张博, 周雷. 姿势引导的人类运动视频生成综述[J]. 智能计算机与应用,2026,16(8):226-235. DOI:10.20169/j.issn.2095-2163.25020802

姿势引导的人类运动视频生成综述

张 博, 周 雷
(上海理工大学 健康科学与工程学院, 上海 200093)

摘 要: 随着深度学习和计算机视觉的快速发展,人工智能生成内容(AI-Generated Content,AIGC)在计算机视觉领域取得了显著成果。生成模型的最新进展为视频生成奠定了坚实基础。基于生成模型,给定输入人体图像与描述运动序列的参考视频,其目标是合成一个视频,使输入图像中的人物复制参考视频中的动作,从而输出一个可控且时间连贯的人类运动视频。尽管最近在姿势引导的人类运动视频生成技术方面取得了显著进展,但生成高质量、连贯且与条件信号一致的人类视频仍面临诸多挑战。本综述梳理了人类运动视频生成领域最新技术进展,旨在为研究社区提供深入见解,并激发新的思考,以推动人类运动视频生成技术的应用和发展。

关键词: 人工智能;深度学习;扩散模型;生成模型;数字人类;文献综述;视频生成;姿势引导
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0226-10

A survey on posture-guided human motion video generation

ZHANG Bo, ZHOU Lei
(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: With the rapid development of deep learning and computer vision, AIGC (AI-Generated Content) has achieved remarkable results in this field. The latest developments in generative models have laid a solid foundation for video generation. Based on generative models, given an input human body image and a reference video depicting the motion sequences, the goal is to synthesize a video in which the actions of the human figure in the image are replicated from the reference video, thereby outputting a controllable and time-coherent human motion video. Although significant progress has been made in posture-guided human motion video generation technology recently, it remains challenging to produce high-quality, coherent human videos consistent with conditional signals. This review summarizes the latest technological advancements in the field of human motion video generation, aiming to provide in-depth insights for the research community and stimulate new thinking to promote the application and development of this technology.

Key words: artificial intelligence; deep learning; diffusion models; generative models; digital humans; literature review; video generation; posture guidance

0 引 言

随着深度学习与计算机视觉技术的快速发展,AIGC 在近年来取得了显著的进步,并在多个领域产生了深远影响。其中,数字人视频生成技术作为 AIGC 的一个重要分支,正在逐渐成为虚拟现实、娱乐、教育和医疗等领域的重要工具。

数字人视频生成的核心目标是基于生成模型,以给定的输入人体图像与描述运动序列的参考视频为条件,合成一个视频,其中图像中的人物 复制参考视频中的动作,输出兼具可控性且时序连贯的人类运动视频。这一技术要求生成的视频与输入序列的动作高度一致。尽管在这一领域已经取得了一些显著的成果,但生成高质量的数字人视频仍然面临诸多挑战,特别是在确保生成视频的质量和连贯性方面。

生成模型的最新进展为人类视频生成带来了重要突破,其中,扩散模型[1]发挥了至关重要的作用。

基金项目: 国家自然科学基金(61906121)。
作者简介: 张 博(1998—),男,硕士。主要研究方向:基于深扩散模型的图像和视频生成算法研究及应用。
通信作者: 周 雷(1986—),男,博士,副教授。主要研究方向:基于深度学习的图像和视频处理算法研究及应用。
收稿日期: 2025-02-08

该模型凭借其强大的生成能力，在图像和视频生成领域已基本取代 GAN[2] 方法成为主流方案，为数字人视频生成奠定了坚实的基础。该模型在图像与视频生成领域的技术突破，使得研究人员能够生成更具真实感的数字人形象，并推动了该领域的发展。然而，由于人类运动的复杂性及其与条件信号之间的隐含关系，使得高质量的数字人视频生成仍面临很大的挑战性。

本文旨在对数字人视频生成技术进行全面综述，重点聚焦姿势引导的人类运动视频生成技术的最新进展，梳理前沿技术与方法，解析算法架构，探讨关键技术、现存挑战以及未来的发展方向。期望通过本综述，为研究社区提供深入的见解，并激发新的研究思路，以推动数字人视频生成技术在各个应用领域中的进一步发展。

如图 1 所示，该领域的技术发展可以从 3 个维度进行理解：基础技术演进、生成能力演进和姿势控制演进。在基础技术维度上，从最初的扩散模型[3]，再到具备条件控制机制的视频扩散模型，为该领域奠定了坚实的技术基础。在生成能力维度上，经历了从早期的静态图像生成到解耦控制增强，再到统一架构与多模态融合的演进过程。在姿势控制维度上，从最初的 2D 骨架姿势发展到密集姿势表示，进而到 3D 参数化模型，最终实现多模态姿势融合控制。从最初的姿势引导图像生成到复杂的多种姿势引导视频生成模型，该领域经历了从静态图像到动态视频、从单一姿势控制到多种姿势融合的技术演进过程。本文将按照这一发展脉络，详细阐述相关技术的进步与创新。



图 1　姿势引导人类运动视频生成技术路线整体架构图

Fig. 1　Overall architecture of posture guided human motion video generation technical roadmap

1　技术基础与发展

1.1　用于图像生成的扩散模型

扩散模型是一类概率生成模型，其核心是通过学习扩散过程来生成与给定数据集分布相符的新数据。如图 2 所示，扩散模型涉及两个马尔可夫链：一个正向链（前向过程）将原始数据逐步扰动为噪声；另一个反向链（反向过程）将噪声逐步转换回原始数据。在视觉内容生成任务中，通过训练神经网络来逆转向真实数据添加噪声的过程，从而能够从随机噪声开始逐步生成新的高质量数据。



图 2　扩散模型概述

Fig. 2　Overview of diffusion model

对于来自特定数据分布 p_{data} 的数据样本 x_0，前向扩散过程被定义为一个固定的马尔可夫链，其会逐渐向数据中添加高斯噪声，定义如下：

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I}) \quad (1)$$

其中，$t = 1, \cdots, T$，T 为扰动总步数，x_t 表示在真实数据 x_0 上加 t 步噪声后的噪声数据。该过程由一个噪声调度序列 $\{\beta_t\}_{t=1}^{\mathrm{T}}$ 控制，该序列由噪声步长 t 参数化，通常 β_t 从较小值逐渐增加到较大值。基于正态分布的封闭性质，可以对上述扩散过程进行简化，直接用 x_0 计算任意时刻 x_t：

$$q(x_t \mid x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \quad (2)$$

其中，$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ 且 $\alpha_t = 1 - \beta_t$。这种简化减少了计算复杂度，使训练更加高效。在 DDPM 模型之

后,参数化的去噪函数 $\epsilon_\theta(x_t,t,c)$ 通常由神经网络实现,通过最小化均方误差损失来训练,训练过程的损失函数定义为:

$$\mathcal{L}_{\text{simple}}=\mathbb{E}_{t,x_0,\epsilon}[\|\epsilon-\epsilon_\theta(x_t,t,c)\|^2] \quad (3)$$

式中:c 是可选条件(如文本描述),x_t 是通过添加 t 步噪声得到的真实数据 x_0 的扰动版本,ϵ 是从标准正态分布采样的噪声。通过这种方式,去噪函数 ϵ_θ 可以在可选的条件 c 下,通过从真实数据分布和时间步长 t 中采样 x_0 来训练至收敛。

1.2 用于图像生成的潜在扩散模型

由于在像素域直接运行扩散过程存在信息冗余和计算成本高等问题,潜在扩散模型(Latent Diffusion Model, LDM)通过在低维潜在空间中执行扩散过程来解决这些挑战。LDM 的工作流程可分为以下步骤:

首先训练一个变分自编码器(Variational Autoencoder, VAE)[4],编码器 ε 将高维图像 $x\in R^{h\times W\times 3}$ 压缩为低维潜在表示 $z=\varepsilon(x)\in R^{h\times w\times c}$,其中 $h\ll H,w\ll W$,解码器 D 则将潜在表示重构回原始图像空间。直接在 VAE[4] 的潜在空间中训练扩散模型,对潜在表示 z 而非原始图像 x 执行加噪与去噪。前向扩散过程公式为:

$$q(z_t\mid z_{t-1})=N(z_t;\sqrt{1-\beta_t}z_{t-1},\beta_t\boldsymbol{I}) \quad (4)$$

条件生成时,条件信息 c(如文本嵌入)通过交叉注意力机制或其他条件嵌入方式融入扩散模型。采样时,从随机噪声 z_T 开始,通过迭代去噪过程生成潜在表示 z_0:

$$z_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(z_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t,t,c)\right)+\sigma_t\epsilon \quad (5)$$

其中,σ_t 为采样噪声标准差,$\epsilon\sim N(0,1)$。最后通过 VAE 的解码器 D 将生成的潜在表示映射回图像空间:$\hat{x}=D(z_0)$。

这种方法不仅显著提高了模型的计算效率(潜在空间维度通常比原始图像空间低数十倍),而且保持了其生成高保真度图像的能力。LDM 的潜在空间在保留语义和结构信息的同时,剔除了冗余的高频细节,让扩散过程能够聚焦于图像的关键特征。

此外,为增强对视觉内容生成过程的可控性,近期研究如 ControlNet[5]、T2I-Adapter[6] 和 IP-Adapter[7],探讨了将补充编码器层整合模型的方案。这些补充层有助于模型吸收各种类型的控制信号,包括但不限于姿势、深度和边缘信息,甚至允许将图像与文本提示相结合,实现更精准的条件生成。

1.3 用于视频生成的扩散模型

视频生成比图像生成更具挑战性,因为视频具有更高的维度属性,需要模型同时理解空间关系和时间运动模式。为有效建模时空依赖性,Make-A-Video[8]、ModelscopeT2V[9] 等模型采用 3D-UNet 框架,通过集成时间层(如 temporal Transformers)对 2D-UNet[10] 进行时间扩展。这种范式在后续的许多工作中也得到了广泛应用[11-14]。视频扩散模型的数学表示可以扩展为:

$$q(x_t^{1:F}\mid x_{t-1}^{1:F})=\prod_{f=1}^{F}N(x_t^f;\sqrt{1-\beta_t}x_{t-1}^f,\beta_t\boldsymbol{I}) \quad (6)$$

其中,F 表示视频帧数;x_t^f 表示第 f 帧在时间步 t 的状态;β_t 为扩散步长;$\boldsymbol{I}$ 为单位矩阵。

为捕捉帧间的时间依赖关系,视频扩散模型的神经网络结构通常包含空间编码层:处理单帧的空间信息;时间编码层:建模跨帧的运动和一致性;时空混合层:整合空间和时间信息。时间层的数学表示可以简化为:

$$\text{TL}(X^{1:F})=\text{Attention}(Q^{1:F},K^{1:F},V^{1:F}) \quad (7)$$

其中,$Q^{1:F}$、$K^{1:F}$、$V^{1:F}$ 是从帧序列 $X^{1:F}$ 生成的查询、键和值矩阵。

为实现空间和时间上的更高可控性,Gen-1[15]、VideoComposer[16] 等尝试引入额外的指导条件,例如深度图和运动矢量,以实现可控的常规视频合成[17-19]。条件视频生成的数学表示为:

$$\epsilon_\theta(x_t^{1:F},t,c_{\text{text}},c_{\text{motion}},c_{\text{spatial}}) \quad (8)$$

其中,c_{text} 是文本条件;c_{motion} 是运动控制信号;c_{spatial} 是空间布局控制信号。

最近的视频生成模型主要通过两种方式增强生成能力:一是向预训练的图像生成模型添加时间层来利用扩散模型的优势[8,20-21],二是利用 Transformer 结构建模序列关系[22-23]。这一过程可表示为:

$$\varepsilon_\theta(x_t^{1:F},t,c)=\text{Tadapter}(\epsilon_{\theta_{\text{image}}}(x_t^{1:F},t,c)) \quad (9)$$

其中,$\varepsilon_{\theta_{\text{image}}}$ 是预训练的图像扩散模型,Tadapter 是适应视频生成的时间建模层。

Stable Video Diffusion (SVD)[24] 是基于 LDM 构建的最流行的开源模型之一。该模型为基于图像的视频生成提供了一种简单有效的方法,并作为此类任务的强大预训练模型。SVD 将扩散模型与 VAE 的优势相结合,在 VAE 编码的潜在空间中执行扩散过程,显著提高了视频生成的效率和质量。在视频生成的 VAE 应用中,通常先对视频帧序列中的每一帧应用 2D-VAE 编码:

$$z^f = \varepsilon(x^f),\ f = 1,2,\cdots,F \tag{10}$$

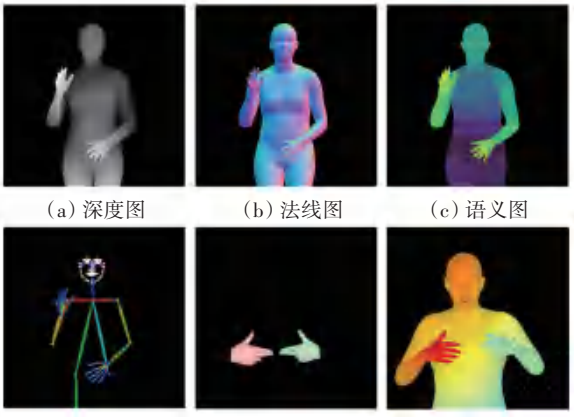
然后在时间维度上添加特定的处理模块(如 3D 卷积或时间注意力机制)来建模帧间关系：

$$\hat{z}^{1:F} = \text{TemporalProcessor}(z^{1:F}) \tag{11}$$

最后，通过解码器重构视频序列：

$$\hat{x}^f = D(\hat{z}^F),\ f = 1,2,\cdots,F \tag{12}$$

这种方法既保持了空间信息的完整性，又能捕捉时间动态，为高质量视频生成奠定了基础。



(a) 深度图　(b) 法线图　(c) 语义图

(d) 骨架图　(e) 3D手部图　(f) SMPL-CS图

图 3　用于引导视频生成的姿态类型

Fig. 3　Pose types used to guide video generation

1.4　用于引导视频生成的姿态类型

人体的关键点和姿态信息通常以人体骨架图的形式表示，如图 3(d)所示，这种表示方式便于在图像或视频生成任务中进行姿态控制和动画生成。骨架图通过连接人体的关键点(如头部、肩膀、肘部、膝盖等)来描绘人体的结构和姿态，为生成模型提供了明确的动作指导。为准确提取这些关键点和姿态信息，研究人员开发了多种先进的姿态估计算法，其中 DWPose[25]、OpenPose[26]是两种广泛应用于骨架图生成的算法。

SMPL(Skinned Multi-Person Linear)[27]是一种用于 3D 人体姿态和形状估计的参数化模型，能够通过低维参数精确描述复杂的人体形状和动作。在将 SMPL 模型应用于生成任务时，需要从参考图像和驱动视频序列中提取相应的模型参数。给定参考图像和驱动动作视频帧，可以利用现有框架(如 4D-Humans)[28]来获取 3D 参数化的 SMPL 模型，从而确保生成内容在形状和动作上与输入条件保持一致。

为提取全面的视觉信息，SMPL 网格会被渲染为 2D 表示形式(如图 3 所示)，包括深度图、法线图和语义分割图等。深度图提供了每个像素到相机的距离信息，这对于 3D 结构重建至关重要，能够帮助模型准确理解场景的深度关系和空间布局。法线图则描述了表面的方向，捕捉与人体表面相关的几何信息，如曲率和凹凸变化，为生成模型提供了丰富的细节信息。此外，语义分割图为每个像素提供了类别信息，有助于模型准确处理人体不同部分之间的交互和关系，如区分手臂、腿部等部位，从而提高生成内容的准确性和真实感。

SMPL-CS 是 SMPL 模型的扩展版本，如图 3(f)所示，在原有 3D 人体姿态与形状估计能力的基础上，进一步整合了颜色信息。通过将纹理和颜色细节融入 3D 人体表示中，SMPL-CS 能够生成更加生动逼真的 3D 人类角色。这种结合几何形状与颜色信息的方法，使得数字人类在各种场景中的表现更加真实自然，适用于对真实感受要求较高的应用领域，如电影特效、虚拟现实等。

HaMeR(Hand Mesh Recovery)[29]是一种基于 Transformer 的 3D 手部网格重建方法，专门用于从单张图像中恢复手部的 3D 形状。HaMeR 利用大规模的 Transformer 架构和丰富的手部数据集，能够精确捕捉手部的各种复杂姿势，如图 3(e)所示，实现手部手势的高精度求解。这对生成逼真的手部动画和交互式应用具有重要意义，能够显著提升生成内容的真实感与交互体验，使手部动作在虚拟场景中更加自然、贴合现实。

2　人类运动视频生成模型

制作人类运动视频以及姿势序列驱动是视频生成领域中一项既具挑战性又极具实用价值的任务。扩散模型在文本到图像领域的最新进展，激发了研究人员将其应用于人体图像动画任务的研究兴趣。随着生成网络技术的快速发展，研究人员已经提出了多种创新方法来解决这一问题。近年来，研究者们在扩散模型的基础上开发了人类视频生成架构，这些架构展现了其卓越的训练稳定性以及通过迭代优化流程生成高保真结果的能力。

2.1　姿势引导的视频生成模型进展

姿势引导的视频生成技术在过去几年中迅速发展，从单一姿势控制到多种姿势融合，从静态图像生成到动态视频生成，经历了多个技术迭代。下文将按照技术发展时间线和方法演进关系，对这些关键技术进行综述。

2.1.1　早期模型与基础技术

PIDM[30](2023 年)是早期将扩散模型应用于姿势引导人类图像生成的代表性工作。该方法引入了

一个纹理扩散模块,专门用于对齐源图像和目标图像的纹理模式,增强了生成图像的真实感。PIDM 的主要贡献在于验证了扩散模型在保留人物身份信息的同时,有效实现了姿势变换,为后续视频生成奠定了基础。

随后,DreamPose[31] 利用预训练的 Stable Diffusion[3] 并引入适配器机制,通过整合 CLIP[32] 和 VAE[6] 进行图像编码,使模型能够在生成过程中考虑更多的上下文信息,如人体姿态和动作序列。DreamPose 的创新在于将预训练的文本到图像模型成功应用于特定的人体姿势驱动任务,实现了从静态图像向初步视频生成的过渡。

2.1.2 模型解耦与控制增强

随着技术的发展,研究人员开始关注如何实现对生成过程的更精细控制。Disco[33] 利用基于 ControlNet 的混合扩散框架,实现了对人类前景、背景和姿态的解耦控制,从而产生可组合化的人类舞蹈场景。这种解耦控制使生成结果既保持了人物身份特征,又准确反映了目标姿势,同时能够灵活处理背景元素。

Animate Anyone[34] 进一步发展了这一思路,利用基于 UNet 的 ReferenceNet 从参考图像中提取特征,并通过轻量级的姿态引导器注入姿态信息。其扩展了 AnimateDiff[35] 引入的原理,将时间层集成到去噪 UNet 中,以增强时间一致性。这使得生成的视频更加流畅自然,有效减少了时间不连贯问题。

MagicAnimate[36] 则在姿势指导方面作出了创新,其采用为 DensePose[37] 量身定制的 ControlNet,而非更常用的 OpenPose 关键点,提供了更精确的姿势指导。这种方法能够捕捉更复杂的人体形态信息,提高了动作生成的精确度和自然度。

2.1.3 模型统一与性能优化

随着研究的深入,统一架构和性能优化成为关注重点。UniAnimate[38] 提出了统一的视频扩散模型,消除了对额外参考网络的依赖,降低了训练难度与模型参数数量。同时处理参考图像和噪声视频,促进特征对齐并确保时间上一致的视频生成。该模型引入了参考图像的姿态图作为额外的参考条件,允许随机噪声视频或第一帧条件视频作为视频合成的输入,保证了视频平滑性。

Champ[39] 则代表了多类型姿态输入融合的技术进步,采用集成自注意力机制的多层运动融合模块,在空间域中融合形状和运动潜在表示。这种多模态融合方法显著提升了生成视频的质量和动作准确性,在复杂动作场景中表现尤为出色。

2.1.4 专项化优化与细节增强

最新的研究方向转向了针对特定问题的专业化解决方案。MimicMotion[40] 引入了置信度感知的姿态引导和基于姿态置信度的区域损失放大,确保了视频帧具备高质量,减少图像失真并提升时间平滑性。这种方法特别关注了姿态估计可能存在的不确定性,通过置信度机制提高了模型的鲁棒性。

RealisDance[41] 则在手部动作生成方面取得了突破,除了在主 UNet 中使用时间注意力外,还将时间注意力插入姿势指导网络,从生成和条件两个方面减轻视频抖动,并采用自适应姿势门控模块融合 3 类不同姿势序列确保动作正确生成。这种方法特别强调了手部动作的精确生成,这在之前的模型中常常是薄弱环节。

这些模型代表了姿势引导的人类运动视频生成技术的演进路径,从单一姿势控制发展到复杂的多模态融合,从解决基础生成问题到关注特定细节的优化,形成了一个逐步完善的技术体系。通常,这些基于扩散的生成框架大多构建在类似 ControlNet 的 3D-UNet 架构之上,以确保时间一致性。同时,为了在生成过程中保留参考图像的复杂外观,这些方法普遍引入了参考编码器。该编码器是 3D-UNet 的简化版本,去除了时间 Transformer 层,使其能够专注于捕捉参考图像的精细纹理和特征。

2.2 姿势引导的视频生成模型架构

姿势引导的视频生成模型概述如图 4 所示。

2.2.1 模型输入

模型的输入通常包含参考图像和一种或多种姿势序列。参考图像:作为生成视频的初始人物图像,为视频生成提供外观特征和身份信息,确保生成的视频在视觉上与输入的参考图像保持一致,如人物的面部特征、服装样式等。姿势序列:用于指导视频生成中的动作变化,包含关键帧的姿势信息,如人体关节的位置和姿态。姿势序列可以是预先定义的,也可以是从其他视频中提取的,使生成的视频能够按照预设的动作进行变化,如跳舞、走路等动作[31, 34,39,41]。姿势序列的准确性对视频生成的质量有重要影响,因此一些模型会引入额外的机制来增强姿势引导的效果[39,41]。

2.2.2 外观控制网络

外观控制网络是用于生成高质量、可控角色动画的关键组件之一。该网络在多种模型中被应用,以确保生成视频中角色的一致性和真实感[33-34,36,38,41]。

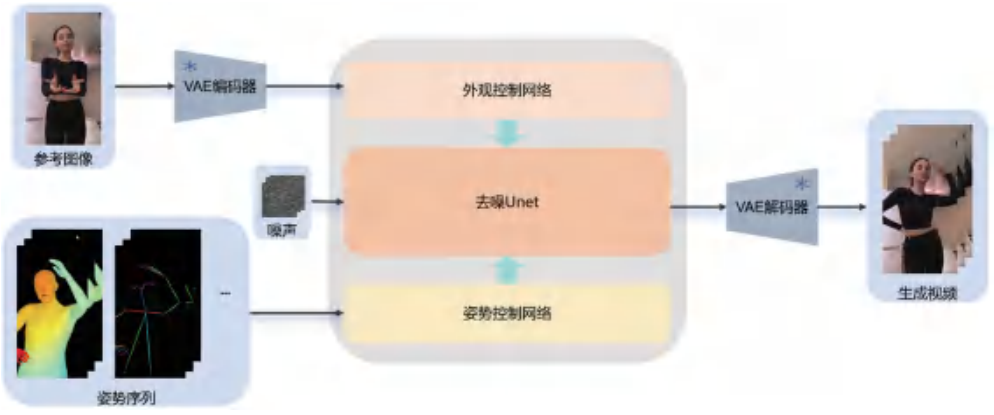


图 4　姿势引导的视频生成模型概述

Fig. 4　Overview of posture guided video generation model

角色一致性：外观控制网络的主要功能是确保生成的视频在角色的外观和身份上与输入的参考图像保持一致[33, 36]。这对于生成高质量的角色动画至关重要，因为角色的服装和面部特征需要在视频中保持稳定；特征提取与合成：通过从参考图像中提取特征，能够捕捉到图像的细节和结构信息，这些信息在生成过程中被用于合成新的视频帧，确保生成的视频帧在视觉上与参考图像高度相似[34,39]。

外观控制网络通常采用与去噪 UNet 相同的对称 UNet 结构，但不包含时间层[36,38]。这种设计使得外观控制网络能够利用从 Stable Diffusion 继承的预训练图像特征建模能力，从而获得良好的初始化特征。在去噪 UNet 的每个对应层块内，通过空间注意力机制完成外观控制网络的特征整合[35,37]。具体操作是将外观控制网络输出的特征图复制多份并与去噪 UNet 的特征图沿宽度维度拼接，然后进行自注意力操作，并提取拼接特征图的前半部分作为输出[33-34]。这种设计使得去噪 UNet 能够选择性地学习与外观控制网络相关的特征，同时保持特征空间的一致性。使得生成的视频帧能够更好地与输入的参考图像特征进行对齐和融合[34,36,38]。该机制有助于提高生成视频的细节质量和一致性。

此外，在图像到视频的转换任务中，尤其是人体动画，需要精确地保持参考图像中的细节特征[34,36,41]。现有方法多采用 CLIP[32] 图像编码器来提取参考图像特征，CLIP[32] 更侧重于语义特征匹配[32-34]。外观控制网络除了空间注意力外，还使用 CLIP 图像编码器进行交叉注意力操作，提取参考图像的语义特征[5,37,39]。这为网络训练提供了有益的初始化，加速了整个网络的训练过程。

2.2.3　姿势控制网络

姿势控制网络对输入姿势图进行分辨率对齐。使其与噪声潜在表示进行有效融合，从而在去噪过程中实现对角色动作的精确控制。这对于生成自然、流畅且与预期姿势一致的角色动画至关重要。

一些使用多种姿势控制的模型[39,41]，其姿势控制网络需要综合考虑来自不同层次的运动信息，这些信息各自包含关于人体形状和运动的不同细节。姿势控制网络可有效整合多类姿势信息，使生成模型能够综合利用不同姿势序列的优势生成高质量的角色动画。以实现更精准和自然的动作表现[36,38]。

姿势控制网络使得生成的视频在动作表现上更加自然和连贯，避免了动作不自然或与预期姿势不一致的问题，从而提高了生成视频的整体质量与一致性。作为一个轻量级的组件，姿势控制网络在实现有效动作控制的同时，并不会显著增加模型的计算复杂度[36,38]。这使得模型能够在保持高效计算的同时，实现高质量的角色动画生成。

2.2.4　去噪 Unet

去噪 UNet 能够在扩散过程中逐步去除噪声，恢复出高质量的图像内容，生成清晰、逼真的角色动画。去噪 UNet 的设计使其能够灵活地处理不同类型的输入和条件信息，如姿势序列、背景信息等，为角色动画生成提供了广泛的适用性和扩展性。

去噪 UNet 是在潜在扩散模型框架下设计的，通常继承 Stable Diffusion 的网络设计和预训练权重，利用其强大的生成能力。在去噪 UNet 的每个层块中，集成了空间注意力、交叉注意力和时间注意力三种注意力机制[34, 36, 41]。空间注意力：用于整合来自外观控制网络的细节特征，确保生成的图像在空间上与参考图像保持一致性，保留参考图像中的细节特征，使得生成的动画在外观上与参考图像高度一致。交叉注意力：利用 CLIP 图像编码器提取的语义特征，进一步增强生成图像的语义一致性。时间

注意力:用于建模视频帧之间的时序关系,确保在生成过程中保持多帧之间的一致性,避免出现帧间抖动或不连贯的问题,确保动画的流畅性,保证生成的视频在时间维度上的连贯性和稳定性[34,36,41]。

3 相关研究的实验方法概述

本节总结了姿势引导人类运动视频生成领域相关研究的实验方法,包括数据集、评价指标和训练策略,旨在为读者提供该领域实验设计的全面视角。

3.1 数据集

表 1 为用于生成人类视频的数据集信息。

TikTok[42]数据集:包含约 350 个舞蹈视频,视频长度 10~15 s,帧率 30 fps,涵盖单人舞蹈动作,部分用于训练,部分用于测试,能较好地评估模型在舞蹈动作生成等任务上的性能[31,33-34,36,38-41,43-44]。

表 1 用于生成人类视频的数据集信息

Table 1 Dataset information for human video generation

数据集名称	公布年份	数据规模
TikTok[42]	2021	约 350 个舞蹈视频,10~15 s,30 fps,超 100 K 张图像
UBC Fashion[45]	2019	339 个训练视频+100 个测试视频,约 12 s,30 fps
TED-talks[46]	2021	1 203 个 YouTube TED 演讲视频片段
COCO[48]	2014	约 328 K 张图像,80 个对象类别
SHHQ[49]	2022	约 50 K 张高质量人体图像
DeepFashion2[47]	2016	约 191 K 张图像,13 种服装类别
LAION[50]	2022	约 50 亿图像-文本对,Disco 预训练使用超 700 K 张图像

UBC Fashion[45]数据集:包含 339 个训练视频和 100 个测试视频,帧率 30 fps,长约 12 s,主要用于训练和测试模型在时尚视频合成任务上的表现,通过随机采样帧对、计算姿态等操作进行训练[31,38-39]。

TED-talks[46]数据集:包含 1 203 个从 YouTube TED 演讲视频中提取的片段,主要用于评估模型在低光照、手势动作为主的演讲视频场景下的表现[34,36]。

DeepFashion2[47]、COCO[48]、SHHQ[49]、LAION[50]等公共数据集:在 Disco[33]中,组合使用这些数据集作为人类属性预训练数据集,先生成人体前景掩码,过滤无人图像后,得到超 700 K 张图像用于预训练,使模型能更好地区分动态主体和静态背景,同时编码复杂的人脸和服装信息。

自建或自收集数据集:部分方法会根据特定需求自建或收集数据[34,38-41,43-44],如 StableAnimator[44]从互联网收集 3 000 个 60~99 s 的视频用于训练,涵盖多样的人物动作表情;Champ[39]中自建了约 5 000 个高保真真实人类视频的数据集,共 100 万帧,来源多样,增强模型对人类运动和服装的分析能力。

3.2 评价指标

3.2.1 图像质量评价指标

L1 误差:计算生成图像与真实图像在像素级的平均绝对差异:

$$L1 = \frac{1}{n}\sum_{i=1}^{n} | x_i - y_i | \tag{13}$$

式中:x_i、y_i 分别表示生成图像和真实图像的像素值。$L1$ 误差值越小,表示生成图像与真实图像越接近[31,33-34,38-41,43-44]。

峰值信噪比(PSNR):衡量图像重建质量的指标,通过计算最大可能信号功率与噪声功率之比来评估生成图像质量:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{14}$$

式中:MAX_I 为图像最大像素值,MSE 为均方误差。PSNR 值越高,表示图像质量越好[30,33-34]。

结构相似性指数(SSIM):考虑人类视觉系统特性的图像质量评价指标:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{15}$$

式中:μ_x、μ_y 是图像的均值,μ_x^2、μ_y^2 为方差,σ_{xy} 为协方差,c_1、c_2 为常数。SSIM 值越接近 1,表示结构相似性越高[36,38-41,43-44]。

LPIPS:基于深度学习的感知图像相似性指标,在多种方法中用于衡量生成图像与真实图像的感知相似性,值越低说明相似度越高,能较好地反映人类视觉感知[31,33-34,36,38-39,43-44]。

FID:衡量生成图像与真实图像之间分布差异的常用指标,值越低表示分布越接近,生成质量越好,被广泛应用[30,33,36,44]。

3.2.2 视频质量评价指标

FVD:专注于衡量生成视频与真实视频在特征分布上的差异,考虑了空间和时间维度,是评估视频保真度的重要指标,被多种方法采用[30,33-34,36,38-40,43]。

FID-VID:结合了图像分布差异和视频感知质量,用于衡量视频生成质量,在 Disco[33]一文中使用了该指标。

AKD、MKR、AED：MagicAnimate[36] 中使用这些指标衡量生成视频中关键点的平均距离、缺失关键点比率、平均欧几里得距离等，用于评估动画的精度和完整性。

3.3　训练策略

相关研究通过优化训练策略提升模型性能，常见的策略包括分阶段训练、预训练与微调、数据增强、模块优化和引导策略。以下详细分析这些策略及其设计逻辑。

分阶段训练：视频生成涉及多维目标（外观、动作、时间一致性），一次性训练可能导致模块竞争或过拟合。分阶段训练降低优化难度，允许模型先掌握基础生成能力，再适应动态序列，从而提高训练稳定性和生成效果。如 DreamPose[31] 的两阶段微调策略，先在完整训练数据集上微调 UNet[9] 和适配器模块，再在特定主题输入图像上进一步微调；disco 中微调是在预训练基础上进行；Champ[39] 分两个阶段，第一阶段仅用图像训练排除运动模块，第二阶段引入运动模块增强时间连贯性；Animate Anyone[34] 也分两阶段，先用单帧训练，再引入时间层训练。这种策略能让模型逐步学习不同方面的知识和技能，先掌握基础的图像生成能力，再增强时间维度上的连贯性和一致性等。

预训练与微调：一些方法会先进行预训练任务，如 disco[33] 中设计预训练任务，通过基于分离的前景和背景区域特征重建完整图像，使模型更好地区分动态主体和静态背景；MimicMotion[40] 采用预训练的 SVD[24] 对图像到视频模型的权重进行初始化。预训练能让模型提前学习到一些通用的特征和知识，在此基础上进行微调，使其更好地适应特定的任务和数据。

数据增强与采样：部分方法在训练过程中采用数据增强技术，如 MagicPose[44] 中随机遮盖面部标志和手部姿态，增强模型对检测结果不完整时的鲁棒性；UniAnimate[38] 在训练阶段随机输入 16 或 32 帧的视频片段到模型中学习时间一致性。这些方法能增加数据的多样性，使模型在训练时接触到更多变化的情况，提高其泛化能力。

特定模块优化与权重调整：在训练过程中，会针对特定模块进行优化或调整权重。如 MagicAnimate[36] 在第二阶段仅优化视频扩散模型中的时间注意力层；Champ[39] 在第二阶段训练运动模块时，保持其他模块权重不变。通过这种方式，能让不同模块在合适的阶段发挥其作用，专注于学习特定的任务，提高模型的整体性能。

引导策略与优化技巧：一些方法会采用引导策略或优化技巧来提升生成质量。如 DreamPose[31] 利用图像和姿态分类器自由引导，通过引导权重控制输出图像与输入图像和姿态的相似度；StableAnimator[43] 在推理阶段引入基于 HJB（Hamilton-Jacobi-Bellman）方程的面部优化策略，提升面部质量。这些方法能对生成过程进行更精细的调控，使生成结果更符合预期。

4　结束语

姿势引导的人类运动视频生成技术作为 AIGC 领域的重要分支，近年来取得了显著进展。扩散模型的引入为生成高质量、时序连贯且与条件约束一致的人类视频提供了坚实的技术基础。本文综述了该领域的最新技术进展，详细介绍了扩散模型在图像和视频生成中的应用，探讨了多种姿势引导的视频生成模型的网络架构、技术特点，以及实验方法中的数据集、评价指标和训练策略。

尽管取得了诸多成果，但生成高质量的人类运动视频仍面临诸多挑战，包括但不限于动作的自然性、视频的时序连贯性以及与输入条件的一致性。未来的研究可以进一步优化模型架构以提高生成视频的质量和连贯性，探索更有效的训练策略以增强模型的泛化能力，以及开发更先进的姿态估计算法以提供更精确的姿势指导为方向。期望通过这篇综述，为研究者提供深入的见解，并激发新的研究思路，以推动姿势引导的人类运动视频生成技术在虚拟现实、娱乐、教育和医疗等领域的广泛应用和发展。

参考文献

[1] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2020: 6840-6851.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[3] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10684-10695.

[4] Kingma D P. Auto-encoding variational bayes[J]. arXiv preprint arXiv, 1312.6114, 2013.

[5] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 3836-3847.

[6] Mou C, Wang X, Xie L, et al. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models [C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2024, 38(5): 4296-4304.
[7] Ye H, Zhang J, Liu S, et al. Ip-Adapter: Text compatible image prompt adapter for text-to-image diffusion models [J]. arXiv preprint arXiv, 2308.06721, 2023.
[8] Singer U, Polyak A, Hayes T, et al. Make-A-Video: Text-to-video generation without text-video data[J]. arXiv preprint arXiv, 2209.14792, 2022.
[9] Wang J, Yuan H, Chen D, et al. ModelScope text-to-video technical report[J]. arXiv preprint arXiv, 2308.06571, 2023.
[10] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.
[11] An J, Zhang S, Yang H, et al. Latent-Shift: Latent diffusion with temporal shift for efficient text-to-video generation[J]. arXiv preprint arXiv, 2304.08477, 2023.
[12] Ceylan D, Huang C H P, Mitra N J. Pix2Video: Video editing using image diffusion [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 23206-23217.
[13] Qing Z, Zhang S, Wang J, et al. Hierarchical spatio-temporal decoupling for text-to-video generation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 6635-6645.
[14] Wang X, Zhang S, Yuan H, et al. A recipe for scaling up text-to-video generation with text-free videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 6572-6582.
[15] Esser P, Chiu J, Atighehchian P, et al. Structure and content-guided video synthesis with diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 7346-7356.
[16] Wang X, Yuan H, Zhang S, et al. VideoComposer: Compositional video synthesis with motion controllability [C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2024: 7594-7611.
[17] Xing J, Xia M, Liu Y, et al. Make-Your-Video: Customized video generation using textual and structural guidance[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(12): 7505-7517.
[18] Zhang Y, Wei Y, Jiang D, et al. ControlVideo: Training-free controllable text-to-video generation [J]. arXiv preprint arXiv, 2305.13077, 2023.
[19] Zhao M, Wang R, Bao F, et al. ControlVideo: Adding conditional control for one shot text-to-video editing[J]. arXiv preprint arXiv, 2305.17098, 2023.
[20] Wang W, Liu J, Lin Z, et al. MagicVideo-V2: Multi-stage high-aesthetic video generation[J]. arXiv preprint arXiv, 2401.04468, 2024.
[21] Wu J Z, Ge Y, Wang X, et al. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 7623-7633.
[22] Yan W, Zhang Y, Abbeel P, et al. Videogpt: Video generation using vq-vae and transformers[J]. arXiv preprint arXiv, 2104.10157, 2021.
[23] Yu L, Cheng Y, Sohn K, et al. Magvit: Masked generative video transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 10459-10469.
[24] Blattmann A, Dockhorn T, Kulal S, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets[J]. arXiv preprint arXiv, 2311.15127, 2023.
[25] Yang Z, Zeng A, Yuan C, et al. Effective whole-body pose estimation with two-stages distillation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 4210-4220.
[26] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7291-7299.
[27] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model [M]//Seminal Graphics Papers: Pushing the Boundaries. ACM, 2023: 851-866.
[28] Goel S, Pavlakos G, Rajasegaran J, et al. Humans in 4D: Reconstructing and tracking humans with transformers [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 14783-14794.
[29] Pavlakos G, Shan D, Radosavovic I, et al. Reconstructing hands in 3D with transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 9826-9836.
[30] Bhunia A K, Khan S, Cholakkal H, et al. Person image synthesis via denoising diffusion model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 5968-5976.
[31] Karras J, Holynski A, Wang T C, et al. Dreampose: Fashion image-to-video synthesis via stable diffusion[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 22623-22633.
[32] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2021: 8748-8763.
[33] Wang T, Li L, Lin K, et al. Disco: Disentangled control for realistic human dance generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 9326-9336.
[34] Hu L. Animate anyone: Consistent and controllable image-to-video synthesis for character animation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 8153-8163.
[35] Guo Y, Yang C, Rao A, et al. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning [J]. arXiv preprint arXiv, 2307.04725, 2023.
[36] Xu Z, Zhang J, Liew J H, et al. MagicAnimate: Temporally consistent human image animation using diffusion model [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1481-1490.
[37] Güler R A, Neverova N, Kokkinos I. DensePose: Dense human pose estimation in the wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7297-7306.

[38] Wang X, Zhang S, Gao C, et al. UniAnimate: Taming unified video diffusion models for consistent human image animation[J]. arXiv preprint arXiv, 2406.01188, 2024.

[39] Zhu S, Chen J L, Dai Z, et al. Champ: Controllable and consistent human image animation with 3D parametric guidance [C]//Proceedings of the European Conference on Computer Vision. Cham: Springer , 2024: 145-162.

[40] Zhang Y, Gu J, Wang L W, et al. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance[J]. arXiv preprint arXiv, 2406.19680, 2024.

[41] Zhou J, Wang B, Chen W, et al. RealisDance: Equip controllable character animation with realistic hands[J]. arXiv preprint arXiv, 2409.06202, 2024.

[42] Jafarian Y, Park H S. Learning high fidelity depths of dressed humans by watching social media dance videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12753-12762.

[43] Tu S, Xing Z, Han X, et al. StableAnimator: High-quality identity-preserving human image animation[J]. arXiv preprint arXiv, 2411.17697, 2024.

[44] Chang D, Shi Y, Gao Q, et al. MagicDance: Realistic human dance video generation with motions & facial expressions transfer [J]. arXiv preprint arXiv, 2311.12052, 2023.

[45] Zablotskaia P, Siarohin A, Zhao B, et al. DwNet: Dense warp-based network for pose-guided human video generation[J]. arXiv preprint arXiv, 1910.09139, 2019.

[46] Siarohin A, Woodford O J, Ren J, et al. Motion representations for articulated animation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13653-13662.

[47] Liu Z, Luo P, Qiu S, et al. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1096-1104.

[48] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[49] Fu J, Li S, Jiang Y, et al. StyLEGAN-Human: A data-centric odyssey of human generation[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2022: 1-19.

[50] Schuhmann C, Beaumont R, Vencu R, et al. LAION-5B: An open large-scale dataset for training next generation image-text models[C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2022: 25278-25294.