基于时序感知潜在扩散模型的人体交互动作生成

石 旭[1)] 孙运莲[1)] 骆岩林[2)] 张鸿文[2)]
[1)](南京理工大学计算机科学与工程学院 南京 210094)
[2)](北京师范大学人工智能学院 北京 100875)

摘 要 近年来,人体动作生成在计算机视觉和计算机图形学领域受到了广泛关注。随着需求的增加,人体交互动作生成逐渐成为一个新的研究热点。然而,相较于单人动作生成,人体交互动作生成尚处于起步阶段,尤其是在生成复杂的交互动作方面。虽然基于文本条件的人体交互动作生成方法在生成符合文本描述的高质量人体交互动作方面已取得一定进展,但现有方法大多在原始动作序列上进行生成模型的学习,导致生成速度较慢。此外,它们普遍沿用对比语言-图像预训练(Contrastive Language-Image Pretraining,CLIP)模型的文本编码器作为动作生成模型的语言指导,这导致动作生成模型缺乏对动作的时序感知,影响了生成动作的质量。为了解决这些问题,针对人体交互动作生成,本文提出一种人体交互动作潜在扩散模型(Human interaction Latent Diffusion Model,HiLDM)。该扩散模型通过在学习到的人体交互动作序列潜在空间中进行去噪,大幅提升生成速度。同时,采用人体交互时序感知文本编码器(Temporal-aware Text Encoder,TTE)作为语言指导,使生成动作更具时序一致性。实验结果表明,在InterHuman数据集上的评估中,所提方法在生成速度和生成质量方面优于现有人体交互动作生成方法,生成速度比ComMDM快57倍,比InterGen快4倍;FID指标比ComMDM改善了36.7%,比InterGen改善了1.7%。

关键词 潜在扩散模型;人体动作生成;人体交互动作生成;人工智能生成内容
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.02226

Human Interaction Generation Based on Temporal-Aware Latent Diffusion Model

SHI Xu[1)] SUN Yun-Lian[1)] LUO Yan-Lin[2)] ZHANG Hong-Wen[2)]
[1)](School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)
[2)](School of Artificial Intelligence, Beijing Normal University, Beijing 100875)

Abstract In recent years, human motion generation has gained significant attention in the fields of computer vision and computer graphics, demonstrating substantial application value in virtual reality, animation production, game development, and sports training. With technological advancements, the demand for generating high-quality and naturally smooth human motions continues to grow. Among them, human interaction generation has emerged as a new research focus, aiming to synthesize dual-person or multi-person motions with temporal consistency, spatial plausibility, and interaction coordination. However, compared to single-person motion generation, human interaction generation is still in its early stages. In particular, ensuring coordination and plausibility between individuals in complex interactions remains a significant challenge.

收稿日期:2024-10-10;在线发布日期:2025-04-15。本课题得到国家自然科学基金面上项目(62476131,62076131,62377004)、中央高校基本科研业务费项目(2233100028)资助。**石 旭**,硕士,主要研究领域为人体动作生成。E-mail:shixu@njust.edu.cn。**孙运莲**(**通信作者**),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为模式识别与计算机视觉。E-mail:yunlian.sun@njust.edu.cn。**骆岩林**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为虚拟现实与可视化。**张鸿文**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为三维数字人体重建与生成。

Current text-guided human interaction generation methods have made progress in generating high-quality interactive motions that align with textual descriptions. However, several critical issues remain unresolved. First, most diffusion methods perform denoising directly on raw motion sequences, which incurs high computational costs and results in slow generation speeds, making them unsuitable for real-time applications. Moreover, these methods often struggle with capturing fine-grained motion details, leading to suboptimal results. Second, the text encoder serves as the linguistic guide for motion generation, yet it lacks temporal awareness, which negatively impacts the quality of generated motions.

To address these issues, this paper proposes a new human interaction generation method, named Human interaction Latent Diffusion Model (HiLDM). The core of this approach lies in three key aspects. First, in order to better capture the interaction and spatial dynamics within human interaction sequences, we introduce a new human interaction representation, i.e., frame-level pose concatenation representation. Second, differing from conventional denoising methods based on raw motion sequences, HiLDM conducts diffusion in latent representations. In other words, denoising is performed in the latent space of human interaction sequences, which significantly improves the generation efficiency. By leveraging a variational autoencoder to learn a low-dimensional latent representation of human interactive motions, this method conducts the diffusion process in a compact latent space. In comparison with direct diffusion on high-dimensional raw motion data, it enables HiLDM to not only reduce substantially computational complexity, but also accelerate inference speed. Lastly, to compensate for the temporal understanding limitations of the CLIP text encoder, we design a Temporal-aware Text Encoder (TTE). Through text-motion contrastive learning, TTE enhances the ability to comprehend implicit temporal and interaction information in text, leading to motions that align better with textual descriptions.

To validate the effectiveness of HiLDM, we conduct extensive experiments on the InterHuman dataset. Experimental results demonstrate that HiLDM outperforms state-of-the-art methods in both generation speed and quality. In terms of generation speed, HiLDM is 57 times faster than ComMDM and 4 times faster than InterGen, significantly improving efficiency. This acceleration makes HiLDM more suitable for practical applications requiring real-time motion synthesis. Regarding generation quality, as evaluated by the Fréchet Inception Distance (FID) metric, HiLDM achieves a 36.7% improvement over ComMDM and a 1.7% improvement over InterGen, indicating that the generated human interactive motions are more natural and adhere more closely to real-world motion dynamics. In qualitative experiments, visual results further demonstrate the superiority of HiLDM in generating complex interactive motions.

Keywords　latent diffusion model; human motion generation; human interaction generation; AIGC

1　引　言

人体动作生成是计算机视觉和计算机图形学中的一个重要研究领域,旨在通过计算机模拟和生成逼真的人类动作。随着虚拟现实、增强现实和电影特效等技术的发展,对于高质量人体动作的需求不断增加。高质量的人体动作生成不仅能够提升用户的沉浸感和互动性,还能够为影视制作、游戏开发、体育训练等领域提供重要的技术支持。为了满足这些需求,近年来出现许多创新的方法和技术,深度学习模型如生成对抗网络(Generative Adversarial Networks,GAN)[1]、变分自编码器(Variational Auto-Encoders,VAE)[2]、扩散模型[3-4]等在人体动作生成中得到了广泛应用。

之前的工作探索了无条件的生成模型[5],以及在各种条件下的生成模型,如动作类别标签[6-7]、文本[8-10]、音乐[11-12]和先验动作[13]等。这些条件生成

模型通过引入额外的信息，使得生成结果更加灵活和可控。例如，基于动作类别标签的生成模型能够生成指定类型的动作，如跳跃、奔跑等；基于文本的生成模型可以通过自然语言描述来生成对应的动作；基于音乐的生成模型则可以根据音乐的节奏和情感生成相应的舞蹈动作；基于先验动作的生成模型则可以利用已有的动作数据生成连续且自然的动作序列。

然而，目前大多数生成模型仅关注于单人动作的生成，忽视了人与人之间复杂多变的互动行为，这导致单人动作生成模型在处理多人互动场景时显得力不从心。与单人动作生成相比，人体交互动作生成面临更多挑战。这类模型不仅需要考虑个体的动作特征，还要捕捉人与人之间的互动和协调，从而生成更加真实、交互自然的人体交互动作。这种复杂的互动行为对生成模型提出了更高的要求，不仅要保证每个个体动作的合理性，还要确保整体互动的连贯性和真实性。在各种条件中，文本包含丰富的语义信息，更易于用户理解和输入，且能够更好地描述人与人之间的交互关系。因此，本文选择使用文本作为条件生成人体交互动作。具体地，本文关注文本引导的人体交互动作生成。

在之前的相关工作中，以生成先验为基础的人体动作扩散模型[14]（Human Motion Diffusion as a Generative Prior，priorMDM）将单人动作扩散模型[9]（Human Motion Diffusion Model，MDM）复制两份，分别生成两个单人动作，并为人体交互动作训练一个简单的通讯块 ComMDM，以注入双人的动作交互。然而，由于缺乏带有文本标注的人体交互动作数据集，priorMDM 采用了单人的动作先验和少量的人体交互动作示例进行训练，导致模型泛化能力较差。对于复杂的人体交互动作，模型容易生成不自然或错误的动作序列。最近，Liang 等人[15]发布了迄今为止最丰富的带有文本标注的人体交互动作数据集 InterHuman，并提出 InterGen 的基线方法。基于扩散模型，InterGen 采用两个共享权重的单人动作去噪器，通过相互注意力机制连接两个去噪过程。得益于 InterHuman 数据集中丰富多样的人体交互动作数据，InterGen 取得了不错的性能。但是，基于扩散模型的人体交互动作生成目前仍然存在很多挑战性的问题。一方面，现有的人体交互动作扩散模型选择在原始动作序列上对单人进行去噪，导致生成速度较慢。另一方面，人体交互动作生成更加复杂，既要保证单人动作的合理性，又要保证动作交互的真实性。现有的动作生成方法普遍沿用文本生成图像领域的对比语言-图像预训练（Contrastive Language-Image Pretraining，CLIP）[16]模型的文本编码器作为动作扩散模型的语言指导。然而，正如 Lu 等人[17]指出，CLIP 文本编码器是在大量的文本-图像对数据集上预训练，对于动作序列和人体交互缺乏足够的理解。这种缺点严重限制了扩散模型生成与文本良好对齐的动作的能力。总体来说主要有两个难点。一方面，现有的人体交互动作生成模型通常采用在原始动作序列上对单个人体进行去噪的方式，导致生成速度较慢。另一方面，现有的动作生成方法普遍沿用 CLIP 文本编码器，对于时序信息缺乏理解，严重限制了模型生成与文本良好对齐动作的能力。

针对人体交互动作扩散模型的上述问题，本文提出一种新的解决方案。其框架包括两个关键模块：人体交互动作潜在扩散模型（Human interaction Latent Diffusion Model，HiLDM）和人体交互时序感知文本编码器（Temporal-aware Text Encoder，TTE）。该解决方案不仅在生成高质量的人体交互动作序列上表现优异，还在生成速度与文本-动作一致性方面有了显著改进。首先，本文提出一种全新的帧级联合人体交互动作表示方法。该表示方法允许通过一个单一的模型同时生成人体交互动作序列，确保两者之间的动作具有良好的互动性和协调性。这种联合表示不同于以往单人的表示方式，更加适用于复杂的人体交互场景。其次，借鉴文本生成图像领域的成功经验，本文引入一种适应人体交互动作生成的潜在扩散模型 HiLDM。潜在扩散模型[18]通过在图像的潜在空间中逐步进行去噪，可以快速生成高质量的图像。本文将这一思想应用到人体交互动作生成领域。通过使用帧级联合人体交互动作表示方法，本文的潜在扩散模型可以仅使用一个去噪器对人体交互动作进行去噪，而无需两个独立的去噪器分别处理两个单人的动作。具体来说，本文设计一个基于 Transformer[19]架构的变分自编码器，以学习人体交互动作序列的潜在分布。同时，基于 Transformer 的去噪网络在人体交互动作的潜在空间中逐步去噪，而非直接作用于原始动作序列。这一设计显著降低了训练和推理阶段的计算复杂度，从而提升了人体交互动作生成的效率。最后，针对 CLIP 文本编码器对动作序列理解不足的问题，特别是人体交互动作生成中对人物之间交互的高度理解需求，本文提出一种人体交互时序感知文本编码

器 TTE。具体地,本文将基于文本的单人动作检索模型(Text-to-Motion Retrieval,TMR)[20]扩展到人体交互动作领域,使用对比学习的方式在文本-人体交互动作数据集上进行预训练。该编码器能够捕捉文本中的细微时序信息,为人体交互动作潜在扩散模型提供具有时序感知的文本特征,从而大大提高生成的人体交互动作与文本描述的一致性。

综上所述,本文的主要贡献总结如下:(1)提出一个简单有效的帧级联合人体交互动作序列表示方式;(2)为了更快地生成人体交互动作,设计一个新的人体交互动作潜在扩散模型 HiLDM;(3)为了增强生成人体交互动作序列与文本的一致性,设计一个具有人体交互时序感知能力的文本编码器 TTE。

2　相关工作

2.1　文本引导的单人动作生成

近年来,人们对人体动作生成产生了浓厚的兴趣。之前的工作已经探索了无条件生成模型[5,21]以及使用各种输入条件的生成模型,如文本[8-10,22-24]、先验动作[13]、动作类别标签[6-7]和音乐[11-12,25]等。文本具有显式或隐式地传达各种类型动作细节的非凡能力,使其成为生成人体动作的良好条件。在本文中,重点关注文本到动作的生成。

早期的工作通常使用序列到序列的模型来解决文本到动作的生成任务[26]。随着 GAN 模型在文本生成图像领域优秀的表现,Text2Action[27]首先利用 GAN 完成文本生成各种动作。Guo 等人[8]发布了一个带有丰富文本描述的动作数据集,并使用 VAE 实现了从文本生成动作,大大提高了生成动作的质量和多样性。TEMOS[28]通过基于 Transformer 的 VAE 学习文本和动作之间的联合分布,从而生成各种动作序列。为了解决传统 VAE 中连续潜在空间难以生成高质量动作的问题,TM2T[29]引入了矢量量化变分自编码器[30](Vector Quantized Variational Autoencoder,VQ-VAE),通过使用离散的向量量化潜在空间,提高了生成动作的质量。T2M-GPT[31]将基于卷积神经网络的 VQ-VAE 与生成式预训练 Transformer(Generative Pre-trained Transformer,GPT)相结合,采用自回归的方式生成动作。MoMask[32]采用分层量化将人体动作表示为具有高保真细节的多层离散动作标记,然后通过掩码 Transformer 进行并行预测。注意到扩散模型在文本生成图像领域取得的显著成功,Tevet 等人[9]首次将扩散模型引入到文本到动作领域中,取得了令人印象深刻的结果。为了提高动作生成模型的泛化能力,Azadi 等人[33]和 Lin 等人[22]尝试利用三维人体重建方法[34-37]来收集大规模伪文本-姿态数据集,然后在这些数据集上进行预训练,可以提高模型的泛化能力。FG-MDM[38]则利用大语言模型将文本标注转述为各个身体部位的细粒度描述,大大提高了模型的泛化能力。

随着大语言模型(Large Language Model,LLM)的出现,文本生成动作领域也涌现出了新的方法。Jiang 等人[39]通过微调 LLM,设计了一个预训练的动作语言模型,可以通过文本提示完成与动作相关的各种任务。Lin 等人[40]利用 LLM 设计了一个全身动作和文本的自动标注管道,能为每个动作自动标注全面的语义标签和帧级的精细化姿态描述。基于这个管道,Lin 等人构建了一个多样性丰富的大型全身动作数据集 Motion-X,包含身体、手和面部动作。Lu 等人[17]基于这个数据集,提出了一种全面分层 VQ-VAE 和一个分层 GPT,用于生成包含身体、手和面部的全身动作。

2.2　文本引导的人体交互动作生成

由于带有文本标注的人体交互动作数据集的稀缺性,导致在该领域的研究工作相对有限。priorMDM[14]采用在 HumanML3D[8]数据集上训练的单人动作生成模型 MDM[9]作为动作先验,并设计了一个轻量级通讯模块 ComMDM,在两个 MDM 模型之间传递交互信息。通过少量的人体交互动作训练样例,该方法能够学习并生成合理的人体交互动作。由于人体交互动作数据集较小,他们的方法效果并不理想。令人兴奋的是,Liang 等人[15]发布了一个迄今为止最丰富的带有文本标注的人体交互动作数据集 InterHuman,包含 6K 个人体交互动作序列和 16K 自然语言描述,序列总时长达到 6.56 小时。Liang 等人还提出了一种基线方法 InterGen,是一种基于扩散模型的人体交互动作生成方法,通过协作网络和共享权重机制来处理两人交互的对称性。此外,InterGen 还引入了新的动作表示方法和交互损失函数,以更好地模拟两人交互中的空间关系,从而生成更自然、更逼真的动作。

3　时序感知潜在扩散模型

本文方法包括三个关键模块:(1)人体交互时序感知文本编码器,其为条件扩散模型提供具有时序

感知的条件信号(3.2 节);(2)人体交互动作变分自编码器,其学习人体交互动作序列在低维空间中的潜在分布(3.3 节);(3)人体交互动作潜在扩散模型,其基于文本条件在低维人体交互动作潜在空间上执行扩散和去噪过程,生成人体交互动作序列(3.4 节)。

3.1 帧级联合人体交互动作表示

在现有的动作表示方法中,通常将人体交互动作序列视为两个独立的单人动作序列,各自进行分析和处理。这种方法虽然在单人动作生成上表现良好,但在处理人体交互动作时,往往容易忽略两者之间的时序关系和空间关系。为了解决这个问题,本文提出了一种帧级联合人体交互动作表示方法,旨在更准确地捕捉人体交互动作序列中的相互作用和空间动态。

对于一个人体交互动作序列 $\boldsymbol{X}$,可以将其表示为两个单人的动作序列组合:$\boldsymbol{X}=\{\boldsymbol{x}_a,\boldsymbol{x}_b\}$。其中 $\boldsymbol{x}_a$ 和 $\boldsymbol{x}_b$ 分别表示两个人 a 和 b 的动作序列。每个单人动作序列又由一系列的姿势状态 x^i 组成,可以表示为 $\boldsymbol{x}_a=\{x_a^i\}_{i=1}^n$,$\boldsymbol{x}_b=\{x_b^i\}_{i=1}^n$,其中 n 表示动作序列的帧数。对于每一帧的姿势,$x^i\in\mathfrak{R}^{J\times D}$,其中 J 表示关节的数量,D 表示关节表示的维度。在 InterHuman[15] 数据集中,采用了一种非规范化的表示方式,以有效表示人体交互过程中人与人之间的空间关系。这种表示方式将关节位置和速度保持在世界坐标系中,而不是像规范化表示那样将它们转换到根节点坐标系中。这样做可以避免累积误差导致的轨迹漂移问题,从而更准确地描述两个人之间的相对位置关系。该非规范化表示为

$$x^i=[j^p,j^v,j^r,c^f] \tag{1}$$

其中,$j^p\in\mathfrak{R}^{J\times 3}$ 表示关节在世界坐标系中的全局位置。$j^v\in\mathfrak{R}^{J\times 3}$ 表示关节在世界坐标系中的全局速度。$j^r\in\mathfrak{R}^{J\times 6}$ 表示关节在根节点坐标系中的旋转,采用 $6D$ 旋转表示。$c^f\in\mathfrak{R}^4$ 表示二值化的脚部接触特征。

在本文中,提出了一种新的帧级联合人体交互动作表示方法,以改进人体交互动作序列的生成。具体而言,考虑一个人体交互动作序列 $\boldsymbol{X}=\{\boldsymbol{x}_a,\boldsymbol{x}_b\}$,其中 $\boldsymbol{x}_a=\{x_a^i\}_{i=1}^n$ 和 $\boldsymbol{x}_b=\{x_b^i\}_{i=1}^n$ 分别表示两个人 a 和 b 的动作序列,将同一帧的双人姿势信息整合为一个整体 $w^i=\{x_a^i,x_b^i\}$,$w^i\in\mathfrak{R}^{2J\times D}$ 的表示形式为

$$w^i=[j_a^p,j_a^v,j_a^r,c_a^f,j_b^p,j_b^v,j_b^r,c_b^f] \tag{2}$$

最终,帧级联合人体交互动作表示为:$\boldsymbol{X}=\{w^i\}_{i=1}^n$。

3.2 人体交互时序感知文本编码器

以 MDM[9] 为代表的一系列文本引导的动作扩散模型,普遍采用了文本-图像模态的 CLIP[16] 文本编码器作为动作扩散模型的语言指导。这类传统的文本嵌入方法本身是为了研究如何在文本和图像模态之间建立联系,因此作为文本生成图像扩散模型的语言指导是合理的选择。然而,在文本生成动作领域,继续沿用文本-图像领域的文本编码器来指导动作扩散模型显然不够合理。CLIP 等传统的文本编码器在大型的文本-图像对数据集上进行预训练,主要关注文本和图像内容的语义理解,而缺乏对动作序列的理解,例如人体动作的含义、顺序性和动态变化等。这使得它们难以有效地捕捉文本描述中与动作相关的信息,导致生成的动作与文本描述之间缺乏一致性。

相比之下,Petrovich 等人[20] 提出的单人动作检索模型 TMR,专门针对文本到单人动作序列的检索任务,能够更准确地捕捉文本中的时序语义信息。TMR 模型通过联合训练文本到动作的检索任务,并采用对比学习策略,构建了一个更精细的跨模态文本-动作空间。这使得模型能够更好地理解文本描述中的动作含义,并将其与动作序列中的关键信息进行匹配。基于其出色的时序感知能力,本文将 TMR 模型扩展到人体交互动作生成领域,并将其称为人体交互时序感知文本编码器 TTE。具体来说,首先对单人动作检索模型 TMR 的模型架构进行调整,使其能够适应帧级联合人体交互动作表示的需求。此外,为了确保生成的人体交互动作在交互性上更加自然、在人体动作学上更加合理,本文引入了单人和双人的几何损失。该损失函数通过约束生成动作的几何结构和相对位置,保证个体动作的姿势连贯性以及人体交互动作的协调性和合理性。这一改进不仅增强了模型对人体交互动作之间复杂交互关系的理解,还提高了生成动作的自然性和动作学准确性,从而使生成的人体交互动作更加符合真实场景中的动态表现。最后,在 InterHuman 的训练集上,使用本文提出的帧级联合人体交互动作表示方式对模型进行对比学习训练。通过这一训练过程,模型能够在人体交互动作序列和对应的文本描述之间建立紧密的跨模态对齐关系。训练完成后,只保留与人体交互动作序列在隐空间内对齐良好的文本编码器,并用其替代原有的在图文隐空间内对齐较好的 CLIP 文本编码器。这样,改进后的文本编码器 TTE 能够更有效地捕捉人体交互动作序列中的时序语义信息,从而提升动作生成的

质量和一致性。

如图 1 所示,TTE 由三部分组成,分别是文本编码器、人体交互动作编码器和人体交互动作解码器。文本编码器使用冻结权重的预训练 DistilBERT[41]模型将文本编码为特征向量序列,然后通过基于 Transformer 的编码器预测高斯分布的均值和方差,经采样得到文本特征 z^T。人体交互动作编码器使用 Transformer[19]对人体交互动作序列 $\boldsymbol{X}=\{w^i\}_{i=1}^n$ 进行编码,经采样得到动作特征 z^M。两个特征被映射到一个共享的隐空间中,通过对比学习,模型能够将语义相似的文本特征和人体交互动作特征映射到空间中靠近的位置,从而实现文本和动作的匹配。在文本和动作编码器中,输入的均值和方差用于提供初始化分布信息,而输出的均值和方差是编码器基于输入特征重新估计和调整的结果,以服务于后续的潜在空间采样和生成任务。人体交互动作解码器则将文本特征 z^T 或者动作特征 z^M 解码为人体交互动作序列 $\overline{\boldsymbol{X}}=\{\overline{w}^i\}_{i=1}^n$。动作解码器的作用是将潜在空间中的特征解码为人体交互动作序列,进而通过重建损失来衡量生成动作与原始动作之间的差异。TTE 可以使模型学习到更全面的文本特征,其损失函数包括对比损失 $\mathcal{L}_{NCE}$、跨模态特征损失 $\mathcal{L}_E$、KL 散度(Kullback-Leibler Divergence)损失 $\mathcal{L}_{KL}$ 和重建损失 $\mathcal{L}_R$。公式如下:

$$\mathcal{L}_{\mathrm{NGE}}=-\frac{1}{2N}\sum_i\left(\log\frac{\exp S_{ii}/\tau}{\sum_j\exp S_{ij}/\tau}+\log\frac{\exp S_{ii}/\tau}{\sum_j\exp S_{ji}/\tau}\right),\quad \text{where}\ 1\leqslant i,j\leqslant N \tag{3}$$

$$\mathcal{L}_E=|z^T-z^M| \tag{4}$$

$$\mathcal{L}_{\mathrm{KL}}=KL(\phi^T,\phi^M)+KL(\phi^M,\phi^T)+KL(\phi^T,\psi)+KL(\phi^M,\psi) \tag{5}$$

$$\mathcal{L}_R=|\boldsymbol{X}-\bar{\boldsymbol{X}}^T|+|\boldsymbol{X}-\bar{\boldsymbol{X}}^M| \tag{6}$$

其中,N 表示批处理中正样本对的数目。S 是一个相似度矩阵,计算批处理中所有文本-动作特征对之间的余弦相似度。τ 是超参数,控制损失函数的平滑程度。$\phi^T=\mathcal{N}(\mu^T,\sum^T)$ 表示文本特征的分布。$\phi^M=\mathcal{N}(\mu^M,\sum^M)$ 表示动作特征的分布。$\psi=N(0,I)$ 是一个标准正态分布。$\boldsymbol{X}$ 表示原始动作序列,$\bar{\boldsymbol{X}}^T$ 表示由文本特征 z^T 解码得到的动作序列,$\bar{\boldsymbol{X}}^M$ 表示由动作特征 z^M 解码得到的动作序列。

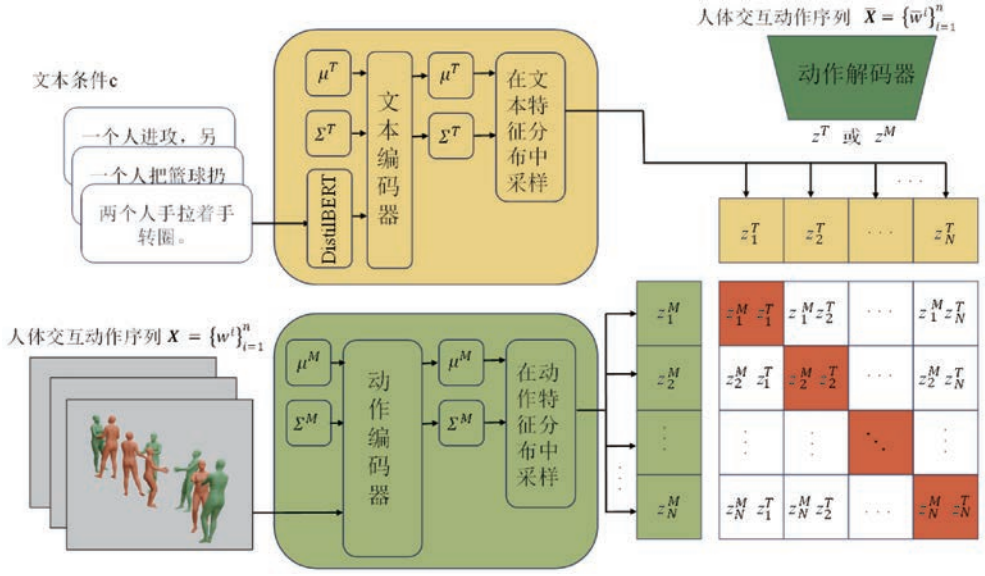

图 1　TTE 框架图(整个模型由文本编码器、动作编码器和动作解码器组成。使用对比学习的方式进行训练,对比学习的目标是最大化相似性矩阵的对角线。)

在对比学习中,相似性矩阵通常用来衡量不同样本之间的相似性。在 TTE 框架中,文本编码器和动作编码器分别生成文本和动作的特征,之后可以计算它们之间的相似性,进而构建相似性矩阵。最大化相似性矩阵的对角线是指通过优化模型,使得相似性矩阵中的对角线元素(即同一对文本和动作之间的相似度)尽可能高。同时,非对角线元素(即不同文本-动作对之间的相似度)则需要尽量较低。通过这种方式,模型学习到的表示不仅能最大化同一对文本和动作的相似度,还能够最小化不同文本-动作对之间的相似度,从而实现有效的对比学习。

实验过程中发现，仅有的重建损失并不能很好地生成令人满意的人体交互动作，生成的动作可能存在脚滑、瞬移、身体错位等各种问题。因此，引入一些人为规定的人体动作学几何损失是有必要的。本文首先对两个单人分别引入了足部损失和速度损失[9]。足部接触损失用于保证生成的动作中足部与地面的接触关系保持合理，避免出现脚部悬空或不自然滑动的情况。速度损失用于控制人体关节的运动速度，确保生成的动作流畅且具有连贯性。然后引入了对两个人的整体的掩码关节距离图(Masked Joint Distance Map，DM)损失和相对方位(Relative Orientation，RO)损失[15]。损失函数如下：

$$\mathcal{L}_{foot}=\frac{1}{n-1}\sum_{i=1}^{n-1}\left\|(FK(\bar{x}^{i+1})-FK(\bar{x}^{i}))\cdot f_i\right\|_2^2 \tag{7}$$

$$\mathcal{L}_{vel}=\frac{1}{n-1}\sum_{i=1}^{n-1}\left\|(x^{i+1}-x^{i})-(\bar{x}^{i+1}-\bar{x}^{i})\right\|_2^2 \tag{8}$$

$$\mathcal{L}_{DM}=\left\|(M(\bar{\boldsymbol{x}}_a,\bar{\boldsymbol{x}}_b)-M(\boldsymbol{x}_a,\boldsymbol{x}_b))\odot I(Mxz(\boldsymbol{x}_a,\boldsymbol{x}_b)<\bar{M})\right\|_2^2 \tag{9}$$

$$\mathcal{L}_{RO}=\left\|O(FK(\bar{\boldsymbol{x}}_a),FK(\bar{\boldsymbol{x}}_b))-O(FK(\boldsymbol{x}_a),FK(\boldsymbol{x}_b))\right\|_2^2 \tag{10}$$

其中，$FK(\cdot)$ 表示将关节旋转转换为关节位置的正向动作学函数。对于每一帧 i，$f_i\in\{0,1\}^J$ 表示二进制脚部接触掩码。$M(\cdot)$ 表示由全局关节位置得到的人体交互关节距离图，$I(\cdot)$ 表示在 xz 平面上应用 2D 距离阈值来掩盖损失的指示函数。$\bar{M}$ 表示距离阈值，$\odot$ 表示哈达玛积。$O(\cdot)$ 表示两个人绕 y 轴旋转得到的相对方位。

最后，训练人体交互时序感知文本编码器总的目标损失如下：

$$\mathcal{L}_{\mathrm{TTE}}=\mathcal{L}_R+\lambda_{\mathrm{NCE}}\mathcal{L}_{\mathrm{NCE}}+\lambda_E\mathcal{L}_E+\lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}+\lambda_{\mathrm{foot}}\mathcal{L}_{\mathrm{foot}}+\lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}}+\lambda_{\mathrm{DM}}\mathcal{L}_{\mathrm{DM}}+\lambda_{\mathrm{RO}}\mathcal{L}_{\mathrm{RO}} \tag{11}$$

其中，$\lambda_{\mathrm{NCE}},\lambda_{\mathrm{E}},\lambda_{\mathrm{KL}},\lambda_{\mathrm{foot}},\lambda_{\mathrm{vel}},\lambda_{\mathrm{DM}},\lambda_{\mathrm{RO}}$ 是平衡系数。

3.3 人体交互动作变分自编码器

为了有效地使用低维特征表示人体交互动作序列，本文使用基于长连接[42]增强的 Transformer 构建了一个人体交互动作变分自编码器。该编码器旨在将人体交互动作编码为低维潜在特征，以便在高信息密度的低维空间中表示高维的动作序列。解码器则将这些低维潜在特征解码为人体交互动作序列，并尽可能地与编码前的动作序列一致。在训练过程中，使用重建损失和 KL 散度损失进行训练。为了进一步提高重建动作的质量与合理性，同样添加了单人和双人的人体动作学几何损失。

如图 2 左侧所示，动作编码器可以将任意长度为 n 的人体交互动作序列 $\boldsymbol{X}=\{w^i\}_{i=1}^n$ 作为输入，得到一个均值为 μ^M，方差为 $\sum^M$ 的正态分布 $\phi^M=\mathcal{N}(\mu^M,\sum^M)$，然后采样得到一个动作特征 z^M。动作解码器通过将这个特征解码得到一个新的动作序列 $\bar{\boldsymbol{X}}=\{\bar{w}^i\}_{i=1}^n$。损失函数包括重建损失、$KL$ 散度损失和人体动作学几何损失，人体动作学几何损失与 3.2 节的损失函数一致，损失函数如下：

$$\mathcal{L}_{R'}=\left|\boldsymbol{X}-\bar{\boldsymbol{X}}\right| \tag{12}$$

$$\mathcal{L}_{\mathrm{KL}'}=KL(\mathcal{N}(\mu^M,\sum^M),\mathcal{N}(0,I)) \tag{13}$$

最后，训练人体交互动作变分自编码器总的目标损失如下：

$$\mathcal{L}_{VAE}=\mathcal{L}_{R'}+\lambda_{\mathrm{KL}'}\mathcal{L}_{\mathrm{KL}'}+\lambda_{\mathrm{foot}}\mathcal{L}_{\mathrm{foot}}+\lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}}+\lambda_{\mathrm{DM}}\mathcal{L}_{\mathrm{DM}}+\lambda_{\mathrm{RO}}\mathcal{L}_{\mathrm{RO}} \tag{14}$$

其中，$\lambda_{R'},\lambda_{\mathrm{KL}'},\lambda_{\mathrm{foot}},\lambda_{\mathrm{vel}},\lambda_{\mathrm{DM}},\lambda_{\mathrm{RO}}$ 是平衡系数。

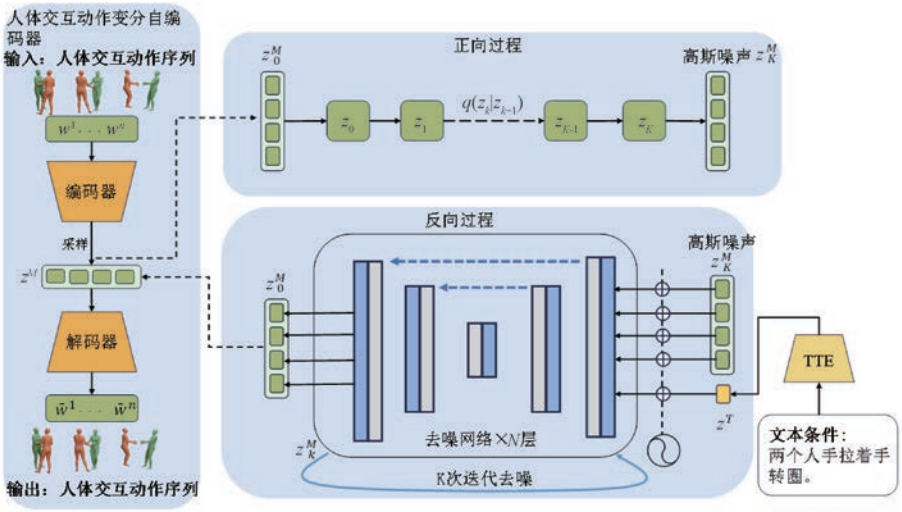

图 2 HiLDM 框架图(左侧是人体交互动作变分自编码器，右侧是人体交互动作潜在扩散模型的主体结构。)

3.4　人体交互动作潜在扩散模型

扩散模型[3-4]的基本思想是学习一个定义良好的随机过程的逆向过程,该过程由正向过程和反向过程组成,两者均遵循马尔可夫链。正向过程涉及添加噪声。输入是潜在空间中的动作特征 z_0^M,输出是添加了 k 次高斯噪声的动作特征 z_k^M。当添加足够的噪声时,动作特征 z_K^M 可以接近标准高斯分布 $\mathcal{N}(0,\mathrm{I})$。反向过程旨在减少高斯噪声 $z_K^M \sim \mathcal{N}(0, I)$ 中的噪声。在去噪过程中,在扩散步骤 k,一部分噪声被消除,从而产生噪声较小的动作特征 z_{k-1}^M。通过迭代地重复此步骤,直到噪声完全消除,生成一个干净的动作特征 $\bar{z}_0^M$。

如图 2 右侧所示,只保留 TTE 的文本编码器部分,为人体交互动作潜在扩散模型提供具有时序感知的文本特征 z^T。本文使用 Transformer 编码器架构实现去噪网络,在文本条件引导下,学习在时间步长 k 的动作特征 z_k^M 到时间步长 $k-1$ 的动作特征 z_{k-1}^M 的去噪过程。在扩散模型的采样过程中,从标准正态分布中对初始随机噪声 z_K^M 进行采样,然后进行 K 次去噪以生成干净动作特征 $\bar{z}_0^M$。最后,通过人体交互动作解码器将其解码为帧数为 n 的人体交互动作序列 $\bar{\boldsymbol{X}}=\{\bar{w}^i\}_{i=1}^n$。

最后,训练人体交互动作潜在扩散模型总的目标损失如下:

$$\mathcal{L}_{\mathrm{LDM}}=E_{z_0^M\sim q(z_0^M|c),k\sim[1,K]}\left[\left\|z_{k-1}^M-G(z_k^M,k,c)\right\|^2\right] \tag{15}$$

其中,c 表示文本条件,z_0^M 表示一组从条件分布 $q(z_0^M\mid c)$ 中采样得到的真实动作序列。$k\sim[1,K]$ 表示从时间步长 1 ～ K 中随机采样的一个时间点,K 表示扩散模型的总时间步数,$G(z_k^M,k,c)$ 是模型预测的在 k 时间步去噪后的动作特征。

4　实验结果与分析

在本节中,将详细介绍本文所使用的数据集、评估指标、实验设置,以及对提出的方法 HiLDM 进行定量与定性实验评估,并与当前最先进的技术方法进行深入比较。首先,4.1 节中对人体交互动作数据集 InterHuman 进行了详细描述,包括其构成、标注方式及使用比例。接下来,4.2 节介绍了使用的评估指标,通过使用多种评估指标,确保从多个角度全面评估模型的性能。随后,4.3 节介绍了实验的具体设置,包括模型的训练参数、模型大小等。在 4.4 节首先进行了定量实验。将本文提出的 HiLDM 方法与当前最先进的人体交互动作生成方法进行比较。然后,进行定性实验,通过可视化展示生成的动作序列与最先进的人体交互动作生成方法进行对比。最后,在 4.5 节进行了消融实验,以评估不同模块和组件对最终结果的贡献。

4.1　数据集介绍

本文采用 InterHuman[15]数据集对本方法进行训练和评估。InterHuman 数据集是一个专为人体交互动作生成任务设计的大规模数据集,包含丰富的动作类型和高质量的标注信息。该数据集涵盖日常动作和专业动作两大类,涉及 6022 个人体交互动作序列,总时长达到 6.56 小时,充分展示人与人之间不同的交互模式与动态变化。为了便于自然语言驱动的动作生成任务,每个动作序列都提供多个自然语言描述,这些描述由不同的标注人员从多个角度进行标注,进一步增强数据的多样性和表达的准确性。数据集中共计有 16756 条文本描述,涵盖 5656 个独特词汇。将数据集的 80%用作训练集,20%用作测试集,确保模型在多种人体交互动作场景下的学习与泛化能力。这一数据集的多样性和规模为本研究的训练和评估提供坚实基础。

4.2　评价指标介绍

本文采用与 Liang 等人[15]相同的五种评估指标来评估模型的能力:R Precision、FID、Multimodal Dist、Diversity 和 Multimodality。R Precision 和 Multimodal Dist 用于评估生成动作与输入文本之间的相关性。Diversity 衡量生成样本之间的多样性。Multimodality 衡量的是同一文本输入或同一动作输入下生成结果的多样性。FID 用于测量生成动作与真实数据在潜在空间中特征分布的差异,从而评估生成动作的质量。

4.3　实验设置

在人体交互时序感知文本编码器 TTE 部分,文本编码器由 6 层和 4 个头的 Transformer 构成。人体交互动作编码器和解码器均由 6 层和 4 个头的 Transformer 构成,动作序列的维度为 $524\times n$,动作和文本特征的维度均为 512。批量大小设置为 16。模型使用 AdamW 优化器,10^{-4} 的固定学习率进行训练。在单个 NVIDIA GeForce RTX 3090 GPU 上训练模型大约需要 1 天时间。

人体交互动作潜在扩散模型 HiLDM 使用 TTE 阶段训练完成冻结后的文本编码器。人体交

互动作编码器和解码器均由 9 层和 8 个头的 Transformer 构成，动作序列的维度为 $524 \times n$，中间特征维度为 512。去噪器由 9 层和 8 个头的 Transformer 构成，去噪总次数 K 设置为 50。加噪过程采用线性噪声策略，去噪采用非马尔可夫的采样策略[43]，去噪过程中 guidance 强度设置为 7.5，生成动作序列的最大长度为 300。在人体交互动作变分自编码器学习阶段批处理大小为 32，在人体交互动作潜在扩散模型学习阶段批处理大小为 64。所有模型使用 AdamW 优化器，10^{-4} 的固定学习率进行训练。在两个 NVIDIA GeForce RTX 3090 GPU 上训练模型，人体交互动作变分自编码器训练阶段大约需要 2 天时间，人体交互动作潜在扩散模型训练阶段大概需要 0.5 天时间。

4.4 定量实验

为了评估本文方法在文本条件下人体交互动作生成方面的性能，将本文方法与最近的 5 种动作生成方法进行了比较：TEMOS[28]、T2M[8]、MDM[9]、ComMDM[14]和 InterGen[15]。在表 1 中，给出了在 InterHuman 数据集上的实验结果，对于 5 种 SOTA 方法，本文引用 Liang 等人[15]报告的结果。对于本文方法的结果，进行 5 次评估，取平均值。由表 1 可以看出，在处理基于文本的人体交互动作生成任务时，本方法取得了最先进的性能。特别是，在 Multimodal Dist 指标上，本文方法相较于 InterGen 方法有了显著的改进，从 5.108 提升至 3.827，证明本文方法生成的动作与输入文本具有更好的一致性。

如表 1 所示，本文方法在时序感知能力和生成动作的精确度上表现突出，特别是在 MM Dist 和 FID 指标上优于其他方法。这表明模型能够更准确地捕捉输入文本的信息，生成的动作更加符合文本的预期，且生成动作的质量更接近真实数据。特别的，本文方法在 Multimodality 指标上的相对较低，这可能是由于本文方法强大的时序感知能力，模型在统一文本条件下更倾向于生成一致性较高的动作序列，这牺牲了生成动作的多样性和变化性。这一现象表明本文方法在生成精确、一致的动作方面具有优势，但在多模态性上存在一定的权衡。

表 1 InterHuman 数据集上的定量结果

Method	R Precision ↑			FID ↓	MM Dist ↓	Diversity →	Multimodality ↑
	Top 1	Top 2	Top 3				
Real	0.452	0.610	0.701	0.273	3.755	7.948	—
TEMOS[28]	0.224	0.316	0.450	17.375	6.342	6.939	0.535
T2M[8]	0.238	0.325	0.464	13.769	5.731	7.046	1.378
MDM[9]	0.153	0.260	0.339	9.167	7.125	7.602	**2.355**
ComMDM[14]	0.223	0.334	0.466	7.069	6.212	7.244	1.822
InterGen[15]	0.371	0.515	**0.624**	5.918	5.108	7.387	2.141
HiLDM	**0.377**	**0.517**	0.598	**5.801**	**3.827**	**7.898**	0.846

注："Real"行所列指标值为真值，粗体表示不同方法中取得的最优指标值，"↑"表示指标值越高越好，"↓"表示指标值越低越好，"→"表示指标值越接近真值越好。下表同。

为了评估不同方法的平均推理时间及 FID 得分的比较，如图 3 所示，从 InterHuman 数据集中随机选取 10 个文本，分别使用各个方法生成动作序列，并计算生成单个动作序列的平均耗时。图中的每种颜色的点代表一种方法，X 轴表示单个文本的平均推理时间（越小越好），Y 轴表示衡量生成动作质量的 FID 指标（越小越好）。所有实验均在单个 NVIDIA GeForce RTX 3090 GPU 上进行推理，忽略了模型加载时间。由图 3 可以看出，TEMOS 和 T2M 方法生成速度较快，但生成质量较差；MDM 和 ComMDM 生成质量较好，但生成速度较慢。本文方法在 FID 上领先所有其他方法的同时，单个动作的平均生成时间为 0.38 秒，显著快于 InterGen 方法的 1.86 秒和 ComMDM 方法的 23.15 秒。

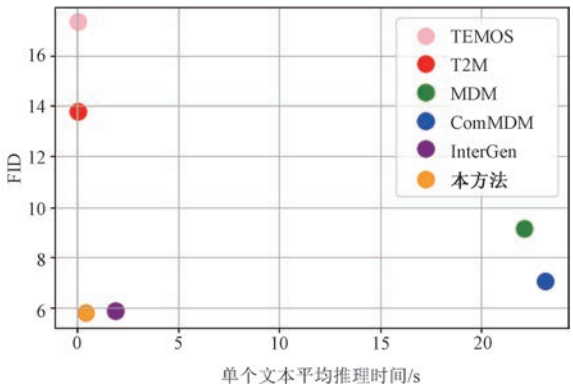

图 3 不同方法在 FID 得分与单个文本平均推理时间之间的关系（图中每种颜色的点代表一种方法，这些点越接近原点表现越好）

4.5 定性实验

如图 4、图 5 和图 6 所示，展示了由 ComMDM

方法、InterGen 方法生成的人体交互动作和本方法生成的人体交互动作的可视化比较。在图 4 中,文本描述为:“第一个人把绳子拉向自己,而第二个人试图把它拉开。最终,第一个人在拔河比赛中获胜。”对比生成结果,ComMDM 方法生成的动作缺乏细节和连贯性,整体效果不如其他两种方法。其他两种方法均能够捕捉拔河的基本动作特征,但本方法在动作细节呈现和整体连贯性方面优于 InterGen 方法。特别是在对抗阶段,InterGen 方法中橙色人物的动作缺乏显著的抵抗表现,整体动作序列缺少力量对抗的张力感。而本方法能够细腻刻画对抗细节,例如橙色人物的拉绳动作不仅与绿色人物形成明确的对抗关系,还展现了双手用力拉开的动态,增强了动作的真实性和互动性。在图 5 中,文本描述为:“两个人手拉着手转圈。” ComMDM 方法在此文本条件下的生成动作效果较差,无法准确展现手拉手转圈的动作。InterGen 方法生成的动作中,两个人物虽然尝试手拉手转圈,但手部接触点

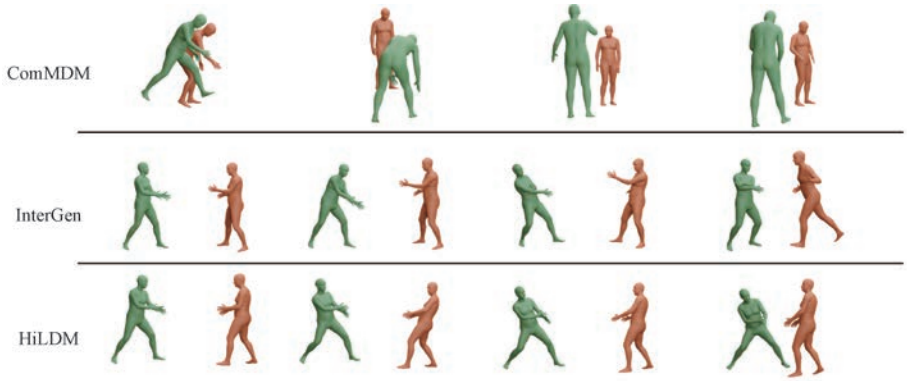

图 4　ComMDM 方法、InterGen 方法与 HiLDM 的可视化比较一(文本描述:第一个人把绳子拉向自己,而第二个人试图把它拉开。最终,第一个人在拔河比赛中获胜。)

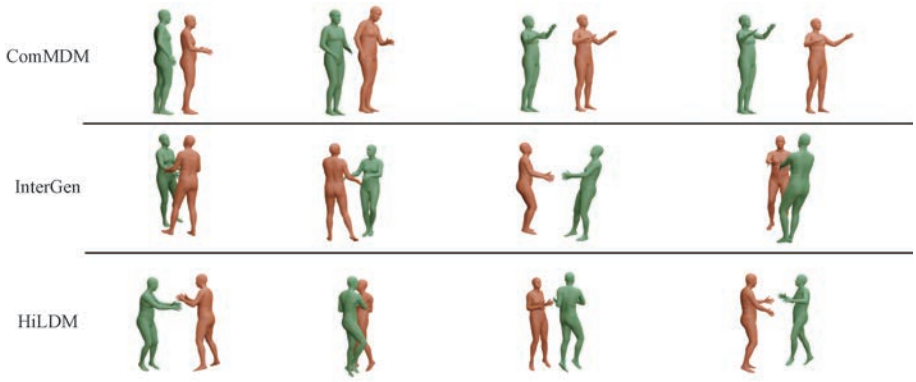

图 5　ComMDM 方法、InterGen 方法与 HiLDM 的可视化比较二(文本描述:两个人手拉着手转圈。)

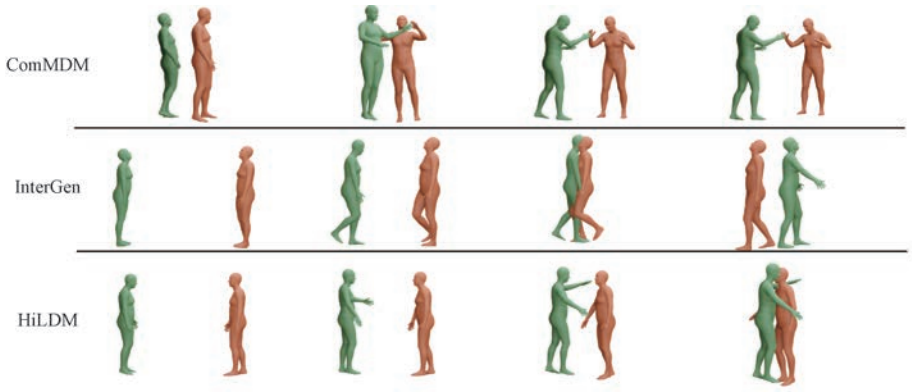

图 6　ComMDM 方法、InterGen 方法与 HiLDM 的可视化比较三(文本描述:两个人拥抱。)

的间距较大，且整体姿态缺乏转圈的动态感，显得较为僵硬。而本方法生成的动作序列表现出更符合文本描述的特征。绿色人物和橙色人物的手部保持紧密接触，展现出稳定的牵手状态。同时，两个人物在转圈时的身体姿态符合旋转运动特征，例如膝盖微弯、身体倾斜角度符合离心力影响，使得整体动作更加自然、生动。在图 6 中，文本描述为："两个人拥抱。"ComMDM 方法的生成效果不如 InterGen 方法，特别是人物之间的距离感没有得到很好的表现。InterGen 方法生成的动作序列虽然表现出拥抱的意图，但存在明显的穿模问题，例如两个人物在靠近时，身体发生互相穿透，尤其是手臂与身体交叠，导致生成结果缺乏真实感和合理性。同时，绿色人物的姿态不自然。而本方法生成的动作序列能够清晰展现拥抱的情感互动，绿色人物和橙色人物的动作幅度及姿态合理，且身体和手臂没有出现穿模现象，更符合拥抱这一情境的语义。

相比之下，本方法在动作细节呈现、连贯性以及符合文本描述的动作特征方面优于 InterGen 方法。特别是在表现动态运动和人物穿模方面，本方法展现出了更高的细腻度和合理性。

4.6 消融实验

为验证本方法各个模块的有效性，本文对各个模块进行了消融实验。实验通过逐步去除或替换模型中的不同组件，来考察每个设计的贡献，并突出本方法的创新之处。如表 2 所示，第一行表示基线方法，人体交互动作变分自编码器阶段的中间特征维度为 256，使用由 9 层和 4 个头组成的 Transformer 结构，且未加入人体几何损失，文本编码器为 CLIP-ViT-L-14。第二行表示特征维度增至 512，Transformer 结构调整为 9 层和 8 个头的情况下的实验结果。当我们将特征维度从 256 增大到 512 时，模型的生成质量有了明显提升。FID 从 12.422 降低至 6.150，MM Dist 从 5.684 下降到 3.882。这说明在人体交互动作变分自编码器中，更高的特征维度为动作序列提供了更丰富的表示，增强了模型对人体交互动作的理解能力。第三行展示了在模型中加入单人和双人人体几何损失后的实验结果。在表 3 中进一步讨论了单人和双人几何损失的作用。加入单人几何损失相较于没有几何损失时，FID 略有增加。MM Dist 略有上升，但多样性略有下降。引入双人几何损失后，生成动作在交互的协调性上得到了提升，FID 值略有改善。进一步增加了全部几何损失，可以观察到生成动作的多样性有了显著提高。第四行则表示将 CLIP 文本编码器替换为 TTE 文本编码器的实验结果。当使用 TTE 文本编码器替换 CLIP 文本编码器后，R Precision(Top 3)从 0.552 提高到 0.598，FID 从 6.265 降低至 5.801，表明 TTE 文本编码器能够显著提升生成结果的文本一致性和生成动作质量。TTE 通过捕捉文本中的时序信息，使得模型更好地理解人体交互动作序列中的交互和时序变化，进而生成与文本描述更为一致的人体交互动作序列。然而，也许是 TTE 文本编码器提供的条件特征更具有确定性，导致多模态距离 Multimodality 指标的下降。这表明 TTE 编码器在时序信息的捕捉和理解方面具有明显优势，能够更精准地感知文本中的时序关系，并生成与文本内容高度一致的动作。然而，由于 TMR 对时序信息的敏锐捕捉，模型倾向于生成较为单一且一致的动作，从而导致 Multimodality 指标上的表现较差。因此，本方法最终选择了 TTE 作为扩散模型的文本编码器。

表 2 InterHuman 数据集上的消融实验

VAE 特征维度	人体几何损失	文本编码器	R Precision (Top 3) ↑	FID ↓	MM Dist ↓	Diversity→	Multimodality ↑
256	无	CLIP	0.382	12.422	5.684	7.362	1.232
512	无	CLIP	0.564	6.150	3.882	7.509	**1.605**
512	有	CLIP	0.552	6.265	3.872	**7.922**	1.322
512	有	TTE	**0.598**	**5.801**	**3.827**	7.898	0.846

注：所有模型在 InterHuman 数据集上进行训练和测评。

表 3 InterHuman 数据集上的消融实验(人体几何损失)。

损失函数	FID ↓	MM Dist ↓	Diversity→
无人体几何损失	**6.150**	3.882	7.509
有单人几何损失	6.325	3.877	7.802
有双人几何损失	6.142	**3.852**	7.786
有全部几何损失	6.265	3.872	**7.922**

如图 7 和图 8 所示，本文继续研究了使用 TTE 文本编码器与 CLIP 文本编码器的扩散模型，在生成人体交互动作序列时的可视化结果。在图 7 中，文本描述为："一个人把篮球扔给另一个人，另一个人站着牢牢抓住它。"对于 CLIP 引导生成的动作，绿色人物和橙色人物的互动较弱，难以清晰地展现"篮球传递"

的关键动作特征。橙色人物的扔球动作缺乏明显的手臂挥动和投掷动态,仅表现出模糊的手部移动。绿色人物腿部姿势僵硬,未能体现抓住篮球的姿态。相比之下,TTE 引导生成的动作能够准确表现“篮球传递”的核心动态。橙色人物在传球时手臂有明显的投掷动作,动作幅度符合传递篮球的特征。绿色人物则及时作出接球动作,腿部微微弯曲稳定核心。在图 8 中,文本描述为:“一个人进攻,另一个人防守。”对于 CLIP 引导生成的动作,防守方绿色人物的动作幅度较小,姿态表现出轻微的站立倾向,未展现出防守应有的紧张和积极性。进攻方橙色人物虽有动作,但手部和身体的动态未能体现进攻的明确性,整体动作显得较为被动。TTE 引导生成的动作中,绿色人物的阻挡动作随着橙色人物的进攻动作而及时调整。橙色人物手部和身体呈现出强烈的进攻意图,特别是朝向绿色人物的进攻角度不停变化。

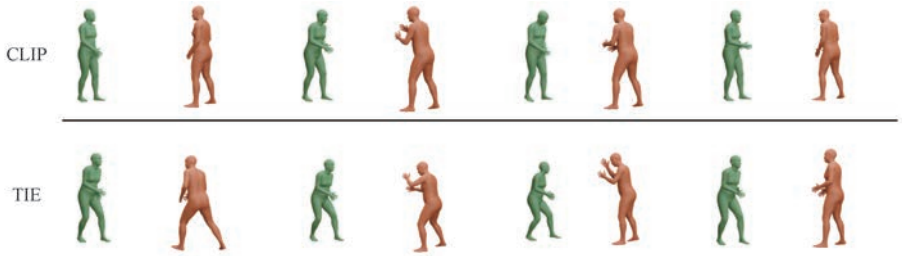

图 7　CLIP 与 TTE 的可视化比较一(文本描述:一个人把篮球扔给另一个人,另一个人站着牢牢抓住它。)

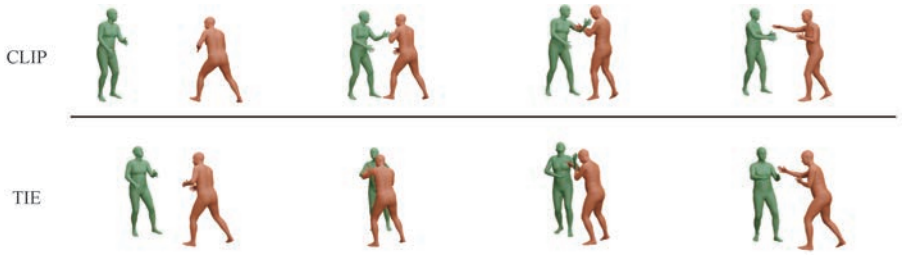

图 8　CLIP 与 TTE 的可视化比较二(文本描述:一个人进攻,另一个人防守。)

正如所观察到的,TTE 引导的扩散模型生成的动作序列与文本条件的匹配度更高,并且 CLIP 引导的扩散模型在某些情况下倾向于生成相对静止的动作序列,尤其是在复杂的交互场景中,动作特征表现不够明确,难以展现出语义上的攻防关系或动态变化。

5　结　语

为了解决当前文本条件下人体交互动作生成过程中存在的生成速度慢,以及生成动作与文本描述一致性差的问题,本文提出了一种新颖的时序感知的人体交互动作潜在扩散模型。该模型通过在人体交互动作序列的潜在空间中进行去噪,显著提升了生成速度。此外,引入人体交互时序感知文本编码器 TTE,有效增强了文本与生成动作之间的一致性与生成动作的质量。最后,实验结果表明,所提出的方法在生成速度和质量方面优于现有方法,验证了该方法在人体交互动作生成任务中的优越性。

然而,本文方法仍存在一定局限性。例如,在一些交互性较强的动作场景中,生成的人体交互动作可能会出现人与人之间动作不完全对齐的情况,尤其是在动作快速变化的情境下。这表明当前模型在捕捉细粒度的交互关系和时序一致性方面仍有待提升。因此,未来的工作可以从以下几个方面进行改进:首先,进一步增强时序建模能力,特别是在处理快速变化和高交互性的人体交互动作时,提高模型对细粒度交互的捕捉能力;其次,可以探索更为复杂的模型架构或算法,以平衡多样性与一致性的要求,从而提升生成效果的灵活性和丰富性。

作者贡献说明　石旭和孙运莲为共同第一作者。

致　谢　感谢编辑和审稿人对本论文的审阅与宝贵建议。感谢上海科技大学虞晶怡教授团队公开 InterHuman 数据集,为本研究提供了重要支持。

参　考　文　献

[1]　Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative

adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014, 27

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

[3] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2256-2265

[4] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2020, 33: 6840-6851

[5] Raab S, Leibovitch I, Li P, et al. Modi: Unconditional motion synthesis from diverse data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 13873-13883

[6] Guo C, Zuo X, Wang S, et al. Action2motion: Conditioned generation of 3d human motions//Proceedings of the ACM International Conference on Multimedia. Seattle, USA, 2020: 2021-2029

[7] Petrovich M, Black M J, Varol G. Action-conditioned 3d human motion synthesis with transformer vae//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10985-10995

[8] Guo C, Zou S, Zuo X, et al. Generating diverse and natural 3d human motions from text//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 5152-5161

[9] Guy Tevet et al. Human motion diffusion model//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023

[10] Chen X, Jiang B, Liu W, et al. Executing your commands via motion diffusion in latent space//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18000-18010

[11] Li R, Yang S, Ross D A, et al. Ai choreographer: Music conditioned 3d dance generation with aist++//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 13401-13412

[12] Tseng J, Castellon R, Liu K. Edge: Editable dance generation from music//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 448-458

[13] Martinez J, Black M J, Romero J. On human motion prediction using recurrent neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2891-2900

[14] Shafir Y, Tevet G, Kapon R, et al. Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418, 2023

[15] Han Liang, Wenqian Zhang, Wenxuan Li, et al. Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision, 2024,132:3463-3483

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 8748-8763

[17] Lu S, Chen L H, Zeng A, et al. Humantomato: Text-aligned whole-body motion generation//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 32939-32977

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

[19] Vaswani A, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017:5998-6008

[20] Petrovich M, Black M J, Varol G. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 9488-9497

[21] Zhao R, Su H, Ji Q. Bayesian adversarial human motion synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6225-6234

[22] Lin J, Chang J, Liu L, et al. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 23222-23231

[23] Kim J, Kim J, Choi S. Flame: Free-form language-based motion synthesis & editing//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 8255-8263

[24] Tevet G, Gordon B, Hertz A, et al. Motionclip: Exposing human motion generation to clip space//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 358-374

[25] Le N, Pham T, Do T, et al. Music-driven group choreography//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 8673-8682

[26] Lin A S, Wu L, Corona R, et al. Generating animated videos of human activities from natural language descriptions. Learning, 2018, 1(2018): 1

[27] Ahn H, Ha T, Choi Y, et al. Text2action: Generative adversarial synthesis from language to action//Proceedings of the Engineers International Conference on Robotics and Automation. Brisbane, Australia, 2018: 5915-5920

[28] Petrovich M, Black M J, Varol G. TEMOS: Generating diverse human motions from textual descriptions//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 480-497

[29] Guo C, Zuo X, Wang S, et al. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 580-597

[30] Van Den Oord A, Vinyals O. Neural discrete representation learning//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017:6306-6315

[31] Zhang J, Zhang Y, Cun X, et al. Generating human motion from textual descriptions with discrete representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14730-14740

[32] Guo C, Mu Y, Javed M G, et al. Momask: Generative masked modeling of 3d human motions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 1900-1910

[33] Azadi S, Shah A, Hayes T, et al. Make-an-animation: Large-scale text-conditional 3D human motion generation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15039-15048

[34] Tian Y, Zhang H, Liu Y, et al. Recovering 3d human mesh from monocular images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15406-15425

[35] Zhang H, Tian Y, Zhou X, et al. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11446-11456

[36] Zhang H, Tian Y, Zhang Y, et al. Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12287-12303

[37] Yao W, Zhang H, Sun Y, et al. STAF: 3d human mesh recovery from video with spatio-temporal alignment fusion. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11):10564-10577

[38] Shi X, Yao W, Luo C, et al. FG-MDM: Towards zero-shot human motion generation via chatgpt-refined descriptions. International Conference on Pattern Recognition. Kolkata, India, 2024: 446-461

[39] Jiang B, Chen X, Liu W, et al. Motiongpt: Human motion as a foreign language//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 20067-20079

[40] Lin J, Zeng A, Lu S, et al. Motion-x: A large-scale 3d expressive whole-body human motion dataset//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 36: 25268-25280

[41] Sanh V. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019

[42] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. Munich, Germany, 2015:234-241

[43] Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020

SHI Xu, master candidate. His research interests include human motion genneration.

SUN Yun-Lian, Ph. D., associate professor. Her research interests include pattern recognition and computer vision.

LUO Yan-Lin, Ph. D., professor. Her research interests include virtual reality and visualization.

ZHANG Hong-Wen, Ph. D., associate professor. His research interests include computer vision and computer graphics.

Background

This paper addresses the problem of human interaction generation within the fields of computer vision and computer graphics. Compared to single-person motion generation, human interaction generation is still in its early stages, particularly when it comes to generating complex interactive motions. It not only requires the individual motion to be plausible but also demands consistency and naturalness in the interaction, making the generation task significantly more challenging. Recently, diffusion models have gained increasing attention in the field of motion generation. However, most of these methods focus on denoising the original motion sequence, which results in lower generation speed. Moreover, existing motion diffusion models typically rely on text encoders derived from pre-trained language-image contrastive models for linguistic guidance. Although these text encoders perform well in the image domain, they fail to adequately capture the temporal dynamics of motion text. This limitation is particularly pronounced in human interaction generation scenarios, where the absence of temporal information further degrades the quality and coherence of the generated motions.

To address these challenges, this paper proposes a Hu-

man interaction Latent Diffusion Model (HiLDM), which performs denoising in the latent space of learned interactive motion sequences, significantly improving generation speed. In addition, the model employs a human interaction Temporal-aware Text Encoder (TTE) as linguistic guidance, ensuring greater temporal consistency in the generated motions. Experimental results demonstrate that, compared to state-of-the-art methods, our model not only achieves a significant advantage in generation speed but also delivers superior performance in terms of motion quality and interaction naturalness, validating the effectiveness and superiority of our approach.

This paper was supported by the National Natural Science Foundation of China (Nos. 62476131, 62076131, 62377004), the Fundamental Research Funds for the Central Universities of Ministry of Education of China (No. 2233100028).