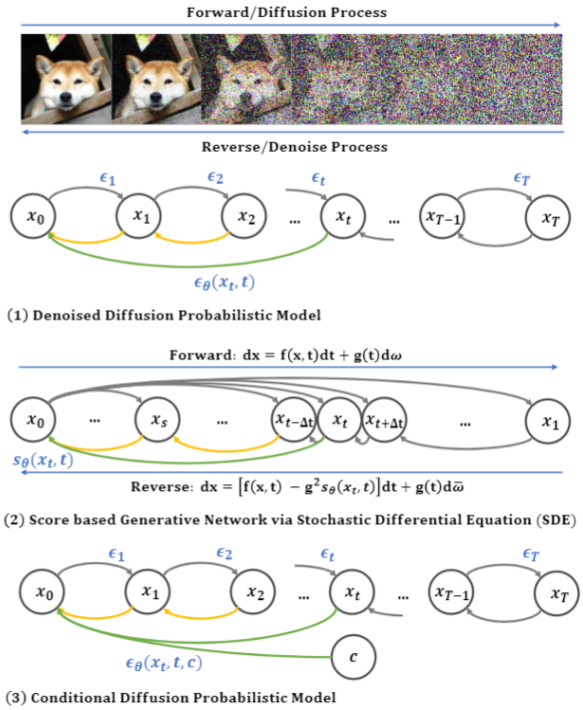

图 1: 扩散模型总览。[79]

预测并消除噪声：

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

上式中的 $\mu_\theta(\mathbf{x}_t, t)$ 表示由逆步分布 p_θ 确定的逆高斯核的可学习均值。在 DDPM 中，Σ_t 表示与预定义的超参数 β_t 对应的固定方差。

训练。为了通过联合概率分布 $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$ 逼近数据分布 p_0，DDPM 需要最大化 $\mathbf{x}_0$ 的对数似然。然而，直接最大化似然函数 $log\ p_\theta(\mathbf{x}_0)$ 是具有挑战性的。因此，DDPM 通过一系列推导，将优化目标重新表述为最大化证据下界（ELBO）。基于这个等价目标，DDPM 构建并优化了一个神经网络来模拟 $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$，并根据后验分布 $q(\mathbf{x}_{t-1}|\mathbf{x}_t, x_0)$ 获得训练目标。

重参数化提供了三种等价的实现方法来建模，即预测噪声 ϵ、原始样本 $\mathbf{x}_0$ 或得分函数 $\nabla log\ p(\mathbf{x})$（详见第 2.1.2 节）。以预测噪声 ϵ 为例，训练噪声预测模型 $\epsilon_\theta(\mathbf{x}_t, t)$ 的目标函数被简化为以下形式：

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

为了高效地从根据 $\mathbf{x}_0$ 获取 $\mathbf{x}_t$，[13] 将扩散过程表示为

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

其中 $\alpha_t = 1 - \beta_t$，$\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$。因此，我们可以简单地采样噪声 ϵ 和 t，通过这个公式直接生成 $\mathbf{x}_t$。

采样。为了生成多样的结果，我们对采样于 $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathcal{I})$ 的序列开始去噪。在采样过程中，我们可以一步步地采样 $\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_t)$，并在最终得到一个干净的生成结果 $\hat{\mathbf{x}}_0$。均值 $\mu_\theta(\mathbf{x}_t, t)$ 可以根据 $\mathbf{x}_t$ and $\epsilon_\theta(\mathbf{x}_t, t)$ 由以下等式得到：

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

最新进展。在 DDPM 之后，出现了许多代表性工作对其进行了改进。DDIM[70] 从理论上将 DDPM 的训练过程由马尔科夫扩散推广到非马尔科夫过程, 进而提出了一种加速采样方法。iDDPM[91] 引入了可学习的方差，以增强 DDPM 的对数似然。Analytic-DPM[92] 证明了在反向过程中的方差存在解析解，通过在采样过程中校正方差项，显著提高了采样速度和质量。LDM[93] 在潜在空间应用扩散过程，显著提高了训练和采样效率，并使高分辨率图像的合成成为可能。DiT[94] 引入了基于 Transformer 的 LDM 模型，具有更好的可扩展性。

2.1.2 Score-based Diffusion Models

DDPM 在前向和逆向阶段均作为离散过程运行。然而，通过引入随机微分方程（SDE），ScoreSDE[71] 将离散时间方案扩展到连续时间框架，允许在不同级别（T）上将 DDPM 视为离散近似。这使得我们可以利用常微分方程（ODE）/SDE 社区中的现有技术对扩散模型进行理论分析，同时在实践中应用适当的离散化。此外，[71] 引入了基于概率流常微分方程（ODE）的确定性采样器，这有助于通过黑盒 ODE 求解器实现快速自适应采样，通过潜在代码灵活处理数据，实现唯一可识别的编码，并且值得注意的是，它还可以进行精确的似然计算。

正向随机微分方程。 ScoreSDE 不再使用有限数量的噪声分布来扰动数据，而是考虑了随时间按照扩散过程演化的连续分布集合。这个过程逐步将数据点扩散成随机噪声，由一个既定的 SDE 描述，该 SDE 不依赖于数据，也没有可训练的参数。这可以被建模为伊藤型随机微分方程的解：

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

在正向随机微分方程（SDE）中，$\mathbf{w}$ 代表标准维纳过程（也称为布朗运动），$\mathbf{f}(\cdot, t)$ 是一个向量值函数，称为 $\mathbf{x}(t)$ 的漂移系数，而 $g(\cdot)$ 是一个标量函数，称为 $\mathbf{x}(t)$ 的扩散系数。只要系数在状态和时间上都是全局 Lipschitz 连续的，SDE 就具有唯一的强解。

逆向随机微分方程。通过反转前向过程，ScoreSDE 能够将随机噪声 $\mathbf{x}_T \sim p(x_T)$ 平滑地生成干净的样本。而这个逆向过程满足一个逆时 SDE，该逆时 SDE 可以从前向 SDE 推导出来。

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

在逆向随机微分方程中，当时间从 T 逆向流动至 0 时，$\bar{\mathbf{w}}$ 代表一个标准维纳过程，而 $\mathrm{d}t$ 表示一个无限小的负时间步长。一旦对于所有时间 t 已知每个边际分布的得分（即梯度），$\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$，我们就能从 eq. (7) 推导出逆向扩散过程，并模拟它以从先验分布 p_0 中采样。这里 p_t 表示边际分布，而 p_0 表示先验分布。

训练目标。为了近似逆时 SDE，训练一个基于时间的分数模型 s 来估计分数 $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$。然而，由于 $\nabla_{\mathbf{x}_t} \log p_t(x)$ 很难获得真实值,ScoreSDE 采用分数匹配目标,使得 $\mathbf{s}_\theta(\mathbf{x}_t, t)$ 预测任意 $\mathbf{x}_t$ 对应的 $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)$，再对 $\mathbf{x}_t$ 求期望:

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

此处，$\lambda(t)$ 是一个正的权重函数，t 在区间 $[0, T]$ 上均匀采样，$q(\mathbf{x}_t|\mathbf{x}_0)$ 是与前向过程 eq. (6) 相关联的高斯转移核。如果 $\mathbf{f}(\mathbf{x}_t)$ 是一个仿射变换，那么 $p(\mathbf{x}t|\mathbf{x}0)$ 也是一个易求解的高斯分布。在拥有足够数据和模型容量的情况下，得分匹配确保了 eq. (8) 的最优解，记作 $\mathbf{s}\theta(\mathbf{x}, t)$，对于几乎所有的 $\mathbf{x}$ 和 t 等于 $\nabla\mathbf{x}_t \log p_t(\mathbf{x}t)$。获得基于得分的模型 $\mathbf{s}\theta(\mathbf{x}_t, t)$ 后，ScoreSDE 可以使用数值 SDE 求解器生成多样化的样本。

概率六常微分方程文献 [71] 提出了一种新的数值方法——概率流常微分方程（ODE），用于求解逆时随机

微分方程（SDE）。对于所有的扩散过程，存在一个相应的确定性过程，它与 eq. (7) 中的 SDE 共享相同的边际概率密度集合 $p_t(\mathbf{x})_{t=0}^{T}$，并且满足一个常微分方程（ODE）：

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

与随机微分方程（SDEs）将 DDPM 扩展到连续时间的思想类似，常微分方程（ODEs）代表了 DDIM 的连续空间解释。当方差被设为零时，SDEs 将退化为 ODEs。与 SDEs 不同，由于缺乏随机性，概率流常微分方程可以采用更大的步长进行求解。因此，一些研究 [95, 96, 97, 98] 通过使用高级的 ODE 求解器实现了更快的采样速度。

2.1.3 条件生成

在前面的部分中，我们专注于对原始数据分布 $p(\mathbf{x})$ 进行建模。然而，我们在实践中往往对建模条件分布更感兴趣，尤其是在动作生成领域。这将使我们能够通过输入条件信号 c，显式控制我们生成的样本。条件生成主要分为分类器引导（Classifier-Guidance[87]）和无分类器引导（Classifier-Free guidance[88]）两种方法，前者只需在预训练好的无条件扩散模型的基础上训练一个分类器，仅在采样时进行梯度引导，大大降低了成本; 而后者加入了条件引导从零开始训练，可以实现更细粒度的控制。

分类器引导。以基于得分的公式化作为示例（如第 2.1.2 节所述），我们可以通过贝叶斯规则推导出以下等价形式 [89]，用于学习 $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c)$。

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

在上面的推导中，我们利用了 $\log p(c)$ 相对于 $\mathbf{x}_t$ 的梯度为零这一事实。

由 eq. (10) 推导出的结果可以解释为，通过将一个无条件得分函数与分类器 $p(\mathbf{x}|c)$ 的对抗梯度相结合，可以建模条件得分函数。因此，在分类器引导方法中，首先基于预训练的无条件扩散模型训练一个分类器，可以接受任意噪声水平的 $\mathbf{x}_t$ 并尝试预测条件信号 y。然后，在采样过程中，总体条件得分函数被计算为无条件得分函数与噪声分类器的对抗梯度的加权和。为了引入细粒度控制来鼓励或阻止模型考虑条件信息，分类器指导通过 γ 实现梯度缩放。在分类器指导下学习的得分函数可以总结为：

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

对于 DDPM，基于上述推导，只需将采样过程修改为：

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

无分类器引导。与分类器引导相比，无分类器引导直接在训练时定义 $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, c, t), \Sigma_t)$，从而将训练目标 eq. (3) 转化为：

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

特别地，无分类器引导方案也模仿分类器引导方案加入了 γ 参数的梯度缩放机制来平衡相关性与多样性。类比 eq. (12)，我们也可以在无分类器引导方案中引入 $\omega = \gamma - 1$，将采样过程修改为：

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

为了得到无条件的输出 $\epsilon_\theta(\mathbf{x}_t, t)$，我们可以新引入一个特定的条件输入 ϕ，对应全体图片，并在训练时按一定概率出现，便可得到 $\epsilon_\theta(\mathbf{x}_t, t) = \epsilon_\theta(\mathbf{x}_t, \phi, t)$。

2.2 运动数据表示

运动数据通常通过运动捕捉 [99, 100, 101, 17]、伪标签 [102, 103, 104, 105] 以及手动标注 [106, 107, 108, 109] 等方法收集等方法进行收集。运动序列通常表示为 $m_{1:N} = m_i{}_{i=1}^{N}$，其中 $m_i \in \mathbb{R}^D$ 表示第 i 帧的姿态状态，N 是总帧数。在这一部分，我们将简要介绍运动数据的表示方法，分别是基于关键点和基于旋转的表示方法 [86]。

基于关键点的表示将身体上的部分关节或其他重要位置视为关键点，并通过这些关键点表示全身姿态。其中，每个关键点由其在像素或世界坐标系中的 2D/3D 坐标表示。**基于旋转的表示**利用关节角度来表示人体姿态，根据身体的分层结构，记录各关节相对于其父级的旋转，并被参数化为欧拉角和四元数等格式。而且，可以使用统计网格模型，如 SMPL[110]，对基于旋转表示的人体进行建模，进一步捕捉身体的形状和运动过程中发生的变形。

这两种表征各有优势。前者可以直接从动捕设备获得，因此更加易得并有更好的可解释性；而后者可以直接利用统计网格模型进行人体建模，更容易被应用于动画和机器人等领域。此外，通过正向运动学（FK）和逆向运动学（IK），我们可以进行这两种表示形式之间的转换。

3 运动扩散模型

由于扩散模型已经在许多生成领域 [72, 111, 112, 113] 取得了令人惊艳的效果，几篇早期工作 [8, 6, 114] 探索了扩散模型在运动生成中的应用。**运动扩散模型（MDM）**[6] 引入了一种无需分类器的基于扩散的生成模型，专门设计用于人类运动。MDM 采用 DDPM 框架，遵循 [73] 的方法，在去噪过程中直接预测原始运动序列。这种方法使得在生成过程中可以直接应用在运动领域验证过的几何损失，分别约束了位置、脚部接触和速度，可以增强物理特性并防止伪影，促进自然和连贯的运动。如图2所示，MDM 采用了基于 Transformer[115] 架构的去噪模型。它首先使用 CLIP 的文本编码器从文本提示中提取特征，然后将这些特征与时间戳编码结合，形成输入令牌之一 z_{tk}。该令牌随后被添加到位置嵌入中，并与其他带噪声的运动令牌一起输入到 Transformer 编码器中。此外，论文还提出了一种基于 inpainting[116, 117] 的运动编辑方法，其中待编辑的部分被加噪声，而其他部分保持不变，然后通过迭代去噪获得修改后的运动序列。由于缺乏相关数据，MDM 在生成长多段运动、多人运动和精细控制（如轨迹和末端执行器跟踪）方面面临挑战。为解决这些问题，**PriorMDM**[118] 引入了三种基于扩散先验（预训练的 MDM）的组合，旨在通过少量微调或采样期间的零样本校正，扩展 MDM 的应用场景。为生成任意长度的运动，PriorMDM 提出了 DoubleTake 方法，该方法可以将 MDM 生成的短运动（10 秒）连接成长序列，并在段之间添加过渡，这使得每个运动段可以由不同的文本提示控制，并在序列长度上有所变化。为生成两人动作，论文提出了一种小样本微调方法，并设计了 ComMDM 模块（单层 transformer），以协调两个预训练 MDM 实例之间的交互。为实现精细调节的运动控制，论文采用了 inpainting 方法来微调特定的控制条件，并使用模型组合来结合不同的单一控制条件。

MDM 的同期工作 **MotionDiffuse**[8] 基于 DDPM 架构设计了预测噪声的 Transformer-based 模型，将文本特征经交叉注意力融入生成过程。为了提高效率，MotionDiffuse 在自注意力层和交叉注意力层均使用线性注意力降低计算复杂度。为了融入扩散的时间条件，MotionDiffuse 将时间 t 编码后与文本特征相加，再输入 Stylization Block。值得一提的是，它还实现了身体部位独立控制（Body Part-independent

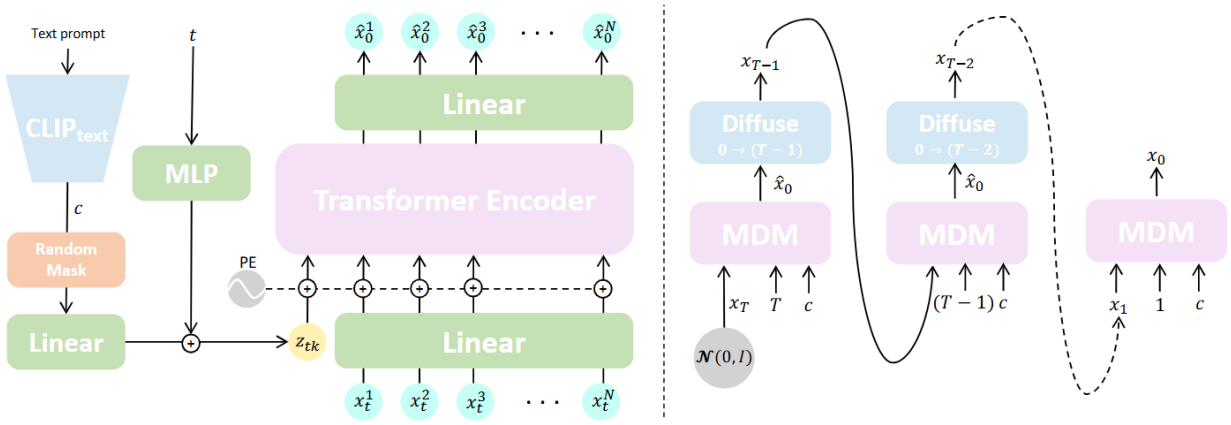

图 2: MDM 总览。[6]

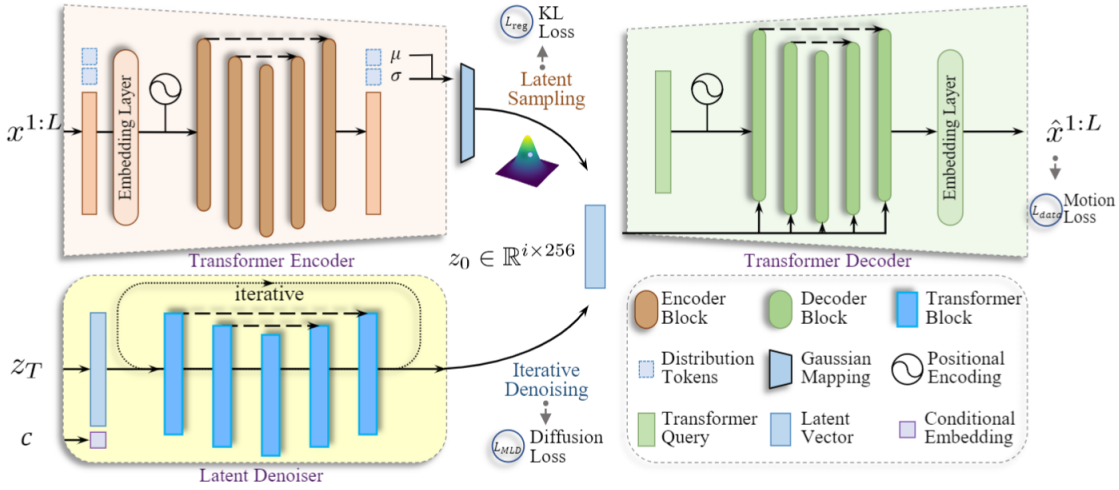

图 3: MLD 的总览.[7]

Controlling）和可变时间控制（Time-varied Controlling）两种细粒度可控生成方法。**FLAME**[114] 也是最早应用扩散模型到动作生成的工作之一，它根据 iDDPM[91] 应用了可学习方差，并且额外设置了 ML token 指明动作序列的长度。

为了提高效率并降低计算开销，**Motion Latent Diffusion (MLD)**[7] 借鉴 LDM 的思想提出潜在运动扩散模型，在降维后的潜在空间进行扩散，从而大大提高了效率。如图3所示，MLD 首先训练一个运动变分自编码器（VAE）。该 VAE 使用基于 transformer 的编码器处理原始运动令牌和分布令牌，预测 μ 和 σ，并使用 Kullback-Leibler 散度损失进行训练。在解码阶段，L 个零运动令牌作为查询，而潜在表示在解码器的交叉注意力层中充当键和值，生成的运动序列将计算 MSE 损失监督。一旦运动 VAE 训练完成，便得到一个高效、低维的潜在空间，可以有效地滤除高频和难以察觉的细节。随后，MLD 在潜在空间中训练了一个基于 transformers 的扩散模型。与在连续特征空间中进行扩散的 MDM, LDM 不同，M2DM[75] 探索了一种在离散空间进行扩散的方法。它先训了一个 VQ-VAE[69]，关于 motion 序列编码了一个信息丰富的 codebook，再对得到的离散潜在表示做扩散，本文还专门为离散扩散设计了特定的加噪和去噪处理。

由于 DDPM 生成的动作不能保证在物理和解剖学上是合理的，会出现运动抖动、非法骨骼和脚滑动等伪影。为了增强运动合理性，**Mofusion**[32] 引入运动学 loss 来约束骨骼运动，确保合成运动中的骨骼

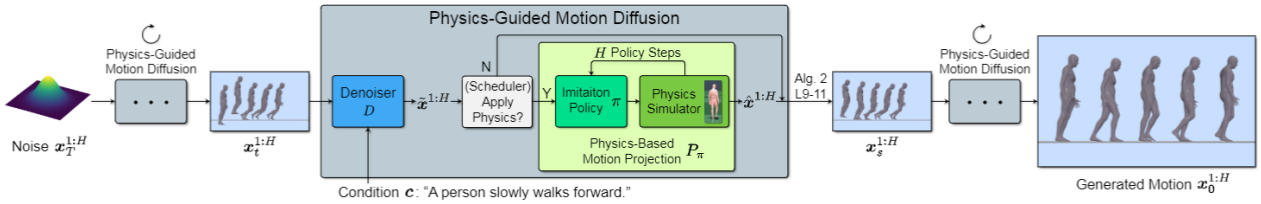

图 4: Physdiff 的总览。[76]

长度随时间保持一致，并且骨骼长度左/右对称。通过重参化，Mofusion 将预测的噪声转换到原始运动形式，从而可以更容易、更直观地应用运动学约束。如图4所示，**Physdiff**[76] 提出了一个简单有效的基于物理规律的投影策略，可以在采样时提高预训练文生动作模型的物理合理性。具体来说，Physdiff 利用了强化学习方法，训练了个运动模拟器，智能体可以在模拟器中根据当前状态和预测的下一帧运动（去噪模型的输出），生成物理合理的下一步运动。通过这样的马尔科夫形式，智能体可以迭代地生成物理合理的运动序列。基于物理规律的投影训练好后，在采样的每一个时间步结束时，Physdiff 利用该投影策略将生成的运动序列修正为物理合理的。还有一些有趣的工作进一步探索动作扩散模型的潜力。**InterGen**[119] 关注具有多人互动的运动生成，提出了一个双人交互的多模态数据集 InterHuman。为了实现双人运动生成，InterGen 引入两个协同 Transformer 用来去噪，他们共享权重，分别处理每个个体，并且借助新颖的相互注意力机制连接不同级别的特征。并且，InterGen 提出了一种有效的运动表示和两个附加正则化损失，以建模人与人交互下的复杂空间关系。**Ude**[29] 提出了一种模态无关的的 Transformer 编码器实现多模态运动一致性，即文本和音频这两种条件的统一。并且，Ude 提出一种结合 VQ-VAE 和扩散模型的生成方法，通过 VQ-VAE 编码离散代码标记以保证长序列的质量和语义一致性，通过扩散模型作解码器以增强多样性。在得到条件标记后，将其作为查询输入 Transformer 以自回归地生成运动的潜在表征序列，经过扩散模型解码器解码即可得到生成的运动序列。加速**自回归运动扩散模型（AAMDM）**[120] 也结合了自回归模型和扩散模型，将去噪扩散 GAN 集成为快速生成模块，并将自回归扩散模型集成为抛光模块，大大提高了生成效率。

4 可控性增强

尽管条件动作生成，包括文本驱动和音频驱动等，取得了飞快的进展，可以生成多样化且逼真的运动序列。但是，由于文本缺乏足够精细的描述和足够的信息密度，使用者难以对合成的运动序列的特定时间的特定姿势细节进行控制。并且，目前的方法难以实现精确控制，而空间控制信号，如轨迹、关键帧等条件，对可控动作生成至关重要。因此，在这一部分，我们将分别讨论基于扩散模型的动作生成中关于增强文本控制和增强空间约束的可控性增强工作。

4.1 文本控制增强

如图5所示，**FineMoGen**[121] 希望可以实现细粒度时空运动生成。为此，它首先设计了可以提供细粒度时空细节的文本描述，从而通过丰富文本输入以增强控制，实现对任一阶段/任一身体部位的控制。同时，FineMoGen 设计了新的注意力模块————时空注意力，显式地建模时间-空间约束。而由于需要关注更细粒度的区域，模型学习的复杂度大大提高，FineMoGen 应用了稀疏激活的 MoE[122] 来增强模型的表征学习能力。并且，由于过去的数据集缺乏对身体部位和时间阶段的细粒度描述，本文扩展 HuMMan[18] 数据集提出了一个新的数据集，将每个序列分成多个时间阶段，并为每个阶段分别提供总体文本描述和针对 7 个身体部位分别的细粒度描述。而为了扩展应用场景，FineMoGen 还应用了大语言模型，可以让用

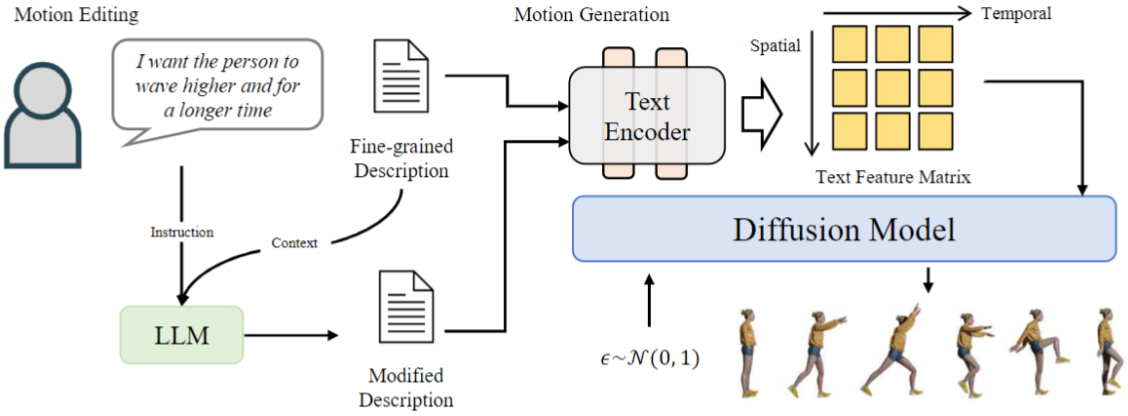

图 5: FineMoGen 的总览。[121]

户用自然语言实现细粒度控制和编辑。

与 FineMoGen 不同，还有一些工作仅针对时间或身体部位其中一方面进行了文本控制增强 [123, 124, 36] 。为了实现长复杂文本控制和可变长度 motion 生成，**自回归动作模型 (AMD)**[123] 将自回归模型与扩散模型结合。为了处理长复杂文本控制，本文提出了一个新的文本-动作数据集，其中文本描述是由短句构成的，每个短句对应一个单运动段。AMD 自回归地调用扩散模型预测每一个运动段，从而迭代预测当前运动段，最终得到整体运动序列。同期的 **FlowMDM**[124] 也是针对长文本动作生成任务，提出 Human Motion Composition 任务，为了生成长的、连续的运动序列，而该长序列是由分别描述每个时间片段的长文本引导。为了既要完全控制目标运动的顺序及其持续时间，又要在动作之间实现无缝且真实的过渡，本文引入了基于扩散模型的 FlowMDM，无需任何后处理或冗余的去噪步骤。为了快速采样，本文引入了混合位置编码，包含绝对和相对位置编码，去噪时首先利用绝对信息来恢复全局运动相干性，然后利用相对位置在动作之间建立平滑且真实的过渡。并且引入了一种专为 Human Motion Composition 任务设计的混合位置编码，确保每个姿势根据其自身条件及其相邻姿势进行去噪。除了针对长文本，**LGTM**[36] 采用大语言模型将全局运动描述分解为特定于部分的叙述，通过丰富文本描述来增强对身体部分的控制能力。丰富后的描述由独立的身体部分运动编码器进行处理，以确保精确的局部语义对齐。最后，使用基于注意力的全身优化器细化运动生成结果并保证整体连贯性。

之前介绍的几篇工作均通过显式地增强文本指令以增强文本控制，而 **Fg-t2m**[125] 通过详细解析文本指令来隐式地增强文本控制。过去的文本生成动作方法大多直接用预训练好的文本编码器提取文本特征，但这难以让模型准确理解文本，而 Fg-t2m 首次将自然语言处理领域的方法引入文本生成动作任务中，提出了一种新的解析文本特征和融合的方法以支持精确的文字描述。具体来说，它设计了语言学-结构辅助模块利用依存分析树和图网络提取文本特征, 从而可以在聚合文本语义的同时保留文本语言学结构。还提出了上下文感知渐进推理模块，模仿人类实现由全局（语义）到局部（单词）的多步推理策略。

4.2 空间约束增强

由于仅通过文本描述难以精确指导模型生成符合人们需求的运动，因此，许多研究通过引入空间约束来提高生成运动的可控性 [42, 43, 80, 126]。**DiffKFC**[42] 是一种在给定稀疏关键帧控制条件下进行文本到运动生成的方法。传统的插值方法在稀疏关键帧条件下表现不佳，因为过于稀疏的关键帧在去噪过程中可能被错误地识别为噪声，从而无法提供有效的生成条件。而 DiffKFC 通过将关键帧意识融入训练过程，实现了关键帧的协同工作。模型结构如图6所示：首先将运动序列添加噪声并编码成标记，然后与时间戳和文本标记一起输入到 Transformer 解码器。同时，稀疏关键帧在扩展至原始序列长度后被编码，输

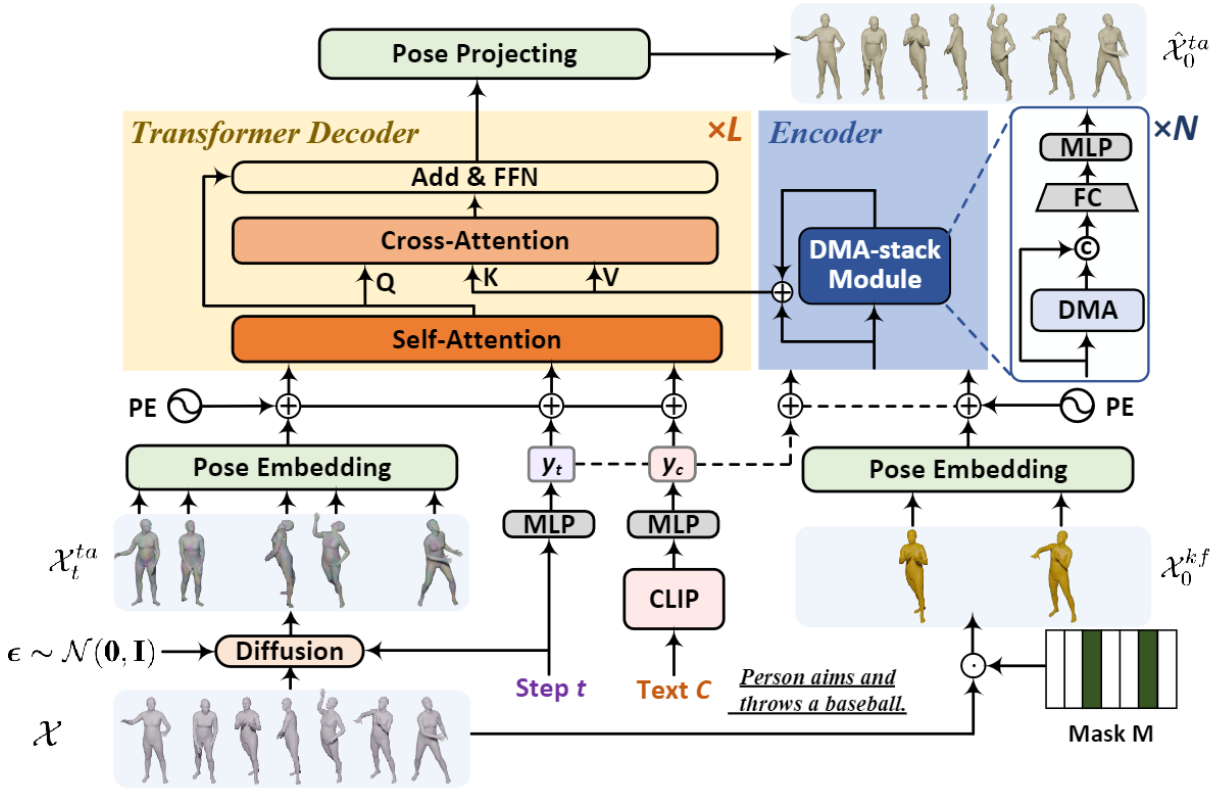

图 6: DiffKFC 的流程图。[42]

入关键帧编码器以提取特征。再以关键帧特征作为条件，在 Transformer 解码器的交叉注意力层与噪声化的运动序列进行融合。为了充分利用关键帧特征，本文提出了一种扩张掩码注意力模块（Dilated Mask Attention, DMA），它通过逐步扩张稀疏关键帧特征，实现在迭代过程中的稠密特征聚合。**条件运动插值（Conditional Motion Diffusion In-betweening, CondMDI）**[43] 针对由关键帧引导的人体运动生成任务，支持任意密集或稀疏关键帧的布局，以及部分关键帧的约束，同时生成多样化且与给定关键帧一致的高质量运动。**引导扩散模型（Guided Motion Diffusion, GMD）**[80] 针对具有空间约束的动作生成，如运动轨迹和障碍物，提出了一种有效的特征投影方案，以增强空间信息和局部姿势之间的一致性。对于稀疏条件，如稀疏关键帧，本文还提出了一种密集引导方法，将其转换为密集信号。

除了轨迹和关键帧，还有研究工作针对更广泛的空间约束进行了探索 [1, 81, 127]。**AGRoL**[1] 专注于 AR/VR 头戴式设备使用场景，针对给定稀疏的上半身信号，预测全身姿势，并设计了一种基于多层感知器（MLP）的扩散模型。**OmniControl**[127] 能够将灵活的空间控制信号整合到不同时间点的不同关节上，例如对关键帧中特定关节的全局位置进行控制，并引入空间和现实感指导，通过精心设计目标函数来计算梯度，引导生成过程。更进一步，**可编程运动生成（Programmable Motion Generation）**[81] 将任意复杂的控制信号分解为基本约束的组合，如轨迹、关键帧、交互。它通过测量误差来评估所有基本约束，并通过累加误差实现开放且完全可定制的运动控制集合。在获得误差函数后，它利用预先训练的运动扩散模型（MDM）并优化其潜在代码，以最小化该误差函数。

5 数据可用性

由于文本-动作数据集难以采集，而大部分运动生成模型都是在在精心标注的数据集上训练的。为了缓解对数据的依赖，将动作生成适应数据不充足的情形，出现了一系列工作提高数据可用性。

提高数据可用性的一种直接方法是扩展数据集。**Make-An-Animation (MAA)**[82] 通过使用文本

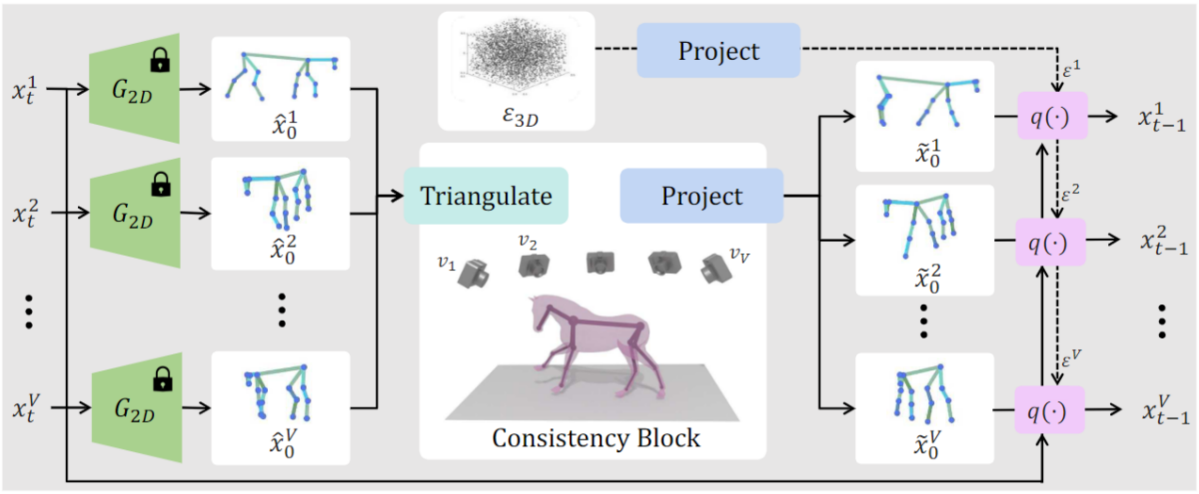

图 7: MAA 总览。[82]

到图像的数据集进行预训练，以扩展文本到运动（T2M）模型生成结果的领域覆盖范围。MAA 首先从文本到图像数据集中筛选出与人体姿态相关的图像，然后使用预训练模型提取人体的关键点，生成 3D 姿态特征，并构建了一个文本伪姿态（Text Pseudo-Pose, TPP）数据集。整个训练过程分为两个阶段，如图7所示：第一阶段，在 TPP 数据集上训练一个文本到姿态的扩散模型；第二阶段，在文本到运动数据集上进行微调，向 ResNet 块和注意力块添加一维时间卷积和时间注意力层作为适配器，并使用没有字幕的运动数据进行无分类器的训练。与 MAA 类似，**Multi-view Ancestral Sampling (MAS)**[84] 利用野外视频数据，提取 2D 姿态进行训练。在训练过程中，同时对表示同一 3D 运动的不同视角的多个 2D 运动序列进行去噪，并在每一步将各视角的 2D 运动序列根据其姿态约束统一成 3D 序列，然后将其投影回原始视角，以确保每个扩散步骤中所有视角的一致性。**MotionMix**[83] 提出了一种简单有效的弱监督扩散模型，用于利用噪声较大（低保真度）和未标注的数据进行训练。通过双阶段训练，模型能够先根据条件生成初始的粗略运动，然后将这些粗略运动细化为高质量的运动，从而有效地从噪声有标注和干净无标注运动序列中学习。

除了扩充数据集，还有一系列创新性的研究工作 [35, 128, 129] 致力于解决数据不足的问题。**开放式词汇运动生成（Open-vocabulary Motion Generation, OMG）**[35] 充分利用了预训练后微调的范式，针对零样本（zero-shot）开放式词汇文本提示下的动作生成。OMG 扩展了 DiT 模型，并在大规模无标注运动数据集上进行预训练，学习到了丰富的跨领域固有运动特征。在微调阶段，OMG 引入了运动控制网络（motion ControlNet[130]），并加入了新提出的混合控制器（Mixture-of-Controllers, MoC）模块。该模块基于文本和运动之间的交叉注意力机制，能够自适应地识别子运动，并采用不同的专家进行处理，有效提高了文本与运动之间的对齐度。为了克服领域依赖性，**检索增强运动扩散模型（Retrieval-Augmented Motion Diffusion Model, ReMoDiffuse）**[128] 采用了一种检索增强方法来生成更多样化的运动。该方法能够高效且有效地从检索样本中探索知识。在检索过程中，首先使用 CLIP 计算文本相似度，然后计算与运动长度相关的系数，结合这两个因素，可以得到与样本最相似的 k 个检索样本。这些检索样本的特征将作为条件特征，在生成过程中提供指导。此外，还有一篇非基于扩散模型的文章 [129]，探索了偏好学习（preference learning）在文本到运动（T2M）中的应用，即不依赖于动作捕捉标注的运动数据，而是利用偏好对（preference pairs）进行学习。由于数据采集更为简单，偏好学习非常有效，可以利用成本效益高的标签训练 T2M 模型，而无需专业标注者。

近期，动作生成领域开始关注更具泛化性的生成对象，即针对不同种类对象的动作生成问题。由于不同主体，如人类、狗、龙等，其骨骼和几何结构等姿态属性存在显著差异，导致基于人类数据训练的动作生

成模型难以直接应用于其他主体，这在很大程度上限制了动作生成模型的应用范围。特别是对于非人类骨骼的动物或虚构生物（如龙），其运动数据极为稀缺。因此， **SinMDM**[85] 提出了一种一次性（one-shot）方法，能够对任意骨骼拓扑结构的动作进行建模，并合成与这些拓扑结构相符的任意长度的动作，即便这些拓扑结构通常只能通过单一的动画序列进行学习。SinMDM 通过结合局部注意力机制和浅层 UNet[131] 进行去噪，以学习局部运动序列，同时缩小感受野以促进生成动作的多样性并减少过度拟合的风险。此外，一项基于变分自编码器的研究——**OmniMotionGPT**[132]，发布了一个专注于动物动作生成的数据集，并提出了一种生成数据受限主体的新方法，该方法利用人类数据学习先验知识，并实现知识迁移，以便在有限的动物数据基础上生成多样化的动物动作。

6 其他值得注意的方向

前面几部分总结的工作主要集中在文本驱动的动作生成领域，然而，基于音频驱动的舞蹈生成在过去几年也取得了显著进展，并涌现了一些采用扩散模型的研究。早期的研究工作 [31] 专注于音频驱动的舞蹈生成以及语音驱动的手势生成，该模型架构是在音频生成模型 DiffWave[111] 的基础上改进而来，同时，研究者还尝试了混合专家模型（MoE）方法。**EDGE**[28] 将运动扩散模型（MDM）应用于舞蹈生成，并采用了音乐预训练模型 Jukebox[133] 来提取音乐特征。**FineDance**[30] 则提出了一个包含多个舞蹈流派的大型 3D 舞蹈生成（编舞）数据集，并开发了一个用于舞蹈生成的模型。该模型的主体结构是在 MDM 的基础上进行修改，特别针对手部和肢体的细粒度生成进行了优化（分别使用专家网络生成这两部分后再进行融合），并引入了流派和连贯性匹配模块（Genre&Coherent aware Retrieval Module），以提高生成舞蹈的流派一致性和动作连贯性。

同样地，场景驱动的运动生成也出现了一些基于扩散模型的方法。Wang 等人的研究 [38] 专注于 3D 场景中的语言引导人体运动生成，使用 RGB 点云表示的床和墙等特征作为输入条件，提供了一个很好解决语言-场景-运动对数据受限下生成的方法。由于场景感知的人体运动生成在泛化到训练分布之外的新对象交互方面存在困难，**ROAM**[37] 通过仅使用一个参考对象训练运动模型，确保了对 3D 对象感知角色合成中新场景对象的鲁棒性和泛化能力。Mir 等人的方法 [39] 在给定一组稀疏关节位置和 3D 场景中的种子运动序列的情况下，能够生成适应不同场景的连续运动。

近期，动作生成领域中有一些新的生成模型崭露头角，它们在特定方面展现出了与传统扩散模型有竞争力的性能。自回归运动模型作为一种新兴技术，已被开发用于提升运动生成的质量，例如 MotionGPT[66], HumanTOMATO[68], and Motiongpt[67]。这些模型首先训练一个向量量化变分自编码器（VQ-VAE），将运动序列编码为离散的代码，这些代码作为标记与条件信息一同输入到 Transformer 模型中，再以自回归方式生成动作。然而，由于自回归模型的单向解码特性可能会限制其表达能力，一些基于生成掩码 Transformer 的研究应运而生，例如 MoMask[33] and MMM[34]。这些模型采用类似于 BERT[134] 的自编码训练方法，通过随机掩码一些标记，并与文本描述一同输入到 Transformer 的编码器中，以预测完整的标记序列。在采样时，输入一个完全掩码的标记序列，可以迭代地生成完整的动作序列。

7 结论与未来工作

扩散模型在动作生成领域的重要性日益凸显。为了充分利用扩散模型的潜力，本文提供了一个全面且最新的综述，涵盖了基于扩散模型的动作生成方法的代表性工作。本综述从扩散模型的基本原理出发，重点探讨了其发展轨迹和条件生成的方法。进一步地，我们将相关工作根据科研贡献进行了分类，并深入讨论了每一类别的任务。我们期望本综述能为读者提供一份全面的指导手册，清晰地展示基于扩散模型的动作生成研究的进展，并指出未来研究的潜在方向。

尽管该领域已取得显著进展，但未来仍面临重大挑战，为未来的研究提供了探索空间。因此，我们从多个视角提出了若干有希望的未来研究方向，旨在激发运动生成领域的新突破。

7.1 更先进的基础生成模型

尽管运动扩散模型已经得到了广泛的研究，但开发更先进的基础生成模型对于动作生成领域仍然至关重要。正如近期涌现的自回归模型和掩码生成模型，未来的研究可以探索如何整合这些先进生成模型的优势。同时，我们也应该从更宏观的视角来审视问题。虽然迭代去噪模型目前代表了最前沿的技术，但这种方法可能并非生成问题的终极答案。毕竟，人类的图像生成过程并非从纯粹的噪声中开始。因此，未来的研究应当持续探索更先进的基础生成模型，旨在提升对数据分布的建模精度和生成效率。

7.2 可控生成

可控运动生成在过去几年得到了广泛探索。但是，目前模型对于输入条件的可控性仍有待进一步提高。另一方面，目前的控制方法（长文本，轨迹，关键帧等）对于大众使用者仍不友好。如何进一步增强模型的可控性，支持更加灵活、交互性更强的动作可控生成，值得未来进一步思考。

7.3 数据可用性

正如在 section 5部分所讨论的，扩散模型在从低质量数据中识别模式和规律方面存在挑战，这限制了它们在新场景或数据集的泛化能力。在实际的动作生成应用中，文本到运动数据往往是有限的，这严重影响了运动生成方法的实际应用。SinMDM 等方法探索了将从丰富的人类骨骼运动数据中学习到的知识迁移到其他物种的运动生成，这一方向在未来将具有显著的价值。我们认为，数据可用性将是数据驱动的运动扩散模型长期面临的一个主要挑战。

7.4 效率提升

尽管扩散模型可以生成高质量的多样化结果，但是它的效率通常较低。由于其迭代去噪的采样策略，运动生成模型较慢的生成速度限制了其应用，未来需要进一步探索提高训练和采样效率的方法。

7.5 评价指标

尽管动作生成领域目前具有大量的评价指标，但是他们都有固有的局限性，提出符合人类标准的的运动生成评估指标仍极具有挑战性。未来的工作可以集中于设计更有原则性的客观评估指标，不仅要与人类的感知紧密结合，而且保持可解释性。

参考文献

[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] 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.

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

[97] 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.

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

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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

[116] 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.

[117] 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.

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

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

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

[125] 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.

[126] 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.

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

[128] 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.

[129] 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.

[130] 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.

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

[132] 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.

[133] 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.

[134] 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.