中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-14

论文引用格式： Sun Renjie, Sun Yubao, Shao Shuai, Shuai Hui, Liu Qingshan. Text driven 3D human motion diffusion generation based on local generation and global fusion[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.250606. (孙仁杰, 孙玉宝, 邵帅, 帅惠, 刘青山. 局部生成和全局融合的文本驱动三维人体动作扩散生成[J/OL]. 中国图象图形学报, XXXX:1-14. DOI: 10.11834/jig.250606.)[DOI: 10.11834/jig.250606]

局部生成和全局融合的文本驱动三维人体动作扩散生成

孙仁杰[1],孙玉宝[1],邵帅[1],帅惠[2],刘青山[2]

1. 南京信息工程大学,计算机学院,南京 210044; 2. 南京邮电大学,人工智能学院,南京 210023

摘 要：目的 根据文本提示生成三维人体动作是多模态生成领域的前沿研究方向。尽管当前已经取得了诸多的研究进展,但现有方法在语义对齐精度、局部动作控制和全局协调性方面存在局限,难以实现从文本到高保真三维资产的一体化生成。针对上述问题,本文提出一种局部生成与全局融合的级联式扩散生成框架。**方法** 首先,利用大语言模型将输入文本自动解耦为头部、四肢及躯干等六个部位的独立语义描述;其次,构建六路并行、梯度隔离的局部扩散编码器,为各部位独立生成动作特征;再次,设计全局融合网络将局部特征融合为符合生物力学的全身姿态,并解码为SMPL(a skinned multi-person linear model)参数化网格;最后,将SMPL网格转换为3D高斯表示,并引入二维扩散模型作为视觉先验,通过分数蒸馏采样优化其外观细节,实现从文本到可实时渲染三维人体的一体化生成。**结果** 在HumanML3D(3D human motion-language Dataset)和KIT-ML(the KIT motion-language dataset)数据集上开展了对比实验,并从FID(Fréchet inception distance)、和CLIP-S(CLIP similarity)两个维度评估分析本文以及基线对比方法的生成结果。相较于基线方法,本文方法在生成质量和动作准确度方面均有提升,消融实验验证了本文设计思路的有效性。**结论** 本文方法能够有效提升所生成人体动作的细节表现力、多样性以及文本语义一致性,为三维人体动作生成提供了高效、可扩展的技术方案。

关键词： 人体动作生成;局部生成;全局融合;扩散模型;三维高斯溅射

Text driven 3D human motion diffusion generation based on local generation and global fusion

Sun Renjie[1], Sun Yubao[1], Shao Shuai[1], Shuai Hui[2], Liu Qingshan[2]

1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract: Objective Text-driven 3D human motion generation has emerged as a frontier research direction in multimodal content creation, holding great promise for applications in virtual reality, film production, and the metaverse. Despite significant progress, existing methods still face fundamental challenges in three aspects: precise semantic alignment between natural language descriptions and generated motions, fine-grained control over individual body parts, and global coordina-

收稿日期： XXXX-XX-XX;**修回日期：** XXXX-XX-XX

基金项目： 国家自然科学基金(项目编号:92470126,U24B20155,62276139),国家重点研发计划项目(项目编号:2022YFC2405602);江苏省重大研究计划(项目编号:BG2024042);江苏省青蓝工程.

Supported by: National Natural Science Foundation of China(Grant No. 92470126,U24B20155,62276139), National Key Research and Development Program of China (Grant No. 2022YFC2405602), Major Research Program of Jiangsu Province (Grant No. BG2024042), Qinglan Project of Jiangsu Province of China.

tion that respects biomechanical constraints. Consequently, current solutions often suffer from semantic leakage unnatural postures, and limited expressiveness. Moreover, most approaches either focus solely on motion synthesis without producing complete 3D assets, or generate static avatars without dynamic pose control. To address these limitations, we propose a novel cascaded diffusion framework that follows a "local-to-global, structure-to-appearance" generation pipeline, enabling end-to-end synthesis from raw text to high-fidelity, real-time renderable 3D human models with precise motion control.

Method Our framework consists of four key stages, each designed to address a specific aspect of the text-to-3D human generation problem. First, a semantic decoupling module leverages a large language model (GPT-4) to automatically parse the input text into independent action descriptions for six anatomical body parts: head, left arm, right arm, torso, left leg, and right leg. This decomposition converts a global motion description into a set of part-specific textual instructions, explicitly separating semantics across different body regions. For body parts not mentioned in the original text, the parser assigns a "do nothing" instruction, preventing unintended movements. This step is crucial because it transforms a loosely coupled global description into a structured, machine-readable format that guides subsequent generation. Second, we construct a local motion generation module composed of six parallel diffusion-based encoders, each conditioned on its corresponding part description. These encoders operate with gradient isolation, meaning that the training and inference processes for different body parts do not share gradients. This design fundamentally prevents semantic leakage—a common issue in prior work where an action described for one body part inadvertently affects others. Each encoder adopts a transformer-based denoising network. Starting from pure Gaussian noise, the network iteratively refines a latent code guided by the corresponding part text embedding produced by a pre-trained TMR encoder. The resulting latent representation captures the fine-grained motion characteristics of that specific body part, such as the trajectory, speed, and joint angles. Importantly, because the six encoders are independent, they can be trained in parallel on part-specific motion data extracted from full-body motion capture datasets. Third, a global motion fusion module integrates the six independent part latents into a coherent full-body pose. Simple concatenation of part latents would ignore the biomechanical dependencies between body regions. To address this, we employ a lightweight feed-forward network with GELU (gaussian error linear units) activation, augmented by the global semantic feature of the complete text. This network learns to enforce biomechanical constraints such as torso leaning backward during a forward kick, natural arm-leg coordination during walking, and maintaining overall balance. The fused latent is then decoded into SMPL parameters, producing a parametric human mesh that respects human skeletal kinematics. Fourth, for appearance enhancement and efficient rendering, we convert the SMPL mesh into a set of 3D Gaussians—a modern explicit representation that supports real-time differentiable rasterization. Each Gaussian is defined by its position, covariance matrix, opacity, and spherical harmonics coefficients for color. To enrich geometric and textural details beyond the smooth SMPL mesh, we adopt a state-of-the-art 2D diffusion model (Flux) as a powerful visual prior. Through SDS (score distillation sampling), gradients from the 2D diffusion model are backpropagated to iteratively optimize the attributes of the 3D Gaussians while keeping their positions fixed to preserve the generated motion. This optimization runs for 4000 iterations, refining details such as skin texture, clothing wrinkles, and lighting effects. The final output is a fully textured 3D human model that can be rendered in real time without any post-processing.

Result We conduct extensive experiments on two standard benchmarks, HumanML3D and KIT-ML, and compare our method against representative baselines including MotionDiffuse, MDM, MLD, DreamFusion, GaussianDreamer, and others. For quantitative evaluation, we employ multiple metrics. FID (Fréchet Inception Distance) measures the realism and diversity of generated motion sequences. CLIP-S (CLIP similarity) evaluates semantic alignment between rendered multi-view images and input text. Additionally, we introduce a Part-FID (part-level Fréchet inception distance), which computes FID separately for each of the six body parts using dedicated feature extractors, providing a fine-grained assessment of local motion quality. Experimental results demonstrate that our method achieves an FID of 0.429, comparable to MotionDiffuse (0.687) and MDM (0.747). In terms of CLIP-S, our method attains 29.41 (ViT-L/14) and 44.39 (ViT-bigG-14), surpassing GaussianDreamer (27.23 and 41.88) and other text-to-3D baselines. The proposed Part-FID yields an average score of 1.26, which is 18.7% better than MotionDiffuse, with the most significant improvement observed on the torso, validating the effectiveness of our global fusion module in enforcing biomechanical coordination. Ablation studies further confirm the contribution of each component: removing gradient isolation increases semantic leakage. Efficiency

analysis shows that our method takes approximately 20 minutes for end-to-end generation, and the final 3D Gaussian representation enables real-time rendering at 24 frames per second, which is two orders of magnitude faster than NeRF-based renderers. **Conclusion** we present a comprehensive framework for text-driven 3D human motion generation that uniquely combines local motion generation with global fusion, supported by efficient 3D Gaussian splatting and a powerful 2D diffusion prior. The method achieves superior performance in motion realism, part-level control accuracy, semantic alignment, and rendering efficiency. It provides an end-to-end solution from natural language to high-quality, real-time renderable 3D human assets, opening new possibilities for interactive virtual human applications. Future work will focus on extending the framework to generate long-sequence motions with temporal consistency and incorporating multimodal control signals.

Key words: Human motion generation; Local generation; Global fusion; Diffusion model; 3D Gaussian Splatting

0 引 言

随着虚拟现实VR(virtual reality)、增强现实AR(augmented reality)、影视动画和游戏产业的快速发展,三维人体动作生成技术成为数字内容创作的核心需求之一(杨航等,2023)。传统的人体动作生成方法主要依赖手工设计或运动捕捉技术,存在成本高、效率低、灵活性差等问题。近年来,生成式人工智能的突破为三维内容自动化生成提供了新的解决方案,尤其是扩散模型在图像、视频和三维生成任务中展现出强大的潜力。文本驱动的人体动作生成因其自然交互性和语义丰富性成为研究热点(冯明涛等,2025),用户通过自然语言描述如"一个人用左腿踢腿"即可生成对应的三维动作,显著降低了创作门槛。为了完成此任务,研究人员开展了深入研究,提出了诸多算法,主要分为三类:1)基于编码器-解码器的方法(Petrovich等,2022;Ghosh等,2021),该类方法的核心思路是将文本描述和动作编码到共享隐空间后解码生成动作,结构简单,但隐空间表达能力有限,难以处理复杂语义,且生成的动作往往趋于平均化,多样性较差。2)基于扩散模型的方法,如MDM(Tevet等,2022)、MotionDiffuse(Zhang等,2024)等,利用逐步去噪生成高质量动作,但其采用全局编码策略,即将整个人体作为一个整体进行去噪生成,所有身体部位共享相同的文本条件和网络参数。这种设计虽能生成合理的整体动作,但存在语义泄漏和局部控制能力弱的缺陷,难以实现精细的部位级控制。3)结合2D/3D扩散模型的方法,如GaussianDreamer(Yi等,2024)、DreamWaltz(Huang等,2023)、DreamFusion(Poole等,2022)等,通过2D扩散模型优化3D生成结果,实现了文本驱动的三维虚拟人生成,但其生成过程中难以根据文本精确控制身体姿态,输出的虚拟人通常为默认的T-pose或A-pose,缺乏姿态多样性。同时,精细动作的三维一致性难以保证,尤其是对于结构复杂的实例。

总体而言,由于人体运动的多样性以及运动描述的复杂性,现有方法仍存在生成质量较低,全局协调性差,语义对齐不足等问题。为此,本文提出一种融合局部生成与全局优化的文本驱动三维人体动作扩散生成框架,使生成的结果既具有3D扩散模型的几何学一致性,又能够从2D扩散模型中获得丰富的细节。该框架的实现流程始于文本输入的语义解耦,将自然语言描述精细化分解为各个身体部位的独立动作语义。在此基础上,通过这些部位的独立描述生成局部动作,并通过全连接层将离散的局部特征拼接融合,最终合成符合人体运动力学的全身静态姿态。该姿态被转换为三角网格后,经点云采样与噪声注入生成初始化的3D高斯人。为进一步提升生成质量,引入2D扩散模型驱动分数蒸馏采样(score distillation sampling,SDS)损失函数,依托多视角渲染和梯度回传的迭代策略优化高斯点的纹理与光照属性。最终生成的3D资产无需转换结构即可进行实时渲染。实验结果显示,本文提出的框架在CLIP-S(CLIP-similarity)指标上得分为29.41,对比基线模型(GaussianDreamer,得分27.23)提升约8%。生成的三维人体表现出更精细的局部细节和整体协调性,在语义一致性和生成质量上均取得了较好的提升。

1 相关工作

1.1 3D人体表示方法

三维动作表示是3D人体生成任务的基础,3D

表示方法直接影响生成任务的表达能力与计算效率。近年来,神经辐射场(representing scenes as neural radiance fields for view synthesis,NeRF)在3D表示方面取得了相当不错的成果,许多文本到3D资产生成方法也采用了NeRF(Mildenhall等,2021)或其变体(Barron等,2021;Müller等,2022)作为表示方法。网格和点云表示等传统方法提供了高精度的几何信息,但也存在局限性,在处理复杂动态动作尤其是人体动作时计算复杂度较高,难以捕捉时间依赖特性。此外,3D高斯(Kerbl等,2023)作为最新提出的创新性三维表示方法,在三维场景的生成与渲染领域表现出显著优势。其核心思想是利用空间中的高斯分布对三维数据进行稀疏表示,每个三维点以中心坐标、方差和颜色分布进行定义。这种方法通过将复杂三维场景分解为稀疏的高斯分布集,显著降低了传统点云或体素表示的存储和计算成本。这种稀疏化处理不仅提升了计算效率,还减少了冗余数据,提高了模型在高维数据处理中的表现,同时3D高斯能通过高斯分布的连续性捕捉时间和空间的细节变化,从而更适合动态变化的三维生成任务。使用3D高斯作为表示方法,与可优化网格表示方法相比,其显著降低了分辨率提升的成本,并具有更快的优化速度,可以根据提示文本在很短的时间内生成详细的3D资产。

1.2 三维资产生成

扩散模型已成为三维资产生成的主流技术之一,相较于其他生成模型如生成对抗网络(generative adversarial networks,GAN)和变分自编码器(variational auto-encoder,VAE),扩散模型在高维数据建模和生成多样性方面更具优势。其逐步去噪生成过程在捕捉复杂动态特性和高维依赖关系方面表现优异。目前该研究方向主要分为两类:一类是三维预训练扩散模型,另一类是将二维扩散模型扩展到三维。两者的区别在于所使用的训练数据是二维还是三维。三维预训练扩散模型如一些工作(Gupta等,2023;Jun等,2022)是基于文本——三维数据对预训练的模型,在预训练之后,只需通过推理即可生成三维资产。例如,Point-E(Nichol等,2022)和Shape-E(Jun等,2023)等模型能够在几分钟内生成三维资产。除了从文本生成三维资产外,还有一些方法利用三维扩散模型基于文本-动作数据生成动作序列(Dabral等,2023;Raab等,2023;Hu等,2023)。这些模型通过预训练能够针对不同的文本生成合理的动作序列,并可将生成的动作序列通过网格表示转换为SMPL模型。由于SMPL模型不包含纹理信息,可通过不同的文本提示对转换后的SMPL模型进行着色。

在文本到三维资产生成方法中,除了使用三维预训练扩散模型外,将二维扩散模型提升到三维同样是一种有效的方法。由于二维图像数据集的丰富性,该方法能够生成具有更高多样性和保真度的三维资产。一些单图像到三维的生成方法也采用了类似的思想。DreamFusion(Poole等,2022)首次提出了SDS方法,即使用二维扩散模型来更新三维表示模型。Wang等人(2022)提出了SJC(score Jacobian chaining)方法,将二维扩散模型扩展到三维。后续的一些方法(Chen等,2023;Li等,2023)在DreamFusion的基础上改进了三维生成的质量。此外,CLIP(Radford等,2021)模型的引入也有效改善了三维生成的语义对齐能力。这些技术突破为扩展扩散模型在三维生成中的应用提供了坚实基础,同时也指出了未来优化的研究方向。

1.3 文本到运动生成技术

文本到运动生成近年来成为计算机动画和人工智能研究的热点方向之一,其目的是通过自然语言描述生成与之对应的人体运动。由于其易于使用和编辑的特点,成为了指导运动生成的一种十分方便用户使用的接口,具有较高的应用价值。

早期的文本到动作生成方法多基于编码器-解码器框架,通过对文本和动作的隐空间对齐实现生成。例如,Language2pose(Ahuja等,2019)中提出了交替编码文本和动作并将其解码回动作的训练策略;TEMOS等工作中则使用联合编码文本与动作的方式,并通过额外的损失函数优化隐空间对齐。这类方法能够生成基本的动作序列,但在处理复杂或长文本描述时表现受限,例如需要手动对长文本分段或设定动作时长(Athanasiou等,2022)。Tevet等人(2022)采用了冻结的CLIP模型对文本编码,利用视觉先验的改进方法更好地对齐文本与动作模态,并将其与动作隐空间对齐。然而,使用随机动作帧渲染的图像进行对齐可能会引入语义混乱。此外,Petrovich等人(2023)指出,动作相关文本在CLIP隐空间中聚集较紧,这种分布特性会限制生成模型对文本细节的捕捉能力。

鉴于扩散模型良好的生成能力，在文本到动作生成问题中也得了广泛应用。模型MDM和MotionDiffuse利用Transformer在扩散过程中基于文本对动作去噪；Chen等人(2023)则采用基于U-Net的DDIM(denoising diffusion implicit models)生成模型，在隐空间中实现高效的动作生成。这些方法生成的动作在质量上获得了显著提升。

当前，文本到动作生成技术已在用户友好性和多样性方面取得了显著进展，但仍面临诸多问题和挑战。首先，现有方法在局部动作语义控制方面表现出明显的不足。全身各部位共享相同的文本信息，导致生成的动作在局部语义匹配上缺乏细致表达，容易引发语义泄漏等问题。此外，基于CLIP(contrastive language-image pre-training)模型的隐空间对齐虽提高了模态匹配性，但对于人体动作描述而言，不同文本在CLIP文本隐空间中的分布往往比较紧密、聚类程度高，从而导致生成模型难以捕捉动作文本描述中的细节，同时，生成模型对长文本的理解能力有限，在处理复杂描述时，往往需要依赖手动筛选，导致生成过程难以自动化。最后，动作生成方法中仍存在控制精度不足和生成结果语义多样性受限等问题。这些问题在实际应用中影响了生成动作的保真度和适用性。

1.4 外观渲染优化

近年来，2D扩散模型的相关研究已从追求基本生成能力，逐步转向追求更高的提示词遵循度、更好的构图能力和更高的生成效率。SD(stable diffusion)(Rombach等，2022)及其变体SDXL(improving latent diffusion models for high-resolution image synthesis)(Podell等，2023)通过将扩散过程在潜在空间中进行，较好地平衡了质量与计算成本，成为当前文本到3D研究中最常用的模型。

然而，SD系列模型在处理复杂文本提示、生成精细细节和避免概念混淆方面仍存在局限。为此，Batifol等人提出的Flux作为新一代生成模型被提出，旨在攻克这些难题。Flux采用了经过大规模、高质量、多模态数据集训练的先进架构(通常基于Transformer或改良的U-Net)，其核心创新在于显著提升了提示词的理解深度与生成图像的语义对齐精度。相较于SDXL，Flux在生成图像的光影合理性、纹理细节(如皮肤质感、织物纹理)以及复杂组合概念的视觉还原方面表现出显著优势。

本研究的渲染优化阶段依赖于预训练的2D扩散模型作为视觉感知先验。鉴于Flux在生成质量和文本忠实度上的突破，它为解决文本到3D生成中的语义鸿沟和细节匮乏问题提供了更强大的先验。本研究采用Flux作为SDS的驱动引擎，旨在将其卓越的2D生成能力蒸馏至3D空间，从而获得纹理丰富、细节逼真且与文本描述高度一致的三维人体动作模型。

2 文本驱动三维人体动作生成方法

2.1 整体框架设计

本文提出一种局部-全局协同的三维人体生成框架，旨在通过端到端的分层处理实现文本到三维动作的高保真生成。如图1所示，该框架由三个阶段构成：首先第一阶段是语义解耦层，在该阶段中利用大语言模型结合所给出的文本提示，将复合动作描述(如“左腿踢腿”)解耦为头部、四肢及躯干等六个部位的独立语义单元；第二阶段中主要为动作生成层，通过并行化的局部特征编码器集群提取部位级运动特征，再经全身运动融合模块融合为符合生物力学的全身静态姿态，并解码为SMPL参数化网格；最终，第三阶段为渲染增强层，通过将输入的SMPL网格转换为3D高斯点表示，借助多视角渲染与2D扩散模型的SDS损失梯度迭代优化纹理与光照物理属性，最终通过实时高斯溅射输出高保真3D人物动作。

文本语义解耦与预处理是框架的初始阶段，其核心任务是将自然语言描述转化为结构化解耦的语义表示。该过程包含以下几个关键步骤：

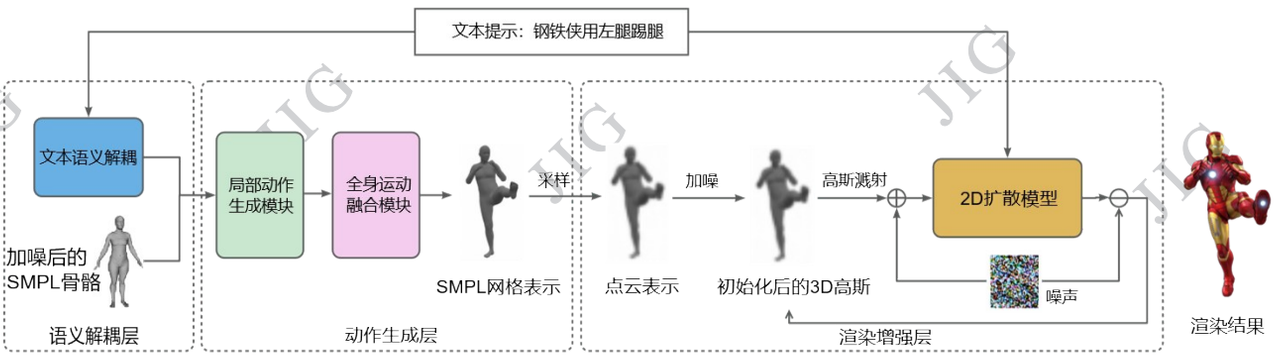
为了确保生成动作的普遍性，避免模型偏向特定角色特征，本文首先对输入文本做一个简化的预处理，通过移除输入文本中的角色信息，将特定角色描述(如“钢铁侠用左腿踢腿”)泛化为标准动作表达(“一个人用左腿踢腿”)。随后进入语义解析阶段，利用GPT-4的知识推理能力，将泛化文本T分解为六部位独立描述T_p：

$$T_p = \left\{ \left(p, \Psi_{GPT4}(T,p) \right) | p \in \boldsymbol{P} \right\} \tag{1}$$

其中部位集合定义为：

$$\boldsymbol{P} = \{ h,la,ra,t,ll,rl \} \tag{2}$$

式中，解析器Ψ_{GPT4}表示使用GPT-4来为每个部位p



2.2 文本语义解耦与预处理

图1 整体模型结构

Fig. 1 Overview of the framework

生成对应的精准动作描述(如左腿对应“左腿向前快速伸展”),未提及部位标注为“什么也不做”。集合$\boldsymbol{P}$则包含了头,左臂,右臂,躯干,左腿,右腿这六个不同的身体部位。

2.3 局部动作生成模块

局部动作生成模块采用人体部位解耦的并行编码架构,本模块为六个身体部位(头部、左/右臂、躯干、左/右腿)分别构建了独立的条件去噪网络,每个网络的功能是:在给定对应部位文本描述的条件下,从随机噪声中逐步去噪,生成符合语义的动作特征。

如图2所示,该模块需要在大规模人体运动数据上训练一组局部运动编码器$E_h, E_{la}, E_{ra}, E_t, E_{ll}, E_{rl}$。每个编码器经过训练可以学习每个部位的高维真实部位姿态数据(如SMPL参数)与对应部位文本之间的映射。具体来说,该编码器首先会采用在AMASS(archive of motion capture as surface shapes)动作语料上预训练的TMR文本编码器(Petrovich等,2022),将每个来自于预处理阶段生成的对应部位文本描述T_p映射为128维特征向量$\boldsymbol{f}_t^p$。该编码器针对运动动词优化,显著增强局部语义的表示精度。至此,原始文本被转化为结构化特征集$\{\boldsymbol{f}_t^h, \cdots, \boldsymbol{f}_t^{rl}\}$。与此同时,将分解后的对应身体部位(如头部、左臂或右腿)的特征通过一个线性变化,变换到128维从而与文本编码器输出的结果对齐。

编码器会将真实的身体部位姿态映射为特征z_p^0,此特征将作为局部扩散模型训练的目标。在推理时,扩散过程从随机采样的高斯噪声z_p^T开始,并通过学习到的去噪过程逐步优化为有意义的特征z_p^0。将一个部位的运动特征记为z_p,对应的文本输入记为$\boldsymbol{f}_t^p$,扩散过程建模为如下马尔科夫加噪过程:

$$q\left(z_p^t|z_p^{t-1}\right)=\mathrm{N}\left(z_p^t;\sqrt{\alpha_t}\,z_p^{t-1},(1-\alpha_t)\boldsymbol{I}\right) \tag{3}$$

其中,$t\in[1,T]$为当前扩散的步数,α_t是扩散步长,其中$\boldsymbol{I}$是单位矩阵。这一过程z_p^0将逐步加入噪声,直到z_p^t服从标准正态分布。去噪过程则是前者的反向过程:

$$p\left(z_p^{t-1}|z_p^t,\boldsymbol{f}_t^p\right)=\mathrm{N}\left(z_p^{t-1};\mu\left(z_p^t,t,\boldsymbol{f}_t^p\right),\Sigma\left(z_p^t,t,\boldsymbol{f}_t^p\right)\right) \tag{4}$$

其中方差Σ通常被设置为单位矩阵,均值μ则由神经网络进行预测得到的均值,该网络以当前带噪声特征z_p^t、时间步t以及条件文本特征$\boldsymbol{f}_t^p$为输入,输出去噪后的均值,用于从z_p^t估计z_p^{t-1}。

最终,每个部位的局部运动编码器输出对应的动作特征$z_h, z_{la}, \cdots, z_{rl}$。这些将作为后续模块的输入。

2.4 全身运动融合模块

全身运动优化模块的主要功能是将离散的局部动作特征融合为协调的完整姿态。如图3所示,该模块首先接收来自局部生成模块的6个不同部位的

动作特征,通过特征拼接操作构建初始全局特征$z_g\in\mathbf{R}^{6\times256}$:

$$z_g=\mathrm{concat}(z_h,z_{la},\cdots,z_{rl}) \tag{5}$$

式中,concat表示特征拼接操作。此向量虽包含各部位动作信息,却尚未建立不同身体部位之间的相互关系。为解决此问题,模块引入一个两阶段的空间融合机制:第一阶段将全局向量与由TMR编码器得到的完整动作文本的语义特征联结,通过全连接层与层归一化操作生成过渡特征z_f,承载动作整体语义,确保合理的动作约束被编码。

第二阶段采用轻量级FFN网络实施空间协调优

化，其GELU（gaussian error linear units）激活函数模拟人体肌腱的非线性弹性特性，在空间域建立跨部位力学耦合，实现“摆臂带动肩部旋转”等人体部位联动效应，最终输出优化后的特征被解码为SMPL姿态参数$\hat{\theta}$。最后，在输出人体网格$\boldsymbol{M}\left(\hat{\theta},\beta\right)$时，为了增强姿态多样性，同时保持核心动作不变，可以添加微量随机噪声$\beta \sim \mathrm{N}(0,0.1)$。

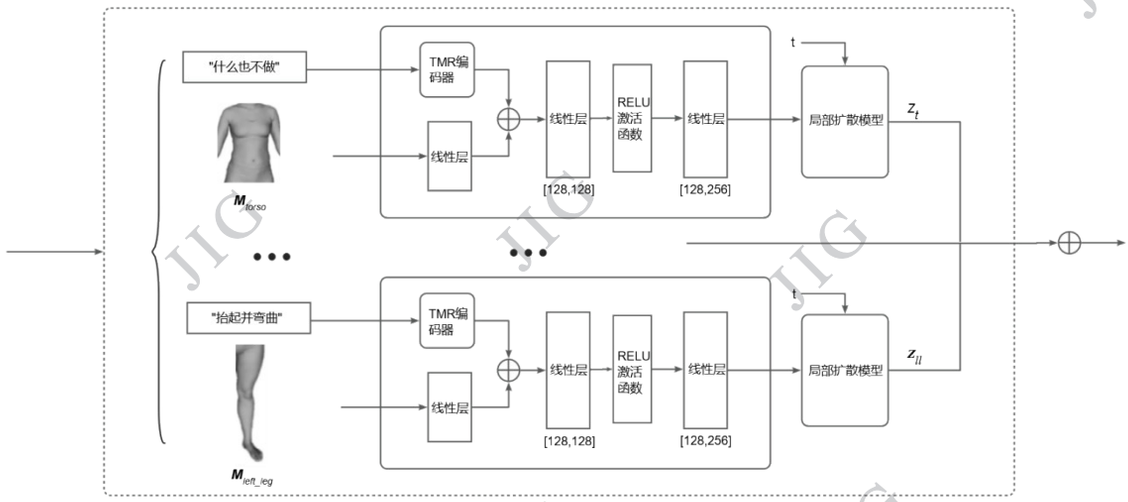


图2　局部动作生成模块

Fig. 2　The structure of the local motion generation module

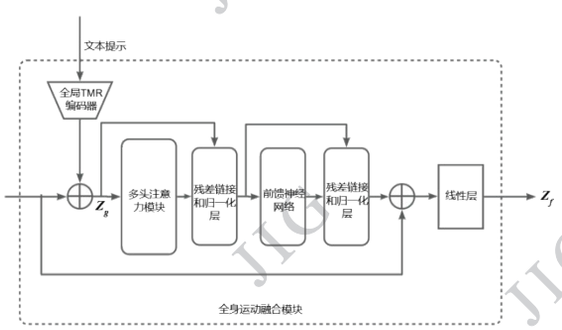


图3　全身运动融合模块

Fig. 3　The structure of the Global motion optimization module

2.5　3D表示与外观优化

3D表示与外观优化部分将参数化SMPL表示的人体网格最终转化为高保真视觉输出。如图1中3D高斯表示与外观细化部分所示，该流程首先需要进行高斯点初始化阶段：从SMPL人体网格$\boldsymbol{M}\left(\hat{\theta},\beta\right)$提取所有顶点坐标$\boldsymbol{V}\in\mathbf{R}^3$，经中心化处理$\boldsymbol{V}_{center}=\boldsymbol{V}-\bar{\boldsymbol{V}}$消除位置偏移；同时随机初始化顶点颜色$\boldsymbol{C}\sim\boldsymbol{\mathcal{U}}(0,1)^{6890\times 3}$作为材质基底。这些几何要素被参数化为3D高斯点集$\boldsymbol{\theta}$。

随后进入2D扩散引导优化阶段。在半径为3单位的球面空间均匀设置8个相机位姿$\{v_k\}_{k=1}^{8}$，通过可微分高斯溅射渲染器生成多视角图像$\boldsymbol{x}_k=g\left(\boldsymbol{\theta},v_k\right)$。以扩散模型作为2D视觉先验，计算SDS损失梯度：

$$\nabla_{\theta}\mathcal{L}_{\mathrm{SDS}}=\frac{1}{8}\sum_{k=1}^{8}E_{t,\epsilon}\left[\omega(t)\left(\widehat{\epsilon_{\phi}}\left(z_t^k;y,t\right)-\epsilon\right)\frac{\partial\boldsymbol{x}_k}{\partial\theta}\right]\tag{6}$$

式中，$\omega(t)$表示时间步t的权重函数，$\widehat{\epsilon_{\phi}}\left(z_t^k;\ y,t\right)$为2D扩散模型预测的噪声$\phi$为扩散模型参数，输入为带噪特征$z_t^k$，条件为文本$y$和时间步$t$，$\epsilon$为实际添加的高斯噪声，$\frac{\partial\boldsymbol{x}_k}{\partial\theta}$为渲染图像$\boldsymbol{x}_k$对3D高斯参数$\boldsymbol{\theta}$的雅可比矩阵。此梯度通过迭代优化3D高斯点参数$\boldsymbol{\theta}$，重点更新颜色$c_i$与透明度$\alpha_i$，使渲染结果在扩散模型的隐空间内逼近文本描述的视觉特征，尤其强化皮肤纹理褶皱、肌肉伸缩形变等细节。

3　实验结果与分析

3.1　实验设置

本文实验采用两个标准数据集：HumanML3D（Guo等，2024）和KIT Motion-Language（Plappert等，2016）作为标准数据集，并通过动作周期峰值采样和固定间隔采样策略提取关键帧，对文本描述进行时态标记移除和主语统一化等预处理。其中HumanML3D数据集包含14，616个动作序列和44，

970条文本描述，KIT-ML数据集包含3，911个动作序列和6，278条文本描述。选取的基线方法包括：GaussianDreamer（CVPR 2024）、MotionDiffuse（2024）、DreamFusion（2022）、MDM（Tevet等，2022）等。所有对比结果均基于原论文报告数据或在统一实验环境下复现得到。

本文方法在PyTorch中使用ThreeStudio（Guo等，2023）开源框架实现的。该框架集成了分数蒸馏采样SDS优化流程，并提供了统一的3D表示抽象层，使得能够高效地集成SMPL模型、3D高斯表示以及先进的2D扩散模型。对于3D高斯，不透明度α和位置μ的学习率分别为10^{-2}和5×10^{-5}。实验在配备NVIDIA GeForce RTX 4090（24GB显存）显卡的计算平台上进行，批大小统一设置为8。

3.2 评价指标

本文提出的框架能够同时生成动作与三维外观，实现了从文本到完整三维人体的端到端生成。因此，我们分别在两个维度上与对应领域的方法进行比较，评估体系包含两个维度：采用FID衡量动作质量，通过CLIP-S，即生成动作渲染帧与文本的CLIP相似度来评估语义对齐性与视觉质量。

FID通过计算生成动作与真实动作在特征空间的分布距离来评估整体质量，其计算公式为：

$$\mathrm{FID}=\left\|\mu_1-\mu_2\right\|^2+Tr\left(\Sigma_1+\Sigma_2-2\left(\Sigma_1\Sigma_2\right)^{\frac{1}{2}}\right) \tag{7}$$

式中，μ_1和μ_2分别表示真实动作和生成动作的特征均值，Σ_1和Σ_2为对应的协方差矩阵。该值越低表明生成质量越高。

为了更精细地评估模型对人体各部位动作生成的质量，本文提出了部位级FID（Part-FID）指标。该指标将传统的全局FID扩展到身体部位层面，通过计算生成动作与真实动作在每个部位特征空间中的分布距离，量化模型在局部动作生成上的表现。其计算公式与FID相同。

CLIP指标计算生成动作的多视角渲染图像与输入文本在CLIP模型隐空间中的余弦相似度，其计算公式为：

$$\mathrm{CLIP-S}=\frac{1}{K}\sum_{k=1}^{K}\frac{E_I\left(\boldsymbol{x}_k\right)\cdot E_T(T)}{\left\|E_I\left(\boldsymbol{x}_k\right)\right\|\cdot\left\|E_T(T)\right\|} \tag{8}$$

式中，E_I和E_T分别为CLIP的图像编码器和文本编码器，$\boldsymbol{x}_k$为第k个摄像机视角下渲染得到的生成动作图像，T为输入文本描述，K为视角总数（本实验K=8）。CLIP-S值越高，代表生成动作的视觉内容与文本描述的语义一致性越好。

3.3 定量分析

表1给出不同方法的CLIP评估值。为了公平比较，本文对所有方法的CLIP-S计算采用基本一致的策略：对于每个文本描述，选取120个间隔均匀的方位角摆放相机（0°～360°，仰角15°，半径为4）得到120个不同视点的图像，为保证评估的全面性和计算效率，再从120个渲染图像中随机抽取10个作为测试图像，使用OpenAI的ViT-L/14和OpenCLIP的ViT-bigG-14模型计算图像与文本的CLIP相似度，取结果的平均值作为最终得分。所有基线方法的CLIP-S值均在此统一策略下重新计算，或取自原论文中采用相似策略的数值。结果表明，本文方法在两个CLIP模型上均取得了最高分：ViT-L/14上达到29.41，ViT-bigG-14上达到44.39，显著优于其他基线方法。这验证了本文采用的视觉先验以及生成框架在提升语义对齐精度方面的有效性。

表1 不同方法的CLIP相似性定量比较

Table 1 Quantitative comparison of CLIP-S with other methods

方法	ViT-L/14↑	ViT-bigG-14↑
Shap-E（Jun等，2023）	20.51	32.21
DreamFusion（Poole等，2022）	23.60	37.46
ProlificDreamer（Wang等，2023）	27.39	42.98
Instant3D（Li等，2023）	26.87	41.77
GaussianDreamer（Yi等，2024）	27.23	41.88
本文	**29.41**	**44.39**

注：加粗字体表示最优结果，“↑”表示值越大越好。

表2评估了各算法生成动作的FID指标。本文使用数据集中文本描述作为输入，生成对应的动作。真实数据由数据集中对应的真实动作构成，生成数据分布由本文的模型生成的动作构成。其他方法FID相关的数据来自原论文中的数据。更低的FID表明该方法会产生更现实和多样的运动。需要说明的是：LGTM（Sun等，2024）是当前文本到动作生成任务的最新SOTA，但其FID值基于序列评估与本文计算方式不同，因此我们未将其在表中列出。事实

上，若将本文的静态姿态扩展为序列，其FID可能会有所变化，但这已超出本文范围。此外，LGTM专注于动作序列生成，输出仅为SMPL参数；而本文方法在生成动作的同时，进一步完成了三维重建与外观优化，实现了从文本到完整三维资产的端到端生成，任务难度更高。因此，在动作质量相近的前提下，本文提供了更丰富的输出结果，体现了方法的综合优势。

表2　不同方法在HumanML3D数据集上FID指标对比

Table 2　Quantitative comparison of FID with other methods

方法	FID ↓
MotionDiffuse（Zhang等，2024）	0.687
MDM（Tevet等，2022）	0.747
MLD（Chen等，2023）	1.753
本文	**0.429**

注：加粗字体表示最优结果，“↓”表示值越小越好

同样的，本文也在KIT-ML测试集上计算了各方法的部位级FID，结果如表3所示。所有方法的生成动作均通过对应官方模型产生，并使用统一的方法裁剪或直接提取对应的身体部位特征用于评估。相较于以往其他方法，各部位的生成精度均有明显提升，这充分验证了局部生成模块和全局融合模块在协调跨部位运动、生成符合生物力学的姿态方面的有效性。

CLIP-S和FID两个维度所选取的对比方法集合有所不同，这是由于不同方法的任务侧重点及适用评价指标不同所致。表1评估的是生成三维模型与输入文本之间的语义对齐程度。CLIP-S指标要求方法能够输出可渲染的三维模型（如网格、NeRF或3D高斯），因此本文选取了以三维模型生成为目标的代表性方法，包括DreamFusion、ProlificDreamer、Instant3D、GaussianDreamer等。这些方法在本任务中能够生成完整的三维人体外观，但通常不输出特定的动作。表2、表3（FID对比）评估的是生成动作的真实性与多样性。因此本文选取了以动作生成为目标的主流方法，包括MotionDiffuse、MDM、MLD。这些方法专注于从文本生成人体动作，通常不包含外观纹理生成模块。

在生成效率方面，本文方法在生成时间上与目前最新的先进方法GaussianDreamer相近，同为分钟级，较ProlificDreamer等基于NeRF的小时级方法具有显著效率优势。Shap-E等方法虽然生成速度快（秒级），但其生成质量低于本文方法，且通常面向通用物体生成，难以处理人体动作的精细部位控制。

表3　不同方法在KIT-ML数据集上的Part-FID指标对比

Table 3　Quantitative comparison of part-FID with other methods

方法	头部FID ↓	左臂FID ↓	右臂FID ↓	躯干FID ↓	左腿FID ↓	右腿FID ↓
MDM（Tevet等，2022）	1.63	1.57	1.59	2.41	1.75	1.80
MotionDiffuse（Zhang等，2024）	1.54	1.36	1.38	2.10	1.41	1.50
Ours	**1.32**	**1.20**	**1.24**	**1.58**	**1.09**	**1.15**

注：加粗字体表示每列最优结果，“↓”表示值越小越好。

3.4　定性分析

在本节中，从人体姿态和最终渲染结果两个维度来定性分析生成质量，并与GaussianDreamer、DreamAvatar（Cao等，2023）等方法进行对比，从而评估本方法生成运动姿态的最终视觉质量。

首先，图4左边展示了在四个较为典型的文本描述下的姿态生成结果，本方法在生成动作的局部精度和全身协调性上均展现出显著优势。在局部精度方面，对于“左腿向前踢腿”这个示例，本方法准确生成了左腿踢腿时身体的自然后仰和手部的抬起。这验证了本方法中局部动作生成模块与梯度隔离机制的有效性。

其次，图4右半部分则展示了经过3D高斯渲染优化后的最终视觉效果。本部分重点评估纹理细节和姿态合理性。在纹理细节方面，本方法表现出更好的材质真实感。特别是在手部、关节等关键部位，本方法准确生成了左腿踢腿时身体的自然后仰和手部的抬起。这验证了本方法中局部动作生成模块与梯度隔离机制的有效

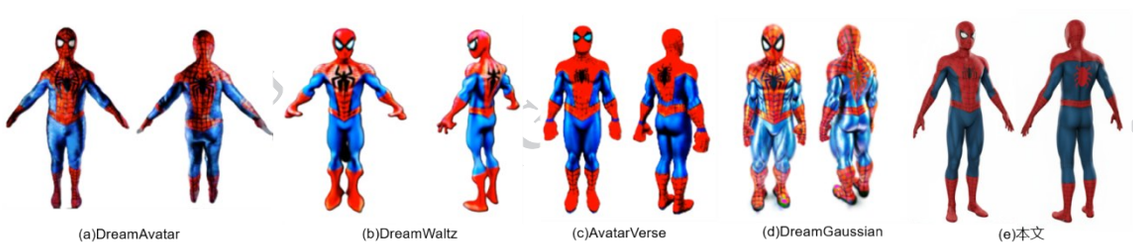
在图5中，本文展示了与DreamAvatar、Dream-

Waltz(Huang等,2023)、AvatarVerse(Zhang

((a)raise both arms(b)kick with left leg(c)bend over and spread arms(d)walk forward and swing arms)

图4　本文算法生成的SMPL姿态和带外观细节的最终渲染结果示例

Fig. 4　Generated different poses of SMPL and results generated by our method



((a)DreamAvatar(b)DreamWaltz(c)AvatarVerse(d)DreamGaussian(e)ours)

图5　生成结果定性对比

Fig. 5　Qualitative comparisons between our method and others

等,2024)和GaussianDreamer的渲染比较结果。其中DreamWaltz的结果图是从GaussianDreamer的论文中下载的,其他图均出自原论文。需要注意的是,其中只有GaussianDreamer和本文使用的文本提示词是“蜘蛛侠张开双臂站立”,而其他方法的提示词则只有“蜘蛛侠”。这是因为其他方法中没有直接的运动生成的部分,而GaussianDreamer和本文的方法则支持更具体的动作描述。与其他方法相比,由于使用了最新的2D扩散模型,本文提出的方法得到了更为精细的最终渲染效果。此外,相比之前的方法,本文的方法能够支持生成具有指定身体姿势的三维人体模型。

3.5　消融实验

本文方法创新采用局部动作生成模块与全身运动融合模块以及三维扩散模型与二维扩散模型结合的生成策略。为验证各模块的有效性并探究关键参数对生成性能的影响,本文设计了一系列消融实验。

1)局部动作生成模块与全身运动融合模块的有效性分析。为了与MDM保持一致,首先将本文提出的方法中使用的TMR文本编码器替换为CLIP,随后将生成结果与MDM中的结果进行比较从而比较局部生成模块与全身运动融合模块的效果。从表4中

可以看到，去除了局部动作生成模块与全身运动融合模块之后生成的运动质量明显下降。

表4　本文方法消融实验结果

Table 4　Ablation studies of text-to-motion generation

方法	FID ↓	CLIP-S ↑
MDM(Tevet等，2022)+ CLIP	0.731	—
DreamWaltz(Huang等，2023)	—	22.76
本文	**0.518**	**29.41**

注：加粗字体表示每列最优结果，“↓”表示值越小越好，“↑”表示值越大越好，“—”表示无法计算，数字空缺。

2)三维扩散模型与二维扩散模型结合的优势分析。如图6所示，对2D扩散模型的细节先验进行消融实验，以验证三维扩散模型与二维扩散模型结合所带来的优势。可以看到左边DreamWaltz在只使用3D扩散模型进行生成时的外观细节更为粗糙，这表明2D扩散模型提供的多视角、高质量视觉先验，能够有效丰富3D模型的表面细节和外观真实感，弥补了纯3D模型在纹理细节上的不足。对应的CLIP相似度结果对比也在表3中体现。

3)3D高斯点数量对渲染质量与效率的影响。表5展示了高斯点数量对视觉质量、渲染速度和显存占用的权衡关系。在SMPL网格顶点数6890基础上，增加高斯点可提升细节表现，但会降低渲染速度并增加显存开销。综合考虑，本文选择6890点作为默认配置。

4)SDS优化迭代次数对生成质量的影响。在2D渲染优化阶段，SDS迭代次数直接影响纹理细节的丰富程度和优化耗时。表6统计了不同迭代次数下的CLIP-S得分和耗时。随着迭代次数增加，CLIP-S评分均逐步提升，但提升幅度逐渐减小。当迭代4000次时已获得较高的视觉质量，继续增加至8000次仅带来微小增益，而优化时间增加约60%。综合考虑质量与效率，本文选择4000次作为默认配置。

图6　2D扩散模型外观细化的消融研究结果

Fig. 6　Ablation studies of the 2D diffusion models

表5　高斯点数量对渲染效率的影响

Table 5　Impact of the number of gaussians on rendering performance

高斯点数	渲染FPS	显存占用
2048	120	6 GB
4096	78	12 GB
6890(本文)	**45**	**18.5 GB**
10000	22	26 GB

注：加粗字体表示本文实验中选取的参数。

4　结　论

本文围绕文本驱动三维人体动作生成这一挑战性任务，针对现有方法在语义对齐精度、全局协调性和渲染效率方面的不足，提出了一个新颖的局部-全局协同生成框架，提出一种从文本直接到渲染之后的三维虚拟人物运动姿态的一体化方法。本研究的主要贡献体现在几个层面：首先，通过引入大型语言模型对输入文本进行人体部位的精细化解析，并设计梯度隔离的并行编码器架构，实现了对身体各部位动作的独立精准控制，提升部位级动作生成的准

表6　SDS优化迭代次数对视觉质量的影响

Table 6　Effect of SDS iteration steps on visual quality and efficiency

SDS迭代次数	CLIP-S↑	优化时间(秒)
1000	24.83	270
2000	26.47	600
4000(本文)	**29.41**	**1080**
8000	29.58	1800

注:加粗字体表示本文实验中选取的参数。

确率,有效解决了语义泄漏问题;其次,通过建立SMPL参数化模型、3D高斯表示与2D扩散模型引导优化的协同流水线,创新性实现了文本到人体动作的端到端生成,极大增强了方法的实用性。实验结果表明,本方法在客观指标和主观视觉质量上均显著优于主流基线模型,有效验证了框架的先进性。

尽管本研究取得了预期成果,但仍存在若干局限性亟待未来探索。在复杂长序列动作生成方面,当前模型对包含时序逻辑与物体交互的复杂描述处理能力有限,未来将引入更强时序建模模块并融合场景上下文信息。在多模态控制与交互层面,需要突破文本单模态限制,探索融合语音、图像等多模态控制信号,向交互式生成方向发展。针对极端条件下的稳定性问题,模型在极端摄像机视角下会出现渲染瑕疵,后续将通过引入显式几何约束或多视角一致性损失增强鲁棒性。此外,当前模型对医疗康复等专业领域的泛化能力有待验证,未来将探索基于小样本学习的技术以适应特定垂直领域需求。综上所述,本文提出的框架不仅在学术上启发了基于语义解耦的生成范式,更为虚拟人、元宇宙等应用领域提供了高效可靠的技术工具,具有重要的理论意义和实践价值。

参考文献(References)

Ahuja C and Morency L P. 2019. Language2pose: Natural Language Grounded Pose Forecasting. In 2019 International Conference on 3D Vision (3DV). IEEE, 719 - 728.[DOI: 10.1109/3DV.2019.00084]

Athanasiou N, Petrovich M, Black M J, and Varol G. 2022. TEACH: Temporal Action Composition for 3D Humans. In 2022 International Conference on 3D Vision (3DV). IEEE Computer Society, 414 - 423.[DOI: 10.1109/ 3DV57658.2022.00053]

Barron J T, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, and Srinivasan P P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE:5835-5844[DOI: 10.1109/ICCV48922.2021.00580]

Black Forest Labs, Batifol S, Blattmann A, Boesel F, Consul S, Diagne C, et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space.[EB/OL].[2025-06-24].
https://arxiv.org/pdf/2506.15742

Cao Y, Cao Y, Han K, Shan Y and Wong K K. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 958-968

Chen R, Chen Y, Jiao N, Jia K. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 22189-22199, [DOI: 10.1109/ICCV51070.2023.02033]

Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, et al. 2023a. Executing Your Commands via Motion Diffusion in Latent Space// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 18000-18010.[DOI: 10.1109/CVPR52729.2023.01726]

Dabral R, Mughal M H, Golyanik V, and Theobalt C. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 9760-9770.[DOI: 10.1109/CVPR52729.2023.00941]

Ghosh A, Cheema M, Oguz C, Theobalt C, and Slusallek P. 2021. Synthesis of compositional animations from textual descriptions[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1396-1406

Gupta A, Xiong W, Nie Y, Jones I and Oğuz B. 2023. 3dgen: Triplane latent diffusion for textured mesh generation.[EB/OL].[2023-05-09].
https://arxiv.org/pdf/2303.05371

Guo C, Zou S, Zuo X, Wang S, Ji, Li X, et al. 2022. Generating Diverse and Natural 3D Human Motions From Text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5152 - 5161.

Guo Y, Liu Y, Shao R, Laforte C, Voleti V, Luo G, et al. 2023. threestudio: A unified framework for 3d content generation.[EB/OL]
https://github.com/threestudio-project/ threestudio

Hang T, Rui C, Shipeng A, Hao W and Heng Z. 2023. The growth of image-related three dimensional reconstruction techniques in deep learning-driven era: a critical summary. Journal of Image and Graphics. 28(08):2396-2409 (杨航, 陈瑞, 安仕鹏, 魏豪, 张衡. 2023. 深度学习背景下的图像三维重建技术进展综述. 中国

图象图形学报，28(08)：2396-2409)[DOI：10.11834/jig.220376]

Huang Y，Wang J，Zeng A，Cao H，Qi X，Shi Y，et al. 2023. Dreamwaltz：Make a scene with complex 3d animatable avatars. Advances in Neural Information Processing Systems，2023，36：4566-4584.

Hu S，Hong F，Hu T，Pan L，Mei H，Xiao W，Yang L，and Liu Z. Humanliff：Layer-wise 3d human generation with diffusion model.[EB/OL].[2023-08-18].
https:// arxiv/pdf/2308.09712.pdf

Jun G，Tianchang S，Zian W，Wenzheng C，Kangxue Y，Daiqing L，et al. 2022. Get3d：A generative model of high quality 3d textured shapes learned from images. Advances in neural information processing systems，35，31841-31854.[DOI：10.48550/arXiv.2209.11163]

Jun H and Nichol A. 2023. Shap-e：Generating conditional 3d implicit functions[EB/OL].[2023-05-03].
https://arxiv.org/pdf/1207.0580.pdf

Kerbl B，Kopanas G，Leimkühler T and Drettakis G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics，42(4)，pp.1-14.[DOI:10.1145/3592433].

Li J，Tan H，Zhang K，Xu Z，Luan F，Xu Y，Hong Y，et al. 2023 Instant3d：Fast text-to-3d with sparse-view generation and large reconstruction model//The Twelfth International Conference on Learning Representations.

Li W，Chen R，Chen X and Tan P. 2023. Sweetdreamer：Aligning geometric priors in 2d diffusion for consistent text-to-3d.[EB/OL].[2023-10-20].
https://arxiv.org/pdf/2310.02596.pdf

Loper M，Mahmood N，Romero J，Pons-Moll G，and Black M J. 2023. SMPL：A Skinned Multi-Person Linear Model. Seminal Graphics Papers：Pushing the Boundaries，Volume 2(1st ed.. Association for Computing Machinery，New York，NY，USA，Article 88，851–866.[DOI:10.1145/3596711.3596800]

Mildenhall B，Srinivasan P P.，Tancik M，Barron J T，Ramamoorthi R，and Ng R. 2022. NeRF：representing scenes as neural radiance fields for view synthesis. Communications of the ACM，65(1)：99-106.[DOI：10.1145/3503250]

Mingtao F，Junhao S，Zijie W，Weixing P，Hang Z，Yulan G，et al. 2025. Advancements in 3D vision understanding using multimodal large language models. Journal of Image and Graphics，30(6)：1744-1791(冯明涛，沈军豪，武子杰，彭伟星，钟杭，郭裕兰，等. 2025. 多模态大模型驱动的三维视觉理解技术前沿进展. 中国图象图形学报，30(6)：1744-1791)[DOI：10.11834/jig.240588]

Müller T，Evans A，Schied C，and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. 41，4，Article 102(July 2022)，15 pages.[DOI：10.1145/3528223.3530127]

Nichol A，Jun H，Dhariwal P，Mishkin P and Chen M. 2022. Point-e：A system for generating 3d point clouds from complex prompts.[EB/OL].[2022-12-16].
https://arxiv.org/pdf/2212.08751.pdf

Petrovich M，Black M J，and Varol G. 2022. TEMOS：Generating Diverse Human Motions from Textual Descriptions. In Computer Vision – ECCV 2022：17th European Conference，Tel Aviv，Israel，October 23 – 27，2022，Proceedings，Part XXII. Springer-Verlag，Berlin，Heidelberg，480 – 497.[DOI：10.1007/978-3-03120047-2_28]

Petrovich M，Black M J，and Varol G. 2023. TMR：Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. International Conference on Computer Vision，Oct 2023，Paris，France.[DOI：10.1109/iccv51070.2023.00870]

Plappert M，Mandery C and Asfour T. 2016. The KIT Motion-Language Dataset. Big data，2016，4(4)：236-252.[DOI：10.1089/big.2016.0028]

Podell D，English Z，Lacey K，Blattmann A，Dockhorn T，Müller J，et al. 2023. Sdxl：Improving latent diffusion models for high-resolution image synthesis//The Twelfth International Conference on Learning Representations.

Poole B，Jain A，Barron J T，and Mildenhall B. 2022. Dreamfusion：Text-to-3d using 2d diffusion.[EB/OL].[2022-09-29].
https://arxiv.org/pdf/2209.14988.pdf

Raab S，Leibovitch I，Tevet G，Arar M，Bermano A H，and Cohen-Or D. 2023. Single motion diffusion.[EB/OL].[2023-06-13].
https://doi.org/10.48550/arXiv.2302.05905

Radford A，Kim J W，Hallacy C，Ramesh A，Goh G，Agarwal S，et al. 2021. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR，2021：8748-8763.

Rombach R，Blattmann A，Lorenz D，Esser P，and Ommer B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，New Orleans，LA，USA，10674 – 10685.[DOI：10.1109/CVPR52688.2022.01042]

Sun H，Zheng R，Huang G，Ma C，Huang H，and Hu R. 2024. LGTM：Local-to-Global Text-Driven Human Motion Diffusion Model. In ACM SIGGRAPH 2024 Conference Papers (SIGGRAPH '24). Association for Computing Machinery，YorkNew，NY，USA，Article 66，1 – 9.[DOI：10.1145/3641519.3657422]

Tevet G，Raab S，Gordon B，Shafir Y，Cohen-Or，D and Bermano A. H. 2022 .Human motion diffusion model.[EB/OL].[2022-09-29]
https://arxiv.org/pdf/2209.14916.pdf

Tevet G，Gordon B，Hertz A，Bermano A H，and Cohen-Or D. 2022a. MotionCLIP：Exposing Human Motion Generation to CLIP Space//European Conference on Computer Vision. Cham：Springer Nature Switzerland，2022：358-374.

Wang H，Du X，Li，J，Raymond A and Yeh G S. 2022. Score Jacobian Chaining：Lifting Pretrained 2D Diffusion Models for 3D Genera-

tion//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 12619-12629. [DOI: 10.1109/CVPR52729.2023.01214]

Wang Z, Lu C, Wang Y, Bao F, Li C, Su H, et al. 2023 Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. [EB/OL].[2023-05-25]. https://arxiv.org/pdf/2305.16213.pdf

Yi T, Fang J, Wang J, Wu G, Xie L, Zhang X, Liu W, et al. 2024 Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 6796-6807.[DOI: 10.1109/CVPR52733.2024.00649]

Zhang H, Chen B, Yang H, Qu L, Wang X, Chen L, et al. 2024. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 7, pp. 7124-7132).[DOI: 10.1609/aaai.v38i7.28540]

Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L, et al. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence, 46(6), 4115-4128. [DOI: 10.1109/TPAMI.2024.3355414]