层次化文本语义驱动的多粒度人体行为生成

舒祥波[1*],李成建[1],尹 政[1],李朋鹏[1],李泽超[1],唐金辉[2]
(1. 南京理工大学计算机科学与工程学院,江苏南京 210094;
2. 南京林业大学信息科学与技术学院&人工智能学院,江苏南京 210037)

摘 要: 当前的人体行为生成方法在生成文本描述与行为一致的高质量运动方面仍面临挑战。尽管近年来基于扩散模型、自回归模型以及多模态预训练模型的方法在运动自然性和多样性上取得了一定进展,但在复杂文本语义理解和精细动作建模方面仍存在明显不足。其主要原因包括:(1) 缺乏句子成分间层次依赖关系建模会导致模型文本语义理解困难;(2) 现有方法仅在全局级或单词级进行文本-行为之间跨模态对齐,忽视了全局与局部信息之间的互补性导致粗细粒度协同建模困难。为此,本文提出了一种层次化文本语义驱动的多粒度人体行为生成框架(Hierarchical Textual-semantic-driven Multi-Granularity human motion generation framework,HTMG),该框架在全面理解文本语义的同时实现了粗细粒度的跨模态交互,从而实现文本-行为的一致性。具体而言,为了解决文本语义理解难题,本文提出了一种层次化语义捕捉策略(Hierarchical Semantic Capture Strategy,HSCS),该策略通过句法分析构建文本结构树显式建模单词间依存关系并引入双曲图注意力机制(Hyperbolic Graph ATtention mechanism,HGAT)在双曲空间动态捕捉层次语义依赖,从而显著提升模型的语义理解能力。此外,为了实现粗细粒度的跨模态对齐,本文设计一种多粒度跨模态注意力机制(Multi-Granularity Cross-modal Attention mechanism,MGCA),通过将全局级语义表示与单词级局部语义表示分别与人体行为特征进行自适应交叉融合,使模型在生成过程中能够同时关注整体动作意图与局部动作变化,从而实现语义一致的多粒度动作建模。大量实验结果表明,本文提出的HTMG在HumanML3D和KIT-ML数据集上均取得了最优性能,充分验证了该框架在文本语义理解与文本-行为一致性建模方面的有效性。

关键词: 人体行为生成;层次化语义捕捉策略;双曲空间;双曲图注意力机制;文本结构树;多粒度跨模态注意力机制

基金项目: 国家自然科学基金国家重大科研仪器研制项目(No.62427808);国家自然科学基金区域创新发展联合基金(No.U25A20442);国家自然科学基金优秀青年科学基金(No.62222207);国家杰出青年科学基金(No.62425603)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0451-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251089

Hierarchical Text Semantics-Driven Multi-Granularity Human Motion Generation

SHU Xiangbo[1*], LI Chengjian[1], YIN Zheng[1], LI Pengpeng[1], LI Zechao[1], TANG Jinhui[2]
(1. *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu* 210094, *China*;
2. *College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing, Jiangsu* 210037, *China*)

Abstract: Generating high-quality human motions that are semantically consistent with textual descriptions remains a challenging problem. Although recent diffusion-based, autoregressive, and multimodal pre-trained approaches have improved motion naturalness and diversity, they still struggle with complex semantic understanding and fine-grained motion modeling. These limitations mainly stem from two factors: (1) the lack of explicit modeling of hierarchical dependency relationships among sentence components, which hampers accurate textual semantic understanding; (2) the reliance on either global-level or word-level text-motion alignment, while neglecting the complementarity between global and local semantics, making coarse-to-fine collaborative modeling difficult. To address these limits, we propose the hierarchical textual-semantic-driven multi-granularity human motion generation framework (HTMG), which models textual semantics while enabling coarse-to-fine cross-modal interactions to ensure text-motion consistency. Specifically, we introduce a hierarchical semantic capture strategy (HSCS) that constructs a textual structure tree via syntactic parsing and embeds it into hyperbolic space, where hierarchical semantic dependencies are dynamically modeled using a hyperbolic graph attention mechanism. Furthermore, we design a multi-granularity cross-modal attention mechanism (MGCA) that adaptively fuses global-level and word-level semantic representations with motion features, allowing the model to jointly capture overall motion intent

收稿日期:2025-12-13;录用日期:2026-01-06;责任编委:任文琦
*通信作者:舒祥波

and fine-grained action variations. Extensive experiments demonstrate that HTMG achieves state-of-the-art performance on the HumanML3D and KIT-ML benchmarks, validating the effectiveness of our framework in textual semantic understanding and text-motion alignment.

Keywords: human motion generation; hierarchical semantic capturing strategy; hyperbolic space; hyperbolic graph attention mechanism; textual structure tree; multi-granularity cross-modal attention

Foundation Item(s): National Natural Science Foundation of China Major Research Instrument Development Program (No. 62427808) ; National Natural Science Foundation of China Regional Innovation Development Joint Fund (No. U25A20442); National Natural Science Foundation of China Excellent Young Scientists Fund (No. 62222207); National Science Fund for Distinguished Young Scholars (No.62425603)

0 引言

文本驱动的人体行为生成旨在根据文本指令合成高质量的人体动作。它广泛应用于动画制作、游戏开发、人机交互和虚拟现实等领域[1-5]。由于从头开始创建具有各种运动的动态场景既昂贵又耗时,文本到动作生成方法通过将自然语言界面集成到动作创建中提供了一种有效的解决方案。尽管具有潜力,但由于人类运动动力学的复杂结构,该任务仍面临巨大挑战[6]。为此,研究者们提出了自编码器(AutoEncoder, AE)[7-8]、生成对抗网络(Generative Adversarial Network, GAN)[9-10]和自回归模型(Autoregressive Model, AR)[3,11-14]等方法以实现自动化行为生成。这些早期方法主要基于直接生成和序列建模。在动作生成过程中,它们通常面临信息丢失、训练复杂以及误差累积等挑战,从而严重限制了合成动作的质量和多样性[4]。近年来,基于扩散模型(Denoising Diffusion Probabilistic Model, DDPM)[1,15-21]的方法以逐步去噪的概率生成机制在非自回归框架下实现稳定行为生成,显著提升了生成结果的保真度与多样性。

尽管上述方法取得了显著成就,但这些工作在生成文本描述一致的高质量运动过程中仍存在两个关键挑战。(1) 文本语义理解困难。现有的行为生成方法依赖CLIP(Contrastive Language-Image Pre-training)[22]模型在欧氏空间中提取句子级和单词级语义(如图1①所示),这些方法缺乏基于句法结构的层级关系建模,导致文本语义理解不充分。例如,在指令“A man runs to the right then runs to the left then back to the middle”中,句法结构可分为主语“man”(动作主体)、谓语动词“runs”(核心动作)以及状语“right”和“middle”(动作方向)等层级。虽然现有方法能够生成谓语动词对应的核心动作,但对主语和状语等辅助成分之间的依存和层级关系缺乏建模,导致行为结束位置错误。(2)粗细粒度协同建模困难。现有方法通常仅使用句子级语义或单词级语义与行为特征进行跨模态对齐,缺乏全局与局部语义的协同导致粗细粒度行为建模困难(如图1①所示)。具体而言,当仅采用全局句子级语义指导时,粗粒度动作(如“runs”)能够有效生成,但动作方向(如“middle”等细节缺乏建模;反之,仅采用局部单词级语义引导虽能刻画细粒度特征,却往往导致核心动作失真,难以保证运动一致性。

为了解决上述挑战,本文提出一种层次化文本语义驱动的多粒度人体行为生成框架(Hierarchical Text Semantics-Driven Multi-Granularity Human Motion Generation, HTMG),该框架在全面理解文本语义的同时实现了粗细粒度的跨模态交互,从而提升文本-行为的一致性(如图1②所示)。针对挑战(1),本文提出了层次化语义捕捉策略(Hierarchical Semantic Capturing Strategy, HSCS)捕获层次化文本语义增强模型对文本描述的全面理解。具体来说,HSCS首先采用句法依赖分析[23]建模单词间依存关系构建文本结构树,使各节点可基于其依存关系进行信息传递。随后,受到双曲空间在建模树形结构层次关系方面的天然优势的启发[24-26],本文利用双曲图注意力网络(Hyperbolic Graph Attention Network, HGAT)将文本结构树节点映射至双曲空间,动态聚合以动词为核心、状语和副词为辅助的层次依赖关系,从而提升模型文本理解能力。

针对挑战(2),本文提出了多粒度跨模态注意力机制(Multi-Granularity Cross-modal Attention mechanism, MGCA),通过句子级和单词级语义与行为特征跨模态对齐实现语义一致的多粒度动作建模。具体而言,MGCA首先利用HSCS捕获的层次化文本语义(如句子级与单词级)分别与行为特征进行交叉注意力建模,从而实现文本与行为在粗粒度和细粒度层面的语义对齐。随后,MGCA通过动量更新机制对粗细粒度特征进行自适应融合,从而进一步提升文本与行为之间的语义一致性。此外,如图1③所示,HTMG在HumanML3D数据集上的R Top 1和FID获得最优结果,验证了所提方法能够生成文本描述一致的高质量运动。

本研究的主要研究贡献如下:

(1)粗细粒度行为生成框架。本文提出了一种多粒度行为生成框架,通过层次化语义捕捉策略和多粒度跨模态注意力实现层次化文本语义与人体行为的跨

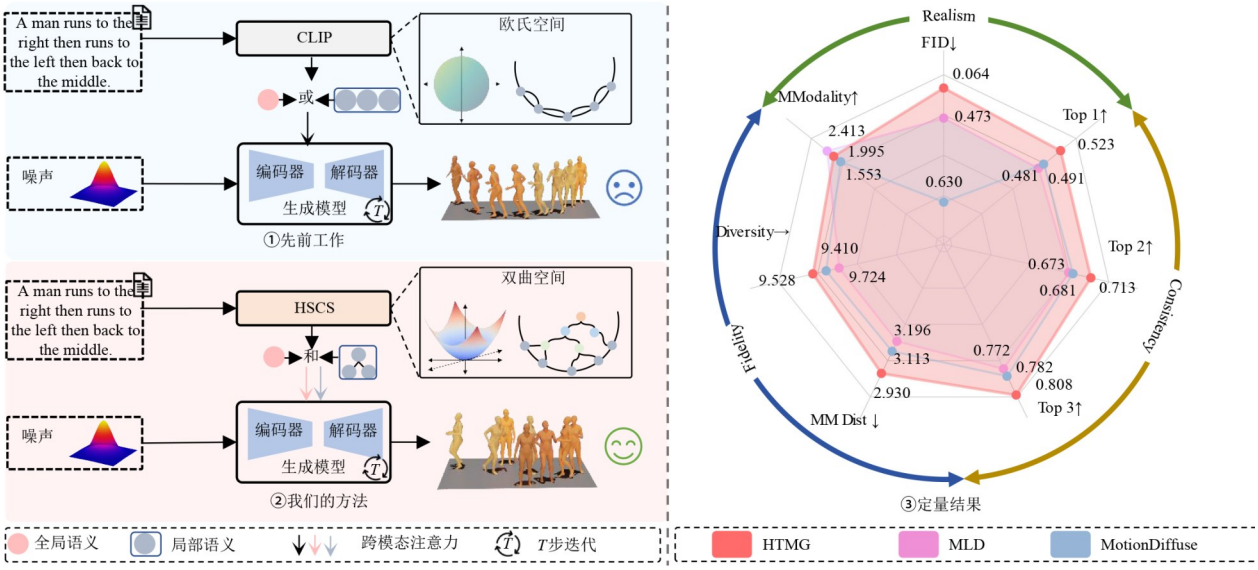

图 1　HTMG 的核心思路及其在 HumanML3D 数据集上的性能

Figure 1　Overview of HTMG and its performance on the HumanML3D dataset

模态对齐,从而生成与文本描述对应的粗细粒度动作。

(2)层次化语义捕捉策略。本文提出了层次化语义捕捉策略,通过对文本句法分析构建文本结构树并映射至双曲空间,利用双曲注意力网络对句子成分中语义关系进行层次化建模,从而增强模型对文本语义的全面理解。

(3)多粒度跨模态注意力机制。本文提出了多粒度跨模态注意力机制,将层次化文本语义与行为特征进行交叉注意力建模,并通过动量更新机制自适应融合粗细粒度特征,从而实现跨模态语义的一致对齐。

1　相关工作

1.1　文本驱动的人体行为生成

文本驱动的人体行为生成旨在生成符合文本指令的自然且高质量的人体行为。根据采用技术的不同分为:基于自编码器的方法[7-8]、基于生成对抗网络的方法[9-10]、基于自回归模型的方法[3,11-14]和基于扩散模型的方法[1,15-21]。

在研究初期,基于 AE 的方法通过将高维动作序列压缩至低维隐空间,实现了对动作数据的鲁棒表示,因此被广泛应用于人体动作生成。例如,TEMOS(Generating diverse human motions from textual descriptions)[7]将 Transformer[27]架构与变分自编码器(Variational Autoencoder,VAE)结合,利用文本编码器生成与描述相对应的动作序列。然而,这类方法在压缩过程中容易丢失文本和细粒度行为信息,导致生成动作质量受限。

为弥补这一不足,基于 GAN 的方法引入了对抗博弈机制,通过生成器与判别器的相互竞争提升动作生成的真实感与多样性。例如,HP-GAN(Probabilistic 3D human motion prediction via GAN)[9]采用改进的 WGAN-GP(Wasserstein generative adversarial network)[28]训练框架,并设计了以先前姿态为条件的运动预测损失函数,以约束未来动作的生成。然而,在条件生成任务中,GAN 的训练过程通常较为复杂,需要精心设计的对抗策略以保持模型稳定性。

为了保证生成过程的稳定性,基于自回归的模型通过在时间维度上逐帧生成未来动作来显式建模时序依赖关系。这种逐步生成机制能够捕捉长时依赖并生成动态连贯的动作序列。例如,TM2D(Bimodality Driven 3D Dance Generation via Music-Text Integration)[14]和 Att-T2M(Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism)[11]在 VQ-VAE(Vector Quantized Variational Autoencoder)[29]的基础上提出了身体部位的时空编码器,以增强离散隐空间的学习,提高模型对行为的表达能力。然而,自回归模型由于生成过程是串行的,容易在长序列生成中出现误差累积,进而影响整体动作质量。

近年来,基于扩散模型的方法在人体动作生成中展现出强大的潜力。其通过全局反向去噪逐步重建完整动作序列,每步基于全局状态更新,从而有效缓解自回归模型的误差累积问题。例如,MotionDiffuse(Text-Driven Human Motion Generation with Diffusion Model)[16]通过条件扩散学习概率文本-动作映射,在文本生成行为任务中取得了优异性能。StableMoFusion(Towards Robust and Efficient Diffusion-based Motion Generation Framework)[1]则基于 DiT 对齐文本与动作特征,通过地面反作用力约束缓解脚滑问题。然

而,这类方法在复杂句法-运动学关联的细粒度建模方面仍存在不足,难以生成语义精准的动作。

针对上述问题,本文基于扩散模型提出了层次化语义捕捉策略,通过对文本句法分析构建文本结构树并映射至双曲空间,对句子成分中语义关系进行层次化建模,从而增强模型对文本语义的全面理解。同时,为实现粗细粒度的跨模态对齐,设计了多粒度跨模态注意力机制,将层次化文本语义与行为特征进行交叉注意力建模,并通过动量更新机制自适应融合粗细粒度特征,从而实现跨模态语义的一致对齐。

1.2 基于双曲空间的表征学习

双曲空间是一种具有恒定负曲率的几何空间[30]。因其指数级的空间容量,使其能以更低的维度高效地表示和嵌入具有天然层次结构和树状拓扑的复杂网络数据,因此被应用于自然语言处理领域(Natural Language Processing, NLP)[25,31]、计算机视觉(Computer Vision, CV)[26,32-33]和大语言模型(Large Language Model, LLM)[34-35]等领域。例如,HGCN(Hyperbolic Graph Convolutional Neural Networks)[25]提出了一种双曲图卷积神经网络,充分利用双曲空间的层次化建模特性,有效缓解了欧氏空间中图神经网络易产生的距离失真问题,从而显著提升了线性预测与节点分类任务的性能。HyCoCLIP(Hyperbolic Compositional CLIP)[32]提出一种基于双曲空间的新型CLIP框架,实现了文本与图像特征[36-37]的高效对齐,并在语义分割和图像识别等任务中取得了优异表现。HELM(Hyperbolic Large Language Models)[24]构建了一个由双曲注意力、双曲归一化以及双曲混合曲率专家模块组成的大语言模型,有效弥补了LLAMA(Open and Efficient Foundation Language Models)[38]和DeepSeek-V3[39]等模型在语言层次信息建模方面的不足。

2 方法

HTMG的整体概览如图2所示。首先,本文通过扩散模型的加噪过程将行为序列x_0映射为噪声分布x_T。随后,在去噪阶段,x_T通过层次化语义融合模块(Hierarchical Semantic Fusion module, HSF)组成的UNet架构进行噪声预测与去除,从而生成与文本指令一致的动作序列$\hat{x}_0$。其中,HSF由HSCS和MGCA组成。HSCS通过对文本句法分析构建文本结构树并映射至双曲空间,并利用双曲图注意力网络对句子成分中语义关系进行层次化建模,从而增强模型对文本语义的全面理解;MGCA将层次化文本语义与行为特征进行交叉注意力建模,并通过动量更新机制自适应融合粗细粒度特征,实现跨模态语义的一致对齐。

特别地,陀螺向量空间中的运算定义在一个开放

2.1 基础知识

扩散模型[40]是一种通过逐步向数据添加噪声并学习逆过程以去除噪声,从而实现从随机噪声生成真实数据的概率生成模型。主要包含三个相互关联的过程:正向扩散、反向去噪和推理生成过程。具体来说,首先正向扩散过程通过将噪声逐步扩散到真实运动x_0中,使得x_0经过T步扩散后变为服从高斯分布的噪声运动x_T,如式(1)所示:

$$q(x_t|x_{t-1})=\mathcal{N}\left(x_t;\ \sqrt{1-\beta_t}\,x_{t-1},\beta_t\boldsymbol{I}\right) \tag{1}$$

其中,β_t控制噪声注入强度;$\boldsymbol{I}$为单位矩阵。随后,反向去噪过程训练一个参数化神经网络$\varepsilon_\theta(x_t,t)$,学习在每个时间步上预测噪声成分,从而反推原始运动分布。如式(2)所示:

$$p_\theta(x_{t-1}|x_t)=\mathcal{N}\left(x_{t-1};\ \mu_\theta(x_t,t),\ \sum\nolimits_\theta(x_t,t)\right) \tag{2}$$

其中,$\mu_\theta(\cdot)$表示模型在反向扩散过程中预测的下一个时间步样本的均值,决定去噪的方向;$\sum_\theta(\cdot)$表示模型预测的协方差或噪声方差,控制生成过程中的随机性。为了提高噪声预测的准确性,训练阶段采用均方误差(MSE)作为损失函数:$L_{\mathrm{MSE}}=E_{x_0,\varepsilon,t}\left[\left\|\varepsilon-\varepsilon_\theta(x_t,t)\right\|_2^2\right]$。最后,推理生成过程从标准高斯噪声$x_T\sim\mathcal{N}(0,\boldsymbol{I})$开始,利用训练好的去噪网络$\varepsilon_\theta(\cdot)$逐步进行反向采样,从而生成最终的动作序列$\hat{x}_0$。

双曲空间。双曲空间是具有恒定负曲率K的黎曼流形,其关键特性是其“空间扩张速度”远快于欧几里得空间[31]。具体而言,考虑一个常曲率$K=-1$的二维双曲空间中的圆盘:当其双曲半径为r时,其周长与面积分别为$2\pi\sin r$和$2\pi(\cos r-1)$,两者都随着r呈指数级增长(即约为e^r)。相比之下,在二维欧几里得空间中,半径为r的圆的周长和面积分别为$2\pi r$和πr^2,它们仅随r作线性与二次方增长。正因为这种指数型扩张特性,一些研究者发现,双曲空间可能是具有层次化结构和幂律分布图的内在几何空间[41]。因此,许多具有层次结构和幂律分布特征的真实世界图(如社交网络、生物网络等)非常适合在双曲空间中进行建模[42]。

陀螺矢量空间。在欧几里得空间中,向量空间构成了几何的代数形式,使研究者能够执行诸如向量加法、减法和标量乘法等常见运算,并据此设计出丰富的算法。然而,这些运算无法直接在双曲空间进行应用。为此,研究者[43]提出在陀螺矢量空间为双曲几何建立对应的代数形式,使得诸如“向量加法”和“标量乘法”等运算可以在双曲空间中自然定义,从而为在双曲空间中设计算法提供了理论基础。

的d维球内,如式(3)所示:

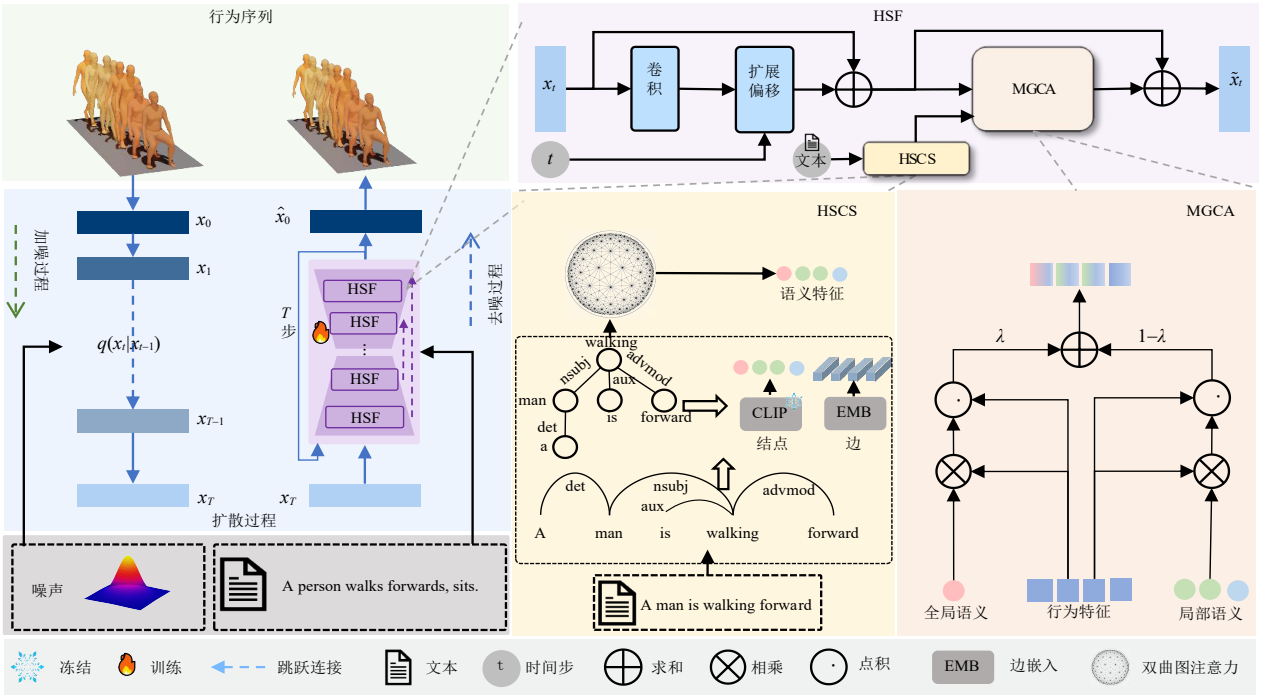

图2　HTMG模型架构图

Figure 2　Architecture of the HTMG model

$$D_c^d=\left\{x\in\mathbb{R}^d:c\|x\|^2<1\right\}\tag{3}$$

其中,$c>0$表示与曲率相关的缩放参数。若$c=0$,则$D_0^d=\mathbb{R}^d$,此时该空间退化为标准的欧氏空间;若$c>0$,D_c^d表示半径为$1/\sqrt{c}$的球;若$c=1$,则退化为常用的d维开球D^d。所有陀螺运算均在该开球内部执行,从而保证了双曲几何中的代数运算具有良好的封闭性和一致性。

2.2　层次化语义捕捉策略

以文本驱动的人体行为生成方法[1-2]通常依赖基于Transformer的CLIP在欧氏空间中提取句子级与单词级语义,但受其线性特性限制,难以建模句法结构中的层级关系,导致对主语、谓语与状语等成分间依存关系的理解不足,从而造成模型文本语义理解困难的问题。为此,本文提出了层次化语义捕捉策略(HSCS),通过句法依赖分析构建文本结构树,并利用双曲图注意力网络在双曲空间中动态建模以动词为核心的层次依赖关系,从而增强文本语义理解能力(如图2所示)。具体来说,本文首先通过spaCy工具对文本描述f^{text}进行句法依赖分析,以获取单词的词性及其依存关系,实现对句子层次结构的建模。随后,根据分析结果将单词映射为节点、依存关系映射为边,构建文本结构树即一种带有层次关系的图,以注入句法层面的先验拓扑信息。此时,文本结构树的欧氏空间图定义为$G_E=\{W_E,\xi\}$,其中E表示欧氏空间,W_E为CLIP提取单词特征,ξ表示节点间依存关系,满足式(4):

$$\xi_{ij}=\begin{cases}1, & \text{节点}i\text{与节点}j\text{存在依存关系}\\ 0, & \text{节点}i\text{与节点}j\text{无关系}\end{cases}\tag{4}$$

此外,由于依存关系的类别数量固定,本文对每种依存关系采用One-hot编码生成对应标签,并将其输入Embedding层以获得向量化表示。其描述过程如式(5)所示:

$$v_e=\beta_i\left(\varphi_e\left(F_{\text{onehot}}(\xi_i)\right)\right)\tag{5}$$

其中,$v_e\in\mathbb{R}^{D_E}$,D_E表示边特征维度;φ_e表示Embedding层;β_i是针对每个ξ_i学习的自适应权重参数,i对应一种特定的词间依赖关系(如主谓关系、动宾关系等)。由于每个单词可能与多个节点相邻,且不同节点对语义的贡献不同,因此在文本特征提取过程中需要加以区分。为了捕获文本结构树中节点间层次关系和贡献,传统方法采用GCN和GAT方法进行节点特征信息聚合实现信息传递。虽然取得了一定的效果,但是由于其欧氏空间建模特性在处理层次化语义信息时存在信息扭曲和丢失问题导致层次化建模能力不足[25]。

为解决上述问题,受HGCN利用双曲空间建模层次化关系以缓解信息丢失的启发,本文引入一种双曲图注意力网络(Hyperbolic Graph Attention Network,HGAT)用于在捕获语义层次结构的同时自适应衡量

不同单词对行为生成的贡献程度(如算法1所示)。具体来说,文本树节点特征范数可能超出陀螺向量空间中定义的开球范围。为了使节点特征能够在双曲空间中使用,本文首先采用指数映射将节点特征 W_i^E 从欧氏空间投影到双曲空间 $W_i^c \in T_x D_c^d$,其中 x 为双曲空间中的一点,$T_x D_c^d$ 表示该点的切空间。指数映射 $\exp_x^c: T_x D_c^d \to D_c^d$,当 $x \neq 0$ 时,定义如式(6)所示:

$$\exp_x^c(W_i^E)=x\oplus_c\left(\tanh\left(\sqrt{c}\,\frac{\lambda_x^c\|W_i^E\|}{2}\right)\right)\frac{W_i^E}{\sqrt{c}\,\|W_i^E\|} \tag{6}$$

当 $x=0$,指数映射定义如式(7)所示:

$$\exp_0^c(W_i^E)=\tanh\left(\sqrt{c}\,\|W_i^E\|\right)\frac{W_i^E}{\sqrt{c}\,\|W_i^E\|} \tag{7}$$

其中,$\oplus_c$ 表示莫比乌斯加法,$\lambda_x^c=\dfrac{2}{1-c\|x\|^2}$ 表示共形因子,主要用于计算欧氏空间特征映射到双曲空间时角度不变长度缩放大小关系。c 表示曲率大小,$\|x\|$ 是点 x 到原点的欧氏距离。此处,假设特征 W_i^E 位于点 $x=0$ 处的切线空间,因此可通过指数映射 $\boldsymbol{p}_i=\exp_0^c(W_i)$ 将其投射到双曲空间,得到双曲空间特征 $\boldsymbol{p}_i \in D_c^d$ 随后,为进一步获取更高层次、更具表达力的潜在表示,本文将向量 $\boldsymbol{p}_i$ 映射至新的表示空间 h_i。为此,我们采用由权重矩阵 $\boldsymbol{M}$ 数化的共享线性变换,并在双曲空间中通过莫比乌斯矩阵-向量乘法实现映射。相应地,h_i 的计算如式(8)所示:

$$h_i=\boldsymbol{M}\otimes_c\boldsymbol{p}_i=\frac{1}{\sqrt{c}}\tanh\left(\frac{\|\boldsymbol{M}\boldsymbol{p}_i\|}{\|\boldsymbol{p}_i\|}\tanh^{-1}\left(\sqrt{c}\,\|\boldsymbol{p}_i\|\right)\right) \tag{8}$$

其中,如果 $\boldsymbol{M}\boldsymbol{p}_i=0$,则 $\boldsymbol{M}\otimes_c\boldsymbol{p}_i=0$。此时,$h_i$ 可以被认为是双曲图注意力隐藏层中的潜在表示。随后,我们在节点上执行自注意力机制。注意力系数 α_{ij} 表示节点 j 对于节点 i 的重要性,如式(9)所示:

$$\alpha_{ij}=f(h_i,h_j) \tag{9}$$

其中,$f(\cdot)$ 表示计算注意力系数的函数。此处本文只计算节点 $j \in N_i$ 的 α_{ij},N_i 是图中节点 i 的邻居。考虑到当节点 j 和 i 具有较高相似度时,其注意力系数 α_{ij} 也应随之增大,本文基于双曲空间的距离函数 $f(\cdot)$ 来度量节点间的相似性。具体而言,当广义双曲度量张量与欧氏度量张量共形(即共形因子为 λ_x^c),对于位于双曲空间 D_c^d 中的两个潜在表示 h_i 和 h_j,其双曲距离定义如式(10)所示:

$$d_c(h_i,h_j)=\frac{2}{\sqrt{c}}\tanh^{-1}\left(\sqrt{c}\,\|\Theta_c h_i\oplus_c h_j\|\right) \tag{10}$$

其中,$d_c(\cdot,\cdot)$ 是距离计算函数,$\oplus_c$ 是 D_c^d 中的莫比乌斯加法,$\Theta_c h_i=-h_i$ 是莫比乌斯加法的逆运算。莫比乌斯加法定义如式(11)所示:

算法1 双曲图注意力神经网络(HGAT)

输入:W_i^E:文本图节点欧氏特征;N_i:图邻居结构;c:曲率;$\boldsymbol{M}$:权重矩阵

输出:h'':层次化局部语义特征

1. /*** 欧氏→双曲空间:指数映射投影 ***/
2. $\boldsymbol{p}_i \leftarrow \exp_0^c(W_i^E)$
3. /*** 双曲空间线性变换 ***/
3. $h_i \leftarrow \boldsymbol{M}\otimes_c\boldsymbol{p}_i$
4. /*** 双曲距离注意力得分计算 ***/
5. $\alpha_{ij} \leftarrow -d_c(h_i,h_j), d_c(h_i,h_j) \leftarrow \frac{2}{\sqrt{c}}\tanh^{-1}\left(\sqrt{c}\,\|-h_i\oplus_c h_j\|\right)$
6. /*** 注意力归一化与特征聚合 ***/
7. $h_i'' \leftarrow \sigma\left(\sum_{j\in N_i}^{\oplus_c}\frac{\tanh\left(\omega_{ij}\tanh^{-1}\left(\sqrt{c}\,\|h_j\|\right)\right)h_j}{\sqrt{c}\,\|h_j\|}\right), \omega_{ij} \leftarrow \frac{\exp(\alpha_{ij})}{\sum_{k\in N_i}\exp(\alpha_{ik})}$
8. /*** 双曲→欧氏空间:对数映射转换 ***/
9. $h'' \leftarrow \log_{h_i}^c(h_j)$
10. 返回 h''

$$h_i\oplus_c h_j=\frac{\left(1+2c<h_i,h_j>+c\|h_j\|^2\right)h_i+\left(1-c\|h_i\|^2\right)h_j}{1+2c<h_i,h_j>+c^2\|h_i\|^2\|h_j\|^2} \tag{11}$$

此时,本文将自注意力系数计算如式(12)所示:

$$\alpha_{ij}=d_c(h_i,h_j) \tag{12}$$

其中,距离越远注意力得分越小,距离越近注意力得分越大。

此时,由于双曲空间属于度量空间,用双曲空间的距离来计算自注意力系数具有两方面优势:(1)满足三角不等式的传递性优势。与欧氏空间中的内积不同,双曲距离天然满足三角不等式,因此基于距离构建的自注意力机制能够保持节点间的语义传递性,更符合图结构中"相邻节点关联可传递"的性质。(2)自注意力权重的数学合理性。根据距离定义,任意节点 i 与自身的双曲距离为 $d_c(h_i,h_i)=0$,因此其注意力系数为 $\alpha_{ii}=-d_c(h_i,h_i)=0$。由于其他节点的距离均为正,此时 α_{ii} 始终是所有注意力分数中最大的,从而确保模型在聚合时优先保留节点自身的表示。这一点在数学上与图表示学习的要求一致,而基于内积的图注意力网络(如GAT)并不能天然保证这一性质[31]。

对于节点 i 的所有邻居(包括自身),为便于比较它们的注意力系数,本文使用 softmax 函数对其进行归一化,如式(13)所示:

$$\omega_{ij}=\frac{\exp(\alpha_{ij})}{\sum_{k\in N_i}\exp(\alpha_{ik})} \tag{13}$$

其中,归一化后的注意力系数 ω_{ij} 用于对 N_i 中所有节点 j 的潜在表示进行线性组合。因此,节点 i 的最终聚合表示 h_i'',如式(14)和(15)所示:

$$h_i'' = \sigma\left(\sum_{j\in N_i}^{\oplus_c} \omega_{ij} \otimes_c h_j\right) \tag{14}$$

$$\sigma(x) = \max(0,x) + \min\left(0, \exp(x) - 1\right) \tag{15}$$

其中,$\sum^{\oplus_c}$ 是莫比乌斯加法的累加,$\sigma(\cdot)$为ELU激活函数兼具ReLU的稀疏性和对负值的平滑处理能力。$\omega_{ij} \otimes_c h_j$可通过莫比乌斯标量乘法实现,如式(16)所示:

$$\omega_{ij} \otimes_c h_j := \frac{\tanh\left(\omega_{ij} \tanh^{-1}\left(\sqrt{c}\,\|h_j\|\right)\right) h_j}{\sqrt{c}\,\|h_j\|} \tag{16}$$

且$\omega_{ij} \otimes_c 0 := 0$。当双曲空间有效建模层次化语义关系后,本文通过对数映射将双曲空间向量转化为欧氏空间,如式(17)所示:

$$\begin{aligned} h'' &= \log_{h_i}^c(h_j) \\ &= \frac{\dfrac{2}{\sqrt{c}\,\lambda_{h_i}^c} \tanh^{-1}\left(\sqrt{c}\,\|\Theta_c h_i \oplus_c h_j\|\right)(\Theta_c h_i \oplus_c h_j)}{\|\Theta_c h_i \oplus_c h_j\|} \end{aligned} \tag{17}$$

随后,在HSCS模块中本文采用两层HGAT结构执行层次化语义特征提取,得到层次化的局部语义特征h''。此外,通过CLIP捕捉文本全局语义特征g''。

虽然HSCS模块能够有效捕捉层次化的文本语义,但由于诸如MDM[44]和Att-T2M[11]等方法在实现文本与动作的交互时通常仅使用单一的局部语义或全局语义作为指导,这种单一语义源的依赖使得粗粒度与细粒度行为的联合建模变得困难。

2.3　多粒度跨模态注意力机制

为了解决上述问题,本文提出了多粒度跨模态注意力机制(MGCA),通过句子级与单词级语义与行为特征进行跨模态对齐,实现语义一致的多粒度动作生成(如图2和算法2所示)。具体来说,本文首先将行为特征x_t经过卷积和时间步t融合后的特征定义为x_t''。随后将全局语义特征g''和局部语义特征h''进行线性映射获得全局特征$\tilde{g}$和局部特征$\tilde{h}$,如式(18)所示:

$$\tilde{g} = \text{Linear}(g''), \tilde{h} = \text{Linear}(h'') \tag{18}$$

其中,Linear(·)为线性映射矩阵。然后,将全局特征$\tilde{g}$和$\tilde{h}$分别与行为特征x_t''进行跨模态注意力,如式(19)所示:

$$x_g = \text{softmax}\left(\frac{x_t''\tilde{g}}{\sqrt{d}}\right)\tilde{g}, x_h = \text{softmax}\left(\frac{x_t''\tilde{h}}{\sqrt{d}}\right)\tilde{h} \tag{19}$$

其中,$\sqrt{d}$为归一化因子,softmax(·)将注意力分数转化为可解释的概率分布,实现模态特征对齐与重要性加权,并为模型提供稳定的梯度以学习复杂关联。为了实现全局信息与局部特征之间的交互,本文引入动量更新机制自适应融合全局和局部信息实现粗细粒度建模,如式(20)所示:

$$x_f = \lambda x_h + (1-\lambda) x_g \tag{20}$$

其中,λ为可学习参数,根据模型对于行为信息的感知自适应地调整局部和全局信息的占比。最后,本文采用残差的方式将粗细粒度特征x_f与x_t''进行融合获得$\tilde{x}_t$,降低信息损失提升行为生成质量。

总而言之,本文提出的HTMG表明,仅依赖单一粒度的文本语义难以充分刻画人体行为生成的复杂性,而显式建模文本的层次结构并在多粒度层面实现跨模态对齐同样至关重要。通过HSCS捕获动词为核心、状语与副词为辅助的层级依赖关系,模型能够获得更完整的语义表达。而MGCA则进一步在句子级与词级实现跨模态交互并自适应融合粗细粒度特征,使生成动作在核心行为和细节属性上同时保持与文本描述的一致性。

算法2　多粒度跨模态注意力机制(MGCA)

输入:x_t:行为特征;h'':层次化局部语义特征;g'':全局语义特征;t:时间步

输出:$\tilde{x}_t$:行为特征

1. /*** 行为特征与语义特征预处理 ***/
2. $x_t'' \leftarrow \text{Conv}(x_t)(1+\text{scale}_t) + \text{shift}_t$
3. $\tilde{g} \leftarrow \text{Linear}(g''), \tilde{h} \leftarrow \text{Linear}(h'')$
3. /*** 跨模态注意力加权 ***/
4. $x_g \leftarrow \text{softmax}\left(\frac{x_t''\tilde{g}}{\sqrt{d}}\right)\tilde{g}$
5. $x_h \leftarrow \text{softmax}\left(\frac{x_t''\tilde{h}}{\sqrt{d}}\right)\tilde{h}$
6. /*** 动量更新融合策略***/
7. $x_f \leftarrow \lambda x_h + (1-\lambda)x_g$
8. /*** 残差融合输出 ***/
9. $\tilde{x}_t \leftarrow x_t'' + x_f$
10. 返回 $\tilde{x}_t$

3　实验结果

3.1　数据集

本文所提方法在HumanML3D[45]和KIT-ML[46]数据集上进行评估,细节如下。

HumanML3D[45]数据集包含14 616条动作序列及44 970条文本描述。每个动作姿态由263维特征向量表示,该向量同时编码了全局运动属性(如根节点速度、高度和脚部接触情况)以及局部关节特征(包括相对于根节点的位置、速度和旋转)。其中,每个人包含22个关节。HumanML3D数据集按0.8∶0.05∶0.15的比例划分为训练集、验证集和测试集。

KIT-ML[46]数据集包含3 911条动作序列及6 278条

文本标注。每个姿态由251维特征向量表示,特征捕获了类似的全局和局部运动属性,其中局部信息来自与SMPL模型对齐的21个关节。KIT-ML数据集按0.8∶0.05∶0.15的比例划分为训练集、验证集和测试集。

3.2 评价指标

参考先前研究MDM[44],本文采用标准评估指标从以下四个关键方面对所提出的方法进行评估:(1)Fréchet Inception Distance(FID),用于评估生成动作的整体质量;(2)R-Precision和MultiModal Distance(MM-Dist),用于衡量输入文本与生成动作之间的语义对齐程度;(3)Diversity,用于量化不同生成动作之间的多样性;(4)MultiModality(MModality),用于衡量同一文本生成的多种动作之间的多样性。

3.3 实验细节

网络架构。HTMG采用类似MDM[44]的扩散模型,通过加噪与去噪过程实现行为生成。在加噪阶段,向输入数据注入高斯噪声以实现特征高斯化;在去噪阶段,利用由HSF构成的UNet架构对噪声进行预测与还原。具体而言,HSF模块由三部分组成:卷积(CNN)、1层HSCS和1层MGCA模块。其中,CNN模块用于捕获局部行为特征和融合时间步;HSCS模块中采用两层HGAT进行层次语义捕获,spaCy版本设置为1.0.0,曲率参数$c=1$;MGCA实现句子级与单词级跨模态特征交互与自适应融合。

可复现性。在训练过程中,本文采用DDPM[47],设定去噪步数$T=1\,000$,并在正向过程中将方差β从0.000 1线性增加至0.02。模型训练使用AdamW优化器,初始学习率为0.000 2,权重衰减系数为0.01。在RTX 4090上以批量大小64进行50 000次迭代训练。学习率每5 000步衰减0.9倍。在分类器自由引导(CFG)中,本文对10%的样本设置$c=\varnothing$。

3.4 对比实验

本节将所提方法HTMG与多种主流行为生成范式进行对比,包括:自编码器方法(如TEMOS[7]),自回归方法(如T2M-GPT[48]、MotionGPT[49]、Att-T2M[11]、MG-MotionLLM[50]、Motion-Agent[51]),以及基于扩散模型的方法(如MLD[21]、MDM[44]、Fg-T2M[2]、MotionDiffuse[16]、Motion Mamba[15]、MotionLCM[52]、DisCoRD[53]和MARDM[54])。为确保对比的公平性,对于已公开代码的基线方法,本文在相同设置下进行复现;而对于未公开代码的方法,则采用其论文中报告的结果进行比较。

定量对比。表1和表2展示了基线方法与本文所提HTMG方法在HumanML3D[45]和KIT-ML[46]数据集上的定量评估结果。从中可以观察到HTMG在R-Precision和MM-Dist关键指标上均取得最优性能。具体来说,在HumanML3D数据集上:(1)HTMG vs AE方法(如TEMOS[7])获得R Top 1提升9.9%,FID降低3.67,MM-Dist降低0.419。其主要原因在于HTMG采用扩散模型迭代去噪过程降低了AE方法压缩过程中带来的信息丢失问题,同时双曲空间的引入也带来了文本一致性的提升。(2)HTMG vs AR方法(如Att-T2M[11])获得R Top 1提升2.4%,FID降低0.048,MM-Dist降低0.108。其主要原因在于,本文提出的基于扩散模型方法天然解决了自回归方法采用码表方式存储行为信息容量受限导致表征不足问题。(3)HTMG vs DDPM方法(如MARDM[54])而言,R Top 1提升2%,FID降低0.047。其主要原因在于,双曲空间在有效地建模层次化文本语义加强文本信息理解能力的同时,多粒度跨模态交叉注意力显式引导行为生成,提升了文本-行为一致性和行为质量。

定性对比。图3展示了本文提出的HTMG与MotionDiffuse[16]、MDM[44]和MLD[21]的定性比较,红色框表示注意区域。从中可以观察到,MDM[44]虽然能够生成基础的语义动作,但在捕捉连续动作之间的细粒度过渡方面存在不足。尽管MLD[21]和MotionDiffuse[16]在该动作上有所改进,但在精准对齐文本描述方面仍显不足。对于指令"Walking forward and steps over an object, and then continue walking.",MLD[21]和MotionDiffuse[16]均未能生成与"跨过物体"相关的动作。而在"rise-walk-lay"的动作序列中,三个方法均遗漏了至少一个关键的动作过渡,HTMG则实现了文本与行为的高度对齐。基于上述观察,本文可以得出以下结论:(1)HSCS的层次化文本描述建模提升了模型对于文本的理解能力;(2)MGCA将HSCS提取的层次化信息与行为进行自适应地对齐,提升了文本-行为的一致性和行为质量。

3.5 消融实验

为了验证HSCS和MGCA模块的有效性,本文进行了表3所示的消融实验。具体来说,当HSCS为⊗时,本文采用CLIP替换HSCS;当MGCA为⊗时,本文采用交叉注意力替换MGCA。如表3所示,B vs A时R Top 1提升1.2%,FID降低0.024。从中可以观察到,HSCS策略的层次化依赖捕捉提升了模型对于文本的理解能力,保证了文本-行为一致性和高质量行为生成。当C vs A时FID降低0.079,MModality提升0.159。这些结果表明,MGCA实现粗细粒度协同指导行为生成时能够提升行为的质量和多模态特性,但是因为缺乏层次关系建模,文本-行为语义一致性难以保证。当D vs A时获得了最优的FID和Diversity。从中可以得出结论:HSCS提取的层次化文本语义与MGCA粗细粒度协同指导行为生成提升了行为生成质量和多样性。

表1　在 HumanML3D 测试集上的定量比较

Table 1　Quantitative comparison on the HumanML3D test set

模型	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
Real	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	—
TEMOS[7]	$3.734^{\pm.028}$	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
MLD[21]	$0.473^{\pm.004}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.081}$	$2.413^{\pm.079}$
T2M-GPT[48]	$0.116^{\pm.004}$	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
MDM[44]	$0.544^{\pm.044}$	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	**$2.799^{\pm.072}$**
Fg-T2M[2]	$0.243^{\pm.019}$	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$	$1.614^{\pm.049}$
Att-T2M[11]	$0.112^{\pm.006}$	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
MotionGPT[49]	$0.232^{\pm.008}$	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$	$2.008^{\pm.084}$
MotionDiffuse[16]	$0.630^{\pm.001}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
Motion Mamba[15]	$0.281^{\pm.009}$	$0.502^{\pm.003}$	$0.693^{\pm.002}$	$0.792^{\pm.002}$	$3.060^{\pm.058}$	$9.871^{\pm.084}$	$2.294^{\pm.058}$
MotionLCM[52]	$0.304^{\pm.003}$	$0.504^{\pm.002}$	$0.698^{\pm.003}$	$0.796^{\pm.002}$	$3.012^{\pm.007}$	$9.634^{\pm.064}$	$2.267^{\pm.082}$
DisCoRD[53]	$0.095^{\pm.011}$	$0.476^{\pm.008}$	$0.663^{\pm.006}$	$0.760^{\pm.007}$	$3.121^{\pm.009}$	—	$1.831^{\pm.048}$
Motion-Agent [51]	$0.230^{\pm.009}$	$0.515^{\pm.004}$	—	$0.801^{\pm.004}$	$2.967^{\pm.020}$	$9.908^{\pm.102}$	—
MG-MotionLLM[50]	$0.303^{\pm.010}$	$0.516^{\pm.002}$	$0.706^{\pm.002}$	$0.802^{\pm.003}$	$2.952^{\pm.009}$	$9.960^{\pm.073}$	—
MARDM[54]	$0.114^{\pm.007}$	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$3.270^{\pm.009}$	—	$2.231^{\pm.071}$
Ours	**$0.064^{\pm.002}$**	**$0.523^{\pm.003}$**	**$0.713^{\pm.003}$**	**$0.808^{\pm.002}$**	**$2.930^{\pm.007}$**	**$9.528^{\pm.097}$**	$1.995^{\pm.078}$

注:对于每个指标,我们重复评估20次,并报告“±”带有95%置信区间的平均值。粗体表示最佳结果,下划线表示次优结果。“↑”表示越高越好,“↓”表示越低越好,“→”表示越接近真实动作越好。“—”表示无具体数值。

表2　在 KIT-ML 测试集上的定量比较

Table 2　Quantitative comparison on the KIT-ML test set

方法	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
Real	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	
TEMOS[7]	$3.717^{\pm.051}$	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$	$0.532^{\pm.034}$
MDM[44]	$0.497^{\pm.021}$	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$9.190^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
MLD[21]	$0.404^{\pm.027}$	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
T2M-GPT[48]	$0.514^{\pm.029}$	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
Fg-T2M[2]	**$0.175^{\pm.047}$**	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$3.114^{\pm.015}$	$10.93^{\pm.083}$	$1.019^{\pm.029}$
Att-T2M[11]	$0.870^{\pm.039}$	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
MotionDiffuse[16]	$1.954^{\pm.062}$	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$2.958^{\pm.005}$	**$11.10^{\pm.143}$**	$0.730^{\pm.013}$
MotionGPT[49]	$0.510^{\pm.013}$	$0.366^{\pm.005}$	$0.558^{\pm.006}$	$0.680^{\pm.005}$	$3.527^{\pm.021}$	$10.35^{\pm.084}$	$2.328^{\pm.117}$
Motion Mamba[15]	$0.307^{\pm.041}$	$0.419^{\pm.006}$	$0.645^{\pm.005}$	$0.765^{\pm.006}$	$3.021^{\pm.025}$	$11.02^{\pm.098}$	$1.678^{\pm.064}$
DisCoRD[53]	$0.541^{\pm.038}$	$0.382^{\pm.007}$	$0.590^{\pm.007}$	$0.715^{\pm.004}$	$3.260^{\pm.028}$	—	$1.928^{\pm.059}$
MARDM[54]	$0.242^{\pm.014}$	$0.387^{\pm.006}$	$0.610^{\pm.006}$	$0.749^{\pm.006}$	$3.374^{\pm.019}$	—	$1.312^{\pm.053}$
Ours	$0.476^{\pm.025}$	**$0.440^{\pm.007}$**	**$0.652^{\pm.006}$**	**$0.769^{\pm.004}$**	**$2.926^{\pm.012}$**	$10.565^{\pm.099}$	$1.494^{\pm.057}$

注:对于每个指标,我们重复评估20次,并报告“±”带有95%置信区间的平均值。粗体表示最佳结果,下划线表示次优结果。“↑”表示越高越好,“↓”表示越低越好,“→”表示越接近真实动作越好。“—”表示无具体数值。

为了验证同一树结构下HGAT的有效性,本文进行了消融对比实验(GCN/GAT vs HGAT),如表4所示。实验结果表明,树结构的主要贡献在于语义对齐,它使欧氏GAT的R Top 1从序列基线的0.517提升至0.522;然而,仅有结构约束仍不足以提升生成质量,从欧氏GAT的FID未出现明显改善可以看出,其表示能力在层级建模方面依然受限。相比之下,双曲HGAT将FID显著优化至0.064。这证明了本文所提方法并非简单的结构堆砌,而是利用树结构捕捉句法依赖,同时依赖双曲几何解决层级嵌入的失真问题,二者协同作用才保证了文本-行为一致性和行为生成质量。

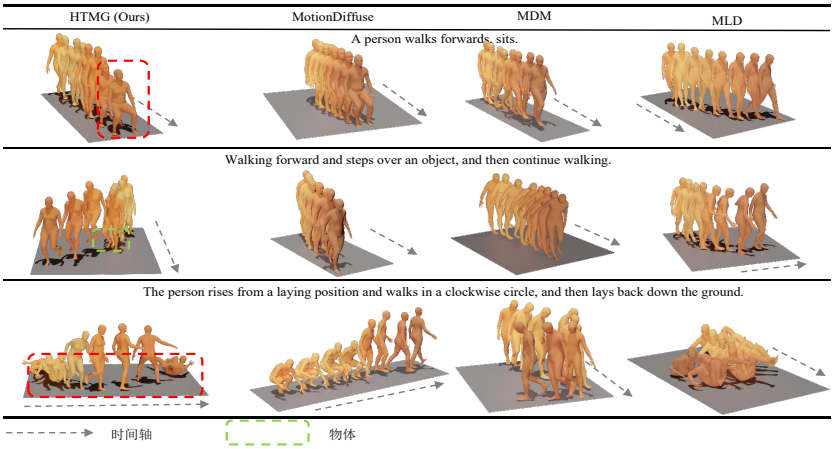

图3 HumanML3D测试集上的定性对比

Figure 3 Qualitative comparison on the HumanML3D test se

表3 HSCS和MGCA模块在HumanML3D测试集上的消融实验

Table 3 Ablation study of the HSCS and MGCA modules on the HumanML3D test set

编号	HSCS	MGCA	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
				Top 1↑	Top 2↑	Top 3↑			
A	×	×	$0.151^{\pm.005}$	$0.546^{\pm.004}$	$0.745^{\pm.004}$	$0.838^{\pm.003}$	$2.801^{\pm.010}$	$9.822^{\pm.075}$	$1.874^{\pm.067}$
B	√	×	$0.127^{\pm.005}$	$\mathbf{0.558}^{\pm.003}$	$\mathbf{0.756}^{\pm.003}$	$\mathbf{0.847}^{\pm.002}$	$\mathbf{2.726}^{\pm.008}$	$9.753^{\pm.092}$	$1.704^{\pm.060}$
C	×	√	$0.072^{\pm.003}$	$0.517^{\pm.002}$	$0.709^{\pm.002}$	$0.806^{\pm.002}$	$2.946^{\pm.007}$	$\mathbf{9.471}^{\pm.095}$	$\mathbf{2.025}^{\pm.061}$
D	√	√	$\mathbf{0.064}^{\pm.002}$	$0.523^{\pm.003}$	$0.713^{\pm.003}$	$0.808^{\pm.002}$	$2.930^{\pm.007}$	$9.528^{\pm.097}$	$1.995^{\pm.078}$

注:“√”表示使用模块,“×”表示不使用模块。对于每个指标,我们重复评估20次,并报告“±”带有95%置信区间的平均值。粗体表示最佳结果。“↑”表示越高越好,“↓”表示越低越好,“→”表示越接近真实动作越好。

表4 HSCS的语义捕捉策略消融实验

Table 4 Ablation study of the semantic capture strategies in HSCS

文本结构	空间几何	方法	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
				Top 1↑	Top 2↑	Top 3↑			
序列	欧氏	Attention	$0.072^{\pm.003}$	$0.517^{\pm.002}$	$0.709^{\pm.002}$	$0.806^{\pm.002}$	$2.946^{\pm.007}$	$9.471^{\pm.095}$	$\mathbf{2.025}^{\pm.061}$
树	欧氏	GCN	$0.074^{\pm.004}$	$0.519^{\pm.003}$	$0.711^{\pm.003}$	$0.806^{\pm.002}$	$2.931^{\pm.008}$	$9.466^{\pm.081}$	$1.990^{\pm.083}$
树	欧氏	GAT	$0.075^{\pm.004}$	$0.522^{\pm.002}$	$\mathbf{0.715}^{\pm.003}$	$0.808^{\pm.003}$	$\mathbf{2.916}^{\pm.009}$	$9.570^{\pm.087}$	$1.957^{\pm.067}$
树	双曲	HGAT	$\mathbf{0.064}^{\pm.002}$	$\mathbf{0.523}^{\pm.003}$	$0.713^{\pm.003}$	$\mathbf{0.808}^{\pm.002}$	$2.930^{\pm.007}$	$\mathbf{9.528}^{\pm.097}$	$1.995^{\pm.078}$

注:对于每个指标,我们重复评估20次,并报告“±”带有95%置信区间的平均值。粗体表示最佳结果。“↑”表示越高越好,“↓”表示越低越好,“→”表示越接近真实动作越好。

为了验证模型曲率敏感性,本文进行了不同曲率$c\in\{0.1, 0.5, 1.0, 2.0\}$下的实验,如表5所示。结果表明,模型性能随曲率变化呈现先升后降的趋势,其中$c=1.0$为最佳平衡点,在此设定下模型取得了最优的FID及多模态指标。分析发现,生成质量(如FID)对曲率变化较为敏感,过大的曲率($c=2.0$)会导致性能显著下降;而语义匹配指标(R-Precision)则表现出较强的鲁棒性,在不同曲率下波动极小。因此,本文最终选取$c=1.0$作为模型的默认设置以确保最佳性能。

为了验证HSCS的层次化语义理解能力,本文在保持MGCA不变的前提下,对比分析了CLIP与HSCS在文本-行为一致性方面的贡献。如表6所示,HSCS vs CLIP时R Top 1提升0.6%,FID降低0.008,MM-Dist降低0.016。如图4所示,采用CLIP方法会遗漏关键动作(如“sits”)和方向信息(如“middle”)。而HSCS则根据动作的层次关系逐一激活单词对应行为。此外,本文对FID进行了Welch's t检验,显著性水平设置为0.05。对于CLIP vs HSCS而言,自由度为33.1,t的统计量为4.349,P值为0.000 123<0.05。结果表明HSCS的性能在统计学上显著优于CLIP方法。从中可以得出结论:将文本描述映射到双曲空间并显式建模单词间的依存结构与层级关系,相比于CLIP在欧氏线性空间中的语义建模,更有助于提升模型对文本语义的理解能力,进而增强文本-行为一致性。

表5　曲率敏感性实验

Table 5　Curvature sensitivity experiment

曲率(c)	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
0.1	$0.081^{\pm.004}$	$0.521^{\pm.003}$	$0.711^{\pm.003}$	$0.806^{\pm.002}$	$2.933^{\pm.007}$	$9.499^{\pm.098}$	$1.947^{\pm.077}$
0.5	$0.073^{\pm.003}$	$0.521^{\pm.003}$	$0.712^{\pm.003}$	$\mathbf{0.808^{\pm.002}}$	$2.933^{\pm.009}$	$9.499^{\pm.107}$	$1.981^{\pm.071}$
1.0	$\mathbf{0.064^{\pm.002}}$	$\mathbf{0.523^{\pm.003}}$	$\mathbf{0.713^{\pm.003}}$	$\mathbf{0.808^{\pm.002}}$	$\mathbf{2.930^{\pm.007}}$	$\mathbf{9.528^{\pm.097}}$	$\mathbf{1.995^{\pm.078}}$
2.0	$0.089^{\pm.004}$	$0.521^{\pm.003}$	$0.710^{\pm.003}$	$0.806^{\pm.002}$	$2.943^{\pm.008}$	$9.483^{\pm.091}$	$1.988^{\pm.075}$

注:对于每个指标,我们重复评估20次,并报告"±"带有95%置信区间的平均值。粗体表示最佳结果。"↑"表示越高越好,"↓"表示越低越好,"→"表示越接近真实动作越好。

表6　HSCS模块在HumanML3D测试集上层次化语义理解能力的有效性消融实验

Table 6　Ablation study on the effectiveness of the HSCS module for hierarchical semantic understanding on the HumanML3D test set

文本理解方法	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
CLIP	$0.072^{\pm.003}$	$0.517^{\pm.002}$	$0.709^{\pm.002}$	$0.806^{\pm.002}$	$2.946^{\pm.007}$	$9.471^{\pm.095}$	$\mathbf{2.025^{\pm.061}}$
HSCS	$\mathbf{0.064^{\pm.002}}$	$\mathbf{0.523^{\pm.003}}$	$\mathbf{0.713^{\pm.003}}$	$\mathbf{0.808^{\pm.002}}$	$\mathbf{2.930^{\pm.007}}$	$\mathbf{9.528^{\pm.097}}$	$1.995^{\pm.078}$

注:对于每个指标,我们重复评估20次,并报告"±"带有95%置信区间的平均值。粗体表示最佳结果。"↑"表示越高越好,"↓"表示越低越好,"→"表示越接近真实动作越好。

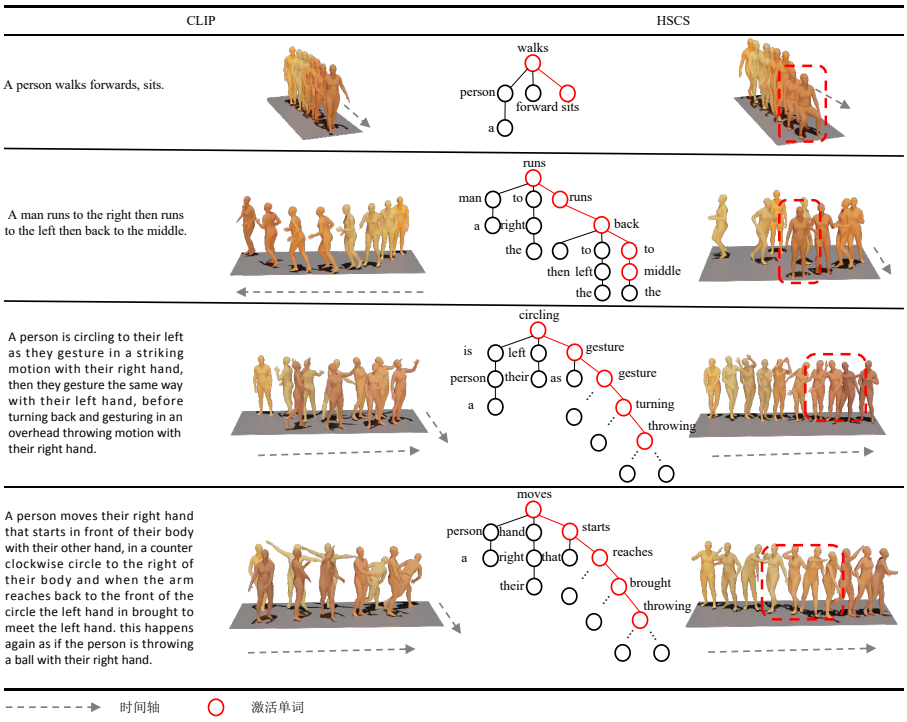

图4　HSCS和CLIP的文本-行为激活对应关系图

Figure 4　Visualization of the text-motion activation correspondence of HSCS and CLIP

为了验证模型对复杂句法结构的建模能力,本文筛选HumanML3D测试集中88个超过40个单词的长文本作为"困难样本"子集进行对比,如表7所示。结果表明,随着文本复杂度的增加,HSCS的层次化建模优势愈发凸显:在CLIP的生成质量因文本过长而显著下降时,HSCS并在R Top 1上获得2.3%的性能提升。如图4所示,第三和第四条指令执行时,CLIP方法会出现方向错误(如"counter clockwise circle")和遗漏动作执行(如"throwing motion with their right hand")。而HSCS则根据动作的层次关系逐一激活单词对应行为。从中可以得出结论:HSCS的层次化语义建模提升了模型处理长文本能力,保证了文本-行为一致性和生成质量。

表7 HSCS在HumanML3D测试集的88条长文本上的对比实验

Table 7 Comparative experiments of HSCS on the 88 long-text descriptions in the HumanML3D test set

方法	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
CLIP	$1.464^{\pm.142}$	$0.331^{\pm.017}$	$0.521^{\pm.019}$	$\mathbf{0.648}^{\pm.018}$	$3.623^{\pm.063}$	$8.067^{\pm.112}$	$\mathbf{2.055}^{\pm.068}$
HSCS	$\mathbf{1.196}^{\pm.087}$	$\mathbf{0.354}^{\pm.013}$	$\mathbf{0.529}^{\pm.017}$	$0.638^{\pm.015}$	$\mathbf{3.583}^{\pm.067}$	$\mathbf{7.894}^{\pm.137}$	$2.009^{\pm.061}$

注:对于每个指标,我们重复评估20次,并报告"±"带有95%置信区间的平均值。粗体表示最佳结果。"↑"表示越高越好,"↓"表示越低越好,"→"表示越接近真实动作越好。

为了验证MGCA的粗细粒度协同指导行为生成的有效性[式(20)],本文对比了两个基线方法(仅使用全局语义或局部语义)。如表8所示,MGCA vs 全局语义时R Top 1提升0.4%,FID降低0.019,MM-Dist降低0.005。MGCA vs 局部语义时FID降低0.135,MModality提升0.159。如图5所示,在引入全局语义情况下能够准确生成"runs"和"middle",但是因为缺乏细节语义导致运动方向混淆。在引入局部语义情况下方向正确但是最终生成位置错误。唯有同时融合全局语义与局部语义,才能实现文本描述与生成行为的精准对齐。这些结果表明:(1)局部语义获得最优的R Top 1,主要归因于HSCS提取的层次化词级语义表示能够更精确地刻画词间依存结构,使生成动作在语义空间中更贴近文本指令。(2)粗细粒度语义的协同建模能够在保持语义一致性的同时显著提高生成质量,验证了多粒度互补信息在动作生成中的关键作用。

表8 MGCA模块在HumanML3D测试集上的粗细粒度语义协同指导有效性消融实验

Table 8 Ablation study on the effectiveness of the MGCA module for coarse-to-fine semantic collaborative guidance on the HumanML3D test set

全局语义	局部语义	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
			Top 1↑	Top 2↑	Top 3↑			
√		$0.083^{\pm.003}$	$0.519^{\pm.003}$	$0.711^{\pm.002}$	$0.808^{\pm.002}$	$2.935^{\pm.008}$	$9.470^{\pm.091}$	$1.963^{\pm.081}$
	√	$0.209^{\pm.006}$	$\mathbf{0.542}^{\pm.003}$	$\mathbf{0.739}^{\pm.002}$	$\mathbf{0.831}^{\pm.002}$	$\mathbf{2.834}^{\pm.009}$	$9.985^{\pm.097}$	$1.756^{\pm.060}$
√	√	$\mathbf{0.064}^{\pm.002}$	$0.523^{\pm.003}$	$0.713^{\pm.003}$	$0.808^{\pm.002}$	$2.930^{\pm.007}$	$\mathbf{9.528}^{\pm.097}$	$\mathbf{1.995}^{\pm.078}$

注:"√"表示使用模块。对于每个指标,我们重复评估20次,并报告"±"带有95%置信区间的平均值。粗体表示最佳结果。"↑"表示越高越好,"↓"表示越低越好,"→"表示越接近真实动作越好。

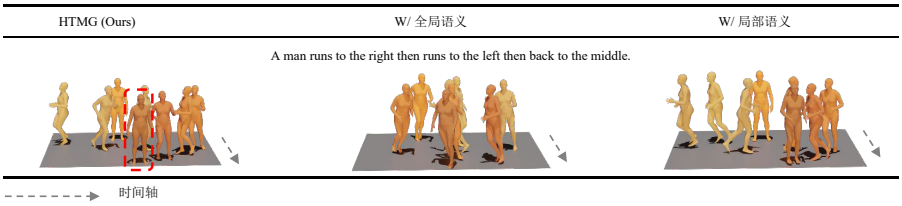

图5 全局语义和局部语义的消融实验可视化

Figure 5 Visualization of the ablation study on global and local semantics

为了验证MGCA的融合策略的有效性[式(20)],本文对比了两个基线方法(相加与拼接)。如表9所示,动量融合 vs 相加时R Top 1提升0.8%,FID降低0.042,MM-Dist降低0.057。动量融合vs拼接时R Top 1提升1.9%,FID降低0.137,MM-Dist降低0.1。从中可以观察到,本文提出的动量融合策略显著提升了文本与行为之间的一致性。其核心原因在于,自适应的跨模态对齐有效减少了简单"相加"或"拼接"融合方式所引入的信息冗余,从而进一步提升了整体的行为生成质量。

表9 MGCA模块在HumanML3D测试集上融合策略的消融实验

Table 9 Ablation study of fusion strategies in the MGCA module on the HumanML3D test set

融合策略	FID↓	R-Precision			MM-Dist↓	Diversity→	MModality↑
		Top 1↑	Top 2↑	Top 3↑			
相加	$0.106^{\pm.004}$	$0.515^{\pm.003}$	$0.707^{\pm.003}$	$0.803^{\pm.002}$	$2.987^{\pm.007}$	$9.745^{\pm.109}$	$\mathbf{2.103}^{\pm.068}$
拼接	$0.201^{\pm.007}$	$0.504^{\pm.003}$	$0.698^{\pm.003}$	$0.795^{\pm.002}$	$3.032^{\pm.001}$	$9.712^{\pm.071}$	$2.086^{\pm.070}$
动量融合	$\mathbf{0.064}^{\pm.002}$	$\mathbf{0.523}^{\pm.003}$	$\mathbf{0.713}^{\pm.003}$	$\mathbf{0.808}^{\pm.002}$	$\mathbf{2.930}^{\pm.007}$	$\mathbf{9.528}^{\pm.097}$	$1.995^{\pm.078}$

注:对于每个指标,我们重复评估20次,并报告"±"带有95%置信区间的平均值。粗体表示最佳结果。"↑"表示越高越好,"↓"表示越低越好,"→"表示越接近真实动作越好。

4 结论

本文提出了一种新颖的行为框架HTMG,该框架在全面理解文本语义的同时实现了粗细粒度的跨模态交互,从而实现文本-行为的一致性。具体而言,为解决文本语义理解困难的问题,本文提出层次化语义捕捉策略,通过句法依存分析构建文本结构树并引入双曲图注意力机制,在双曲空间中动态建模文本的层次依赖关系,从而强化模型的语义解析能力。为实现多粒度语义与动作特征的有效对齐,本文设计多粒度跨模态注意力机制,自适应融合句子级与单词级语义与动作特征,实现语义一致的粗细粒度行为建模。大量实验结果充分验证了HTMG的有效性,该方法在HumanML3D和KIT-ML数据集上均取得了最优性能。

参考文献

[1] Huang, Yiheng, Yang, Hui, Luo, Chuanchen, et al. StableMoFusion: Towards robust and efficient diffusion-based motion generation framework[PP/OL]. V2.arXiv (2024-12-09)[2025-11-24]. https://doi.org/10.48550/arXiv.2405.05691.

[2] Yin Wang, Leng Zhiying, Li F W B, et al. Fg-T2M: Fine-grained text-driven human motion generation via diffusion model[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 21978-21987.

[3] Guo Chuan, Mu Yuxuan, Javed M G, et al. MoMask: Generative masked modeling of 3D human motions[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 1900-1910.

[4] Li Chengjian, Shu Xiangbo, Cui Qingjie, et al. FTMoMamba: Motion generation with frequency and text state space models[PP/OL]. V1.arXiv (2024-11-26)[2025-01-24]. https://doi.org/10.48550/arXiv.2411.17532.

[5] Javed M G, Guo Chuan, Cheng Li, et al. InterMask: 3D human interaction generation via collaborative masked modeling[PP/OL]. V3.arXiv (2025-03-02)[2024-11-28]. https://doi.org/10.48550/arXiv.2410.10010.

[6] Jeong M, Hwang Y, Lee J, et al. HGM3: Hierarchical generative masked mo- tion modeling with hard token mining[C]//The Thirteenth International Conference on Learning Representations (ICLR).2025.

[7] Petrovich M, Black M J, Varol G. TEMOS: Generating diverse human motions from textual descriptions[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 480-497.

[8] Tevet G, Gordon B, Hertz A, et al. MotionCLIP: Exposing human motion generation to CLIP space[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 358-374.

[9] Barsoum E, Kender J, Liu Zicheng. HP-GAN: Probabilistic 3D human motion prediction via GAN[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2018: 1499-149909.

[10] Harvey F G, Yurick M, Nowrouzezahrai D, et al. Robust motion in-betweening[J]. ACM Transactions on Graphics, 2020, 39(4): 1-12.

[11] Zhong Chongyang, Hu Lei, Zhang Zihao, et al. AttT2M: Text-driven human motion generation with multi-perspective attention mechanism[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 509-519.

[12] Pinyoanuntapong E, Wang Pu, Lee M, et al. MMM: Generative masked motion model[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 1546-1555.

[13] Pinyoanuntapong E, Saleem M U, Wang Pu, et al. BAMM: Bidirectional autoregressive motion model[M]//Computer Vision - ECCV 2024. Cham: Springer Nature Switzerland, 2024: 172-190.

[14] Gong Kehong, Lian Dongze, Chang Heng, et al. TM2D: Bimodality driven 3D dance generation via music-text integration[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 9908-9918.

[15] Zhang Zeyu, Liu A, Reid I, et al. Motion mamba: Efficient and long sequence motion generation[M]//Computer Vision - ECCV 2024. Cham: Springer Nature Switzerland, 2024: 265-282.

[16] Zhang Mingyuan, Cai Zhongang, Pan Liang, et al. MotionDiffuse: Text-driven human motion generation with diffusion model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4115-4128.

[17] Wei Mingjie, Xie Xuemei, Shi Guangming. ACMo: Attribute controllable motion generation[PP/OL]. V1.arXiv (2025-03-14)[2025-07-08]. https://doi.org/10.48550/arXiv.2503.11038.

[18] Zheng Bowen, Chen Ke, Yao Yuxin, et al. AutoKeyframe: Autoregressive keyframe generation for human motion synthesis and editing[C]//Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. New York: ACM, 2025: 3730664.

[19] Wu Bizhu, Xie Jinheng, Ding Meidanet al. FineMotion: A dataset and benchmark with both spatial and temporal annotation for fine-grained motion generation and editing[PP/OL]. V1. arXiv (2025-07-26) [2025-07-29]. https://doi.org/10.48550/arXiv.2507.19850.

[20] Zhong Lei, Yang Yi, Li Changjian. SMooGPT: Stylized

motion generation using large language models[PP/OL]. V2.arXiv (2026-01-26)[2025-09-05]. https://doi.org/10.48550/arXiv.2509.04058.

[21] Chen Xin, Jiang Biao, Liu Wen, et al. Executing your commands via motion diffusion in latent space[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 18000-18010.

[22] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C/OL]//2021 International Conference on Machine Learning(ICML). PMLR, 2021: 8748-8763.

[23] Nivre J. Algorithms for deterministic incremental dependency parsing[J]. Computational Linguistics, 2008, 34(4): 513-553.

[24] He N, Anand R, Madhu H, et al. HELM: Hyperbolic large language models via mixture-of-curvature experts[PP/OL]. V2. arXiv (2025-11-06)[2025-10-09]. https://doi.org/10.48550/arXiv.2505.24722.

[25] Chami I, Ying R, Re C, et al. Hyperbolic graph convolutional neural networks[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM, 2019: 4868-4879.

[26] Li Jun, Wang Jinpeng, Tan C L, et al. HLFormer: Enhancing partially relevant video retrieval with hyperbolic learning[PP/OL]. V2. arXiv (2025-07-27) [2025-10-25]. https://doi.org/10.48550/arXiv.2507.17402.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]//Advances in Neural Information Processing Systems.2017.30.

[28] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5769-5779.

[29] Van den oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6309-6318.

[30] Foundations of hyperbolic manifolds[EB/OL]. [2025-10-25]. https://link.springer.com/book/10.1007/978-1-4757-4013-4.

[31] Zhang Y D, Wang X, Jiang X Q, et al. Hyperbolic graph attention network[PP/OL]. V1. arXiv (2019-12-06)[2025-11-07]. https://doi.org/10.48550/arXiv.1912.03046.

[32] Pal A, van Spengler M, di Melendugno G M D, et al. Compositional entailment learning for hyperbolic vision-language models[PP/OL]. V2. arXiv (2025-03-01) [2025-10-08]. https://doi.org/10.48550/arXiv.2410.06912.

[33] Li Yue, Qu Haoxuan, Liu Mengyuan, et al. HyLiFormer: Hyperbolic linear attention for skeleton-based human action recognition[PP/OL]. V1. arXiv (2025-02-09)[2025-10-09]. https://doi.org/10.48550/arXiv.2502.05869.

[34] Peng Zelin, Xu Zhengqin, Liu Qingyang, et al. HyperET: Efficient training in hyperbolic space for multi-modal large language models[PP/OL]. V3. arXiv (2025-12-18)[2025-10-24]. https://doi.org/10.48550/arXiv.2510.20322.

[35] Mandica P, Franco L, Kallidromitis K, et al. Hyperbolic learning with multimodal large language models[PP/OL]. V1. arXiv (2024-08-09)[2025-10-09]. https://doi.org/10.48550/arXiv.2408.05097.

[36] 王彩霞, 安琪, 周鸿策, 等. 基于特征自适应选取的视觉目标跟踪算法[J]. 电子学报, 2025, 53(8): 2879-2898.
Wang Caixiang, An Qi, Zhou Hongce, et al. Visual object tracking algorithm based on adaptive feature selection[J]. Acta Electronica Sinica, 2025, 53(8): 2879-2898. (in Chinese)

[37] 秦钰淑, 杨良怀, 朱艳超, 等. 融合图像与文本特征的组合检索方法[J]. 电子学报, 2025, 53(2): 558-567.
Qin Yushu, Yang Lianghuai, Zhu Yanchao, et al. A combined retrieval method by fusing image and text features[J]. Acta Electronica Sinica, 2025, 53(2): 558-567. (in Chinese)

[38] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[PP/OL]. V1.arXiv (2023-02-27)[2025-10-25]. https://doi.org/10.48550/arXiv.2302.13971.

[39] Liu Aixin, Feng Bei, Xue Bing, et al. DeepSeek-V3 technical report[PP/OL]. V2.arXiv (2025-02-18)[2025-10-25]. https://doi.org/10.48550/arXiv.2412.19437.

[40] 李豪, 郝文宁, 邹世辰, 等. 基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法[J]. 电子学报, 2025, 53(9): 3384-3396.
Li Hao, Hao Wenning, Zou Shichen, et al. Progressive image synthesis method based on diffusion-mamba and scale-invariant loss[J]. Acta Electronica Sinica, 2025, 53(9): 3384-3396. (in Chinese)

[41] Muscoloni A, Thomas J M, Ciucci S, et al. Machine learning meets complex networks via coalescent embedding in the hyperbolic space[J]. Nature Communications, 2017, 8: 1615.

[42] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

[43] Krioukov D, Papadopoulos F, Kitsak M, et al. Hyperbolic geometry of complex networks[J]. Physical Review E, 2010, 82(3): 036106.

[44] Tevet G, Raab S, Gordon B, et al. Human motion diffusion model[PP/OL]. V2.arXiv (2022-10-03)[2025-11-06].

https://doi.org/10.48550/arXiv.2209.14916.
[45] Guo Chuan, Zou Shihao, Zuo Xinxin, et al. Generating diverse and natural 3D human motions from text[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5142-5151.
[46] Plappert M, Mandery C, Asfour T. The KIT motion-language dataset[J]. Big Data, 2016, 4(4): 236-252.
[47] HO J, JAIN A, ABBEEL P. Denoising Diffusion Probabilistic Models[J]. Advances in neural information processing systems, 2020, 33.
[48] Zhang Jianrong, Zhang Yangsong, Xiaodong Cun, et al. Generating human motion from textual descriptions with discrete representations[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 14730-14740.
[49] Jiang Biao, Chen Xin, Liu Wen, et al. MotionGPT: Human motion as a foreign language[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 20067-20079.
[50] Wu Bizhu, Xie Jinheng, Shen Keming, et al. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 27849-27858.
[51] Wu Qi, Zhao Yubo, Wang Yifan, et al. Motion-agent: A conversational framework for human motion generation with LLMs[PP/OL]. V3. arXiv (2024-10-06) [2025-11-17]. https://doi.org/10.48550/arXiv.2405.17013.
[52] Dai Wenxun, Chen Ling-Hao, Wang Jingbo, et al. MotionLCM: Real-time controllable motion generation via latent consistency model[M]//Computer Vision - ECCV 2024. Cham: Springer Nature Switzerland, 2024: 390-408.
[53] CHO J, KIM J, KIM J, et al. DisCoRD: Discrete tokens to continuous motion via rectified flow decoding[C] // 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2025: 14602-14612.
[54] Meng Zichong, Xie Yiming, Peng Xiaogang, et al. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 27859-27871.

作者简介

舒祥波　男,1986年3月出生于湖北省孝感市。现为南京理工大学计算机科学与工程学院副院长,博士生导师。主要研究方向为人体行为计算、视频理解和行为分析。
E-mail: shuxb@njust.edu.cn

李朋鹏　男,1996年11月出生于安徽省合肥市。现为南京理工大学计算机科学与工程学院博士研究生。主要研究方向为手术视频理解。
E-mail: pengpengli@njust.edu.cn

李成建　男,1995年8月出生于河南省周口市。现为南京理工大学计算机科学与工程学院博士研究生。主要研究方向为计算机视觉和人体行为生成。
E-mail: lichengjian@njust.edu.cn

李泽超　男,1985年5月出生于河南省开封市。现为南京理工大学计算机科学与工程学院院长,博士生导师。主要研究方向为智能媒体分析和计算机视觉。中国电子学会会员编号:E190031283S。
E-mail: zechao.li@njust.edu.cn

尹　政　男,2002年4月出生于湖北省洪湖市。现为南京理工大学计算机科学与工程学院硕士研究生。主要研究方向为人体动作预测与生成。
E-mail: alanyz@njust.edu.cn

唐金辉　男,1981年2月出生于江苏省丹阳市。现为南京林业大学副校长,博士生导师。主要研究方向为多媒体分析与检索和计算机视觉。
E-mail: tangjh@njfu.edu.cn