https://crad.ict.ac.cn
63(4): 943−959, 2026

人与场景交互运动生成技术综述

景宇宸[1,2] 温玉辉[2] 李晨光[2] 于 剑[2]

[1](先进轨道交通自主运行全国重点实验室(北京交通大学) 北京 100044)

[2](交通数据挖掘与具身智能北京市重点实验室(北京交通大学) 北京 100044)

(25110183@bjtu.edu.cn)

Review of Human-Scene Interactive Motion Generation

Jing Yuchen[1,2], Wen Yuhui[2], Li Chenguang[2], and Yu Jian[2]

[1](*State Key Laboratory of Advanced Rail Autonomous Operation* (*Beijing Jiaotong University*), *Beijing* 100044)

[2](*Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence* (*Beijing Jiaotong University*), *Beijing* 100044)

Abstract The task of generating human–scene interaction motions involves multiple disciplines, including computer vision, computer graphics, and robotics. The primary goal of this task is to leverage deep learning algorithms to model and learn the relationships between humans and scenes from large-scale interaction motion data, thereby generating diverse human motions interacting with indoor scenes or objects within them. These motions include, for instance, obstacle-avoidance navigation, human–chair interaction, and object grasping. Compared with traditional physics-based simulation approaches, data-driven methods for interaction motion generation are free from reliance on physical simulation engines, providing higher computational efficiency and stronger generalization capability. As a result, these methods hold broad application prospects in areas such as game design, film production, and human–computer interaction. Nevertheless, current research on human–scene interaction motion generation has not yet formed a systematic review or synthesis. To address this gap, we systematically organize and explicate the core advancements in this domain. Specifically, we first introduce data representation methods for 3D humans and scenes. Building on this foundation, we then summarize the different types of interaction tasks and the associated technical challenges, while also presenting the key characteristics of relevant benchmark datasets and evaluation metrics. Finally, we highlight the limitations of existing technical approaches and discuss potential breakthrough directions and development paths for future research.

Key words human motion generation; interaction data; scene interaction; data driven; literature review

摘 要 人与场景交互运动生成任务涉及计算机视觉、计算机图形学和机器人控制学等多个领域。该任务旨在利用深度学习算法从大量交互运动数据中建模并学习人与场景的交互运动关系,生成人与室内场景或其中物体的各种交互运动,包括避障漫游、人椅交互和物体抓取等。相比于传统物理仿真方法,基于数据驱动的交互运动生成方法摆脱了对物理仿真引擎的依赖,具有更高的计算效率和更强的泛化能力,在游戏设计、影视制作和人机交互等领域具有广泛的应用前景。然而,当前对人与场景交互运动生成技

收稿日期: 2025-05-26; **修回日期**: 2025-10-16

基金项目: 国家重点研发计划项目 (2024YFE0202900); 国家自然科学基金项目 (62436001); 北京市科技计划项目 (Z231100005923029); 北京交通大学人才基金项目 (2023XKRC045)

This work was supported by the National Key Research and Development Program of China (2024YFE0202900), the National Natural Science Foundation of China (62436001), the Beijing Science and Technology Plan Project (Z231100005923029), and the Beijing Jiaotong University Talent Foundation (2023XKRC045).

通信作者: 温玉辉(yhwen1@bjtu.edu.cn)

术的研究尚未形成系统性归纳,系统梳理与阐述当前人体与场景交互运动生成技术的核心进展。首先阐释3维人体与场景的数据表示方法;其次在此基础上系统地归纳不同交互任务类型及技术挑战,详述相关基准数据集的核心特征及评估指标体系;最后总结现有技术路线的局限性并分析未来研究的突破方向与潜在发展路径。

关键词 人体运动生成;交互数据;场景交互;数据驱动;研究综述

中图法分类号 TP391.9; TP391

DOI: 10.7544/issn1000-1239.202550352 **CSTR:** 32373.14.issn1000-1239.202550352

随着深度人体运动生成模型[1-4]快速发展,数据驱动的人与场景交互运动生成技术成为研究热点。其核心在于通过构建基于大量交互运动数据的深度学习算法,实现对人与场景交互运动的数据分布建模,从而生成符合物理规律与场景约束的交互运动。生成的动作不仅涵盖基础人体运动,还需要对场景的复杂交互进行建模,例如抓取操作、物体搬运和场景导航等。相比于传统的物理仿真方法[5-9],基于数据驱动的深度学习方法具有更高的计算效率,能够从大量数据中学习到丰富的交互模式,具有更强的跨场景泛化能力,并且能够生成更多样化的交互动作。本文对现有数据驱动人与场景交互运动生成技术进行整理归纳,根据时间维度将基于场景条件的运动生成任务分为帧级静态交互姿态生成[10-14]和序列级交互运动生成[15-19]2类。帧级静态交互姿态生成方法通过场景几何编码对人体姿态和空间位置进行建模,实现基于场景信息的人体姿态推理。相比于帧级静态交互姿态生成,序列级交互运动生成需同时满足时序连续性与物理合理性约束。已有方法主要基于关键帧生成与运动插值的级联架构,根据场景信息和交互标签生成关键帧姿态,进而利用基于条件输入的人体运动生成技术,图1展示了基于关键帧插值的运动补全生成。本文将对基于场景的序列级交互运动生成方法进行梳理和归纳。

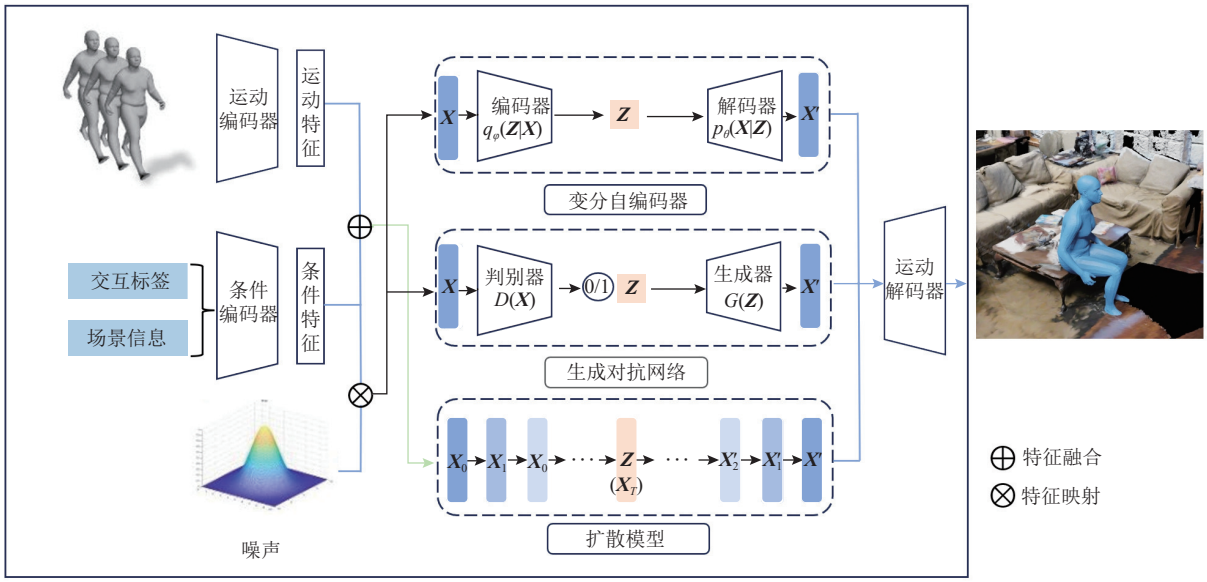


Fig. 1 Conditional human motion generation method

图1 条件人体动作生成方法

此前已有研究[20-21]对该领域进行阶段性总结。例如,文献[20]从仿真平台、交互数据、交互感知和模型算法4个角度对室内场景的拟人交互研究进行总结,但其偏重物理仿真方法。文献[21]围绕文本、音频和场景等多模态条件引导的人体运动生成展开综述,但受限于研究范畴的广泛性,未能深入分析基于场景交互的人体动作生成方法。本研究聚焦数据驱动的人与场景交互运动生成领域,系统地梳理现有方法,并围绕模型架构设计和任务分类展开讨论。本文的主要贡献有3点:

1)从数据表示方法的角度,系统梳理了人与场景交互运动生成任务中常用的3维数据表征形式,

阐明不同表示方法对运动生成效果的影响。

2)针对数据驱动的人与场景交互运动生成方法，结合近年研究成果，分别从生成模型架构与交互任务类型的视角进行分类综述，并梳理该领域技术的发展脉络。

3)系统归纳现有交互运动生成任务的主流数据集与评价指标，并基于当前技术瓶颈对该领域未来发展方向进行展望。

1 3维数据表示

人与场景交互运动生成任务同时依赖场景数据和运动数据。其中，场景数据通过3维几何表征构建交互环境的空间拓扑结构，明确场景中静态物体(如大型家具、墙壁)与动态物体(如可移动道具)的空间位姿(位置及朝向等)；运动数据通过运动学参数描述智能体(如机器人、虚拟角色)与场景交互时的运动状态及动作规律。本节将系统讨论与分析上述2类数据的表示方法及其对运动生成任务的影响。

1.1 3维人体运动数据表示

人体运动数据丰富的表示形式在动画制作、虚拟现实和运动分析等领域具有广泛应用。Zhu等人[21]将运动数据表示分为基于关键点的表示和基于旋转的表示2类。关键点指生物学意义上的人体骨骼关节点，基于关键点的运动数据表示使用关节点在3维坐标系中的连续位置变化来描述人体运动姿态。该类数据通常由光学运动捕捉系统(如Vicon, OptiTrack)或惯性传感器(如Xsens)直接采集，其优势在于能够直观地展现人体运动时的全局空间轨迹。基于旋转的运动表示本质上是利用人体骨骼层级结构[22]描述人体运动。通过定义关节间的父子关系(如髋关节为膝关节的父节点)，采用局部坐标系下的旋转参数描述关节运动。旋转参数的表示通常使用欧拉角、轴角和四元数等，通过运动学链式递归变换实现旋转参数到全局位姿推导。在运动学链式递归中，通过正向运动学[23]和逆向运动学[23]可实现运动旋转表示和关键点表示的相互转换。其中，正向运动学[23]通过关节旋转复合运算生成骨骼末端轨迹，将旋转信息转换为关键点的3维位置坐标；逆向运动学[23]根据关键点的位置坐标来求解满足约束的最优关节角配置，获得关节的3维旋转信息。在机器人控制和数字人驱动等应用领域中，基于关节旋转的运动表示形式能够对控制信号进行直接映射，可实现对复杂运动的精确动力学控制。在交互运动生成任务中，人体参数化模型，如蒙皮多人线性(skinned multi-person linear, SMPL)模型[24]得到广泛应用，SMPL模型用姿态参数和体型参数分别表示运动旋转信息和人体形态，进而可以利用这些参数生成精细的人体网格。为提升模型表现力与场景适应性，SMPL-H[25]和SMPL-X[26]在SMPL模型的基础上对姿态参数进行扩展，添加了手部运动和面部表情参数，使得人体网格得到进一步的细化。详细的人体网格信息有利于准确建模人与场景的交互运动。

1.2 3维场景数据表示

传统的3维场景显式表示形式主要包括点云(point)、体素(voxel)、网格(mesh)和基点集(basis point set, BPS)[27]等。此外，近年来兴起的神经隐式表示方法也为3维场景的建模提供了新的途径，主要包括符号距离场(signed distance field, SDF)[28]、占用网络(occupancy networks)[29]和神经辐射场(neural radiance field, NeRF)[30]等。在基于场景交互的运动生成任务中，场景表达方式的选择通常取决于具体的交互任务。例如进行运动避障或者路径规划时，可以使用点云[15,31]、稀疏体素[17]，或者直接使用RGB图像[16,32]表达场景中的静态物体。在与物体进行交互时，需要采用精细的面片[18,33]、高分辨率体素[34]或基点集[35]等更加细致且可泛化的局部表达方式。

2 人与场景交互运动生成模型

现有的人与场景交互运动生成技术可以归纳为2类：一类基于自回归网络架构，该架构在时间序列预测任务中具有广泛应用。人体运动具有显著的时空连续性[36]，因此相关研究[31,37-39]采用自回归网络进行人体运动序列建模。典型自回归网络架构包含循环神经网络(recurrent neural network, RNN)[40]、长短时记忆(long short-term memory, LSTM)网络[41]和门控循环单元(gated recurrent unit, GRU)[42]等。另一类基于生成式模型架构，通过无监督学习策略建模人体运动的隐式概率分布或条件联合概率分布。代表性深度生成模型包含生成对抗网络(generative adversarial network, GAN)[43]、变分自编码器(variational autoencoder, VAE)[44]和去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)[45]等。本节对当前主流的运动生成模型展开系统性梳理和分析。

2.1 基于自回归网络的交互运动生成

基于自回归网络的人与场景交互运动生成方法[31,37-39]采用监督学习框架进行人体运动预测，其核

心在于将运动生成任务建模为序列预测任务。该方法通过建立输入特征(如历史姿态、环境信息等)与输出运动序列间的直接映射关系实现符合预期条件的运动序列生成。Corona 等人[37]使用图注意力网络(graph attention network, GAT)[46]将人和物体建模为图节点、人与物体的交互运动建模为边,通过迭代更新图注意力层学习具有上下文感知能力的语义图,并结合人体运动预测 RNN 网络[47]来实现未来短期的交互运动预测。Zhang 等人[38]使用基于 LSTM 架构[41]的控制网络(ControlNet)来预测未来手部关节点位置,并利用基于混合专家架构[48]的姿态网络预测全身运动序列。Araújo 等人[31]使用基于 Transformer[49]架构的自回归模型将场景点云特征和初始姿态进行位置编码来预测未来运动姿态。Mao 等人[39]基于门控循环单元(gated recurrent unit, GRU)[42]建模历史运动序列人体关节点与场景点云的接触关系来预测未来人体姿态。然而,基于自回归方式的运动生成模型往往只适用于特定的数据集,在跨场景数据分布差异较大时,其泛化能力通常显著降低。

2.2 基于生成对抗网络的交互运动生成

生成对抗网络[37]通过生成器(generator)和鉴别器(discriminator)的对抗训练框架实现数据分布建模。生成器以服从标准高斯分布的随机向量作为输入,生成与真实数据分布相近的合成样本;鉴别器则通过二元分类区分真实数据与合成数据。其优化目标可表示为零和博弈,目标函数公式为:

$$\min_G \max_D V(G,D) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\ln D(\boldsymbol{x})] + E_{z\sim p_z(z)}[\ln(1-D(G(z)))],$$

其中$\boldsymbol{x}$表示从真实数据分布$p_{\text{data}}(\boldsymbol{x})$中采样的观测样本,$z$表示从噪声分布$z\sim p_z(z)$中采样的噪声向量,$G(z)$表示生成器输出的合成样本。

人体运动生成任务中,HP-GAN[3]将 Seq2Seq[40]模型整合到生成对抗网络中实现运动序列生成。CSGN[50]构建基于图卷积网络的生成器和鉴别器,直接从噪声序列中生成运动序列。Wang 等人[16]在此基础上提出基于场景感知的人体运动合成生成对抗网络,该生成网络通过编码场景的图像特征引导轨迹生成器生成运动轨迹,姿态生成器根据初始姿态、轨迹以及场景特征为条件生成运动序列。如图 2 所示,该方法在训练阶段同时引入轨迹鉴别器、姿态鉴别器、投影鉴别器和上下文鉴别器,其中上下文鉴别器融合了场景的深度信息以提升模型的场景感知能力。投影鉴别器通过将 3 维运动姿态投影到 2 维空间以实现人体运动与场景结构的一致性约束。

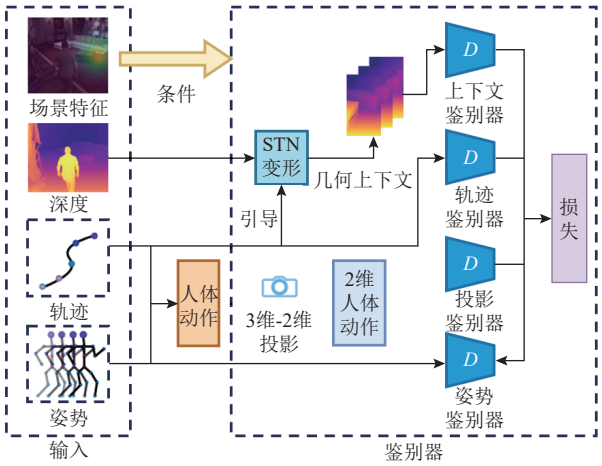


Fig. 2 Overview of the discriminator[16]

图 2 鉴别器概述[16]

尽管生成对抗网络在运动生成中表现出较强的数据建模能力,但是存在训练过程不稳定、收敛性差和模式坍塌现象等固有问题。上述问题在运动生成任务中会导致关节角度失真、运动连续性断裂等异常现象。

2.3 基于变分自编码器的交互运动生成

条件变分自编码器(conditional variational autoencoder, CVAE)[51]是由编码器和解码器组成,其核心原理是使用似然概率来近似难以解析计算的后验概率分布,推导过程采用了变分推断方法。在训练过程中,CVAE 模型的优化目标是最小化近似分布和真实后验分布之间的 KL 散度,具体通过优化证据下界(evidence lower bound, ELBO)实现,损失函数的形式化表示为:

$$Loss(\boldsymbol{x},\boldsymbol{y};\theta,\varphi) = E_{z\sim q_\varphi(z|\boldsymbol{x},\boldsymbol{y})}\left[\ln p_\theta(\boldsymbol{y}|\boldsymbol{x},z)\right] - KL\left(q_\varphi(z|\boldsymbol{x},\boldsymbol{y})\|p_\theta(z|\boldsymbol{x})\right),$$

其中$\boldsymbol{y}$表示观测数据,$\boldsymbol{x}$表示条件变量,z表示隐变量,$q_\varphi(z|\boldsymbol{x},\boldsymbol{y})$表示编码器推导的近似后验分布,$p_\theta(\boldsymbol{y}|\boldsymbol{x},z)$表示解码器生成数据的似然分布。模型能够根据学习到的概率分布生成新数据。近年来,基于 CVAE 架构的生成模型被广泛应用于人与场景交互运动生成领域。Cao 等人[32]提出 3 阶段框架,如图 3 所示。首先构建基于 CVAE 的目标预测网络 GoalNet,该网络以场景图像作为条件输入,将条件信息和历史姿态联合映射至隐空间,并基于变分推断框架估计后验概率分布,采样后与场景条件联合解码生成目标交互点;然后采用基于 Hourglass 架构[52]的路径预测网络 PathNet 来预测运动路径;最后基于 Transformer 网络构建姿态预测网络 PoseNet,根据历史姿态和运

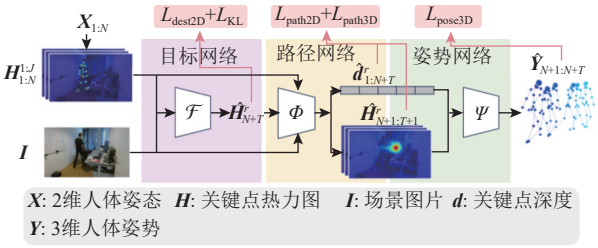


Fig. 3 Three-stage network architecture[32]

图 3 3 阶段网络架构[32]

动路径预测生成运动序列。

Wang 等人[15]将交互运动生成任务分解为子目标姿态生成和运动补全 2 个阶段；通过场景点云特征和子目标位置组合输入 CVAE 生成交互姿态，并根据各个子目标的交互姿态使用双向 LSTM 网络[53]合成完整运动序列。SAMP[17]基于 CVAE 构建 MotionNet 网络，该网络采用状态编码器和交互编码器分别编码历史运动姿态和交互对象的几何信息，编码特征经 2 个独立的全连接层预测近似后验高斯分布参数。推理时结合交互对象特征和历史运动状态生成交互运动。GOAL[35]提出的 GoalNetwork 以人体姿态参数、交互物体的 BPS 表示和位置信息以及接触距离信息作为输入来学习后验分布参数，在推理阶段以交互物体的 BPS 表示和位移信息为条件生成抓取姿态。SAGA[54]提出全身抓取网络 WholeGrasp-VAE 和运动填充网络 MotionFill-VAE。WholeGrasp-VAE 通过残差网络融合身体关键点、物体的几何特征和接触标签以预测后验分布参数。MotionFill-VAE 由 TrajFill 和 LocalMotionFill 子模块级联构成，分别用以生成全局根关节运动信息和关节局部运动姿态，进而生成全局运动姿态。HUMANISE[19]将语言文本描述与场景信息融合作为联合条件，如图 4 所示。通过多层感知机（multi-layer perceptron, MLP）预测高斯分布参数并基于 Transformer 架构生成给定场景和语言描述条件下的交互运动。IMoS[55]提出基于文本驱动的全身抓取运动生成模型，通过 CLIP（contrastive language-image pre-training）文本编码器[56]提取文本标签语义特征，结合目标物体属性与人体体型参数分别生成手臂和身体运动姿态，并设计优化模块生成符合物理规律的物体抓取运动。

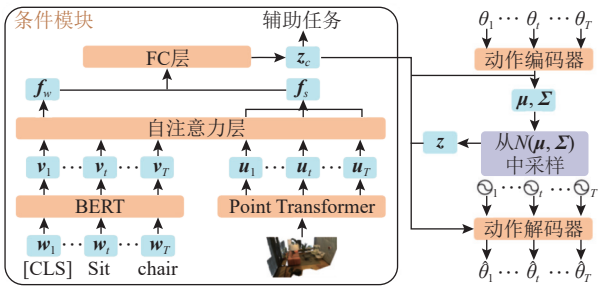


Fig. 4 HUMANISE network architecture[19]

图 4 HUMANISE 网络架构[19]

2.4 基于条件扩散模型的交互运动生成

近期的一些工作将扩散模型用于基于场景条件的交互运动生成任务，扩散模型[45]与变分自编码器类似，同样通过优化 ELBO 来学习数据的分布 p_θ。相比于其他生成模型，扩散模型能够产生高质量的样本并且训练稳定。与扩散模型相同，条件扩散模型的正向扩散过程在每个时间步 t 向原始数据 $\boldsymbol{x}^0 \sim q(\boldsymbol{x}^0)$ 逐步添加噪声 $q(\boldsymbol{x}^t|\boldsymbol{x}^{t-1})$，逆向去噪过程对随机采样的噪声 $\boldsymbol{x}^T \sim N(0,\boldsymbol{I})$ 进行迭代去噪 $p_\theta(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t,\boldsymbol{c})$，经过 T 步去噪得到生成数据 $\hat{\boldsymbol{x}}_0$，具体公式表示为：

$$q\left(\boldsymbol{x}^t|\boldsymbol{x}^{t-1}\right) \sim N\left(\boldsymbol{x}^t;\sqrt{1-\boldsymbol{\beta}_t}\boldsymbol{x}^{t-1},\boldsymbol{\beta}_t\boldsymbol{I}\right),$$

$$p_\theta\left(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t,\boldsymbol{c}\right) \sim N\left(\boldsymbol{x}^{t-1};\mu_\theta\left(\boldsymbol{x}^t,t,\boldsymbol{c}\right),\Sigma_\theta\left(\boldsymbol{x}^t,t,\boldsymbol{c}\right)\right),$$

其中 $\boldsymbol{c}$ 表示输入条件，$\boldsymbol{\beta}_t$ 表示扩散时间表，μ_θ 和 Σ_θ 表示通过神经网络学习的参数。

SceneDiffuser[57]首次使用条件扩散模型解决场景交互运动生成和路径规划任务。OMOMO[58]根据物体运动采用条件扩散模型生成人物交互运动。该方法以交互对象的 BPS 表示为条件，利用扩散模型生成手部关节运动姿态，随后以手部关节运动为条件利用扩散模型生成全身交互运动。Jiang 等人[59]提出自回归扩散策略，将整段交互运动生成划分为多个子阶段以实现任意长度序列生成。该方法以场景信息和运动标签为条件引导扩散模型生成各子阶段运动，并基于 Shafir 等人[4]的方法进行相邻子片段之间的平滑过渡。

Cen 等人[34]提出使用大语言模型先验辅助目标物体的定位，便于模型聚焦于特定物体附近的场景表示，进而提升交互运动生成的真实性。该方法借助 3 维物体检测模型[60]获取文本形式的场景图，并利用 Referit3d[61]模型分析场景中不同物体之间的位置关系，随后借助 ChatGPT-3.5 定位目标物体。该模型以目标物体周围的场景为条件引导扩散模型生成交互运动，降低了场景表示的复杂性，有利于提升模型对于场景和文本信息的理解能力。

Wang 等人[62]引入场景功能图实现文本驱动的人与场景交互运动生成任务，如图 5 所示。该方法采用 2 阶段扩散模型框架实现人与场景交互运动的生成。第 1 阶段以场景与人体的接触信息和交互文本为条件输入，通过扩散模型生成场景功能图，建立

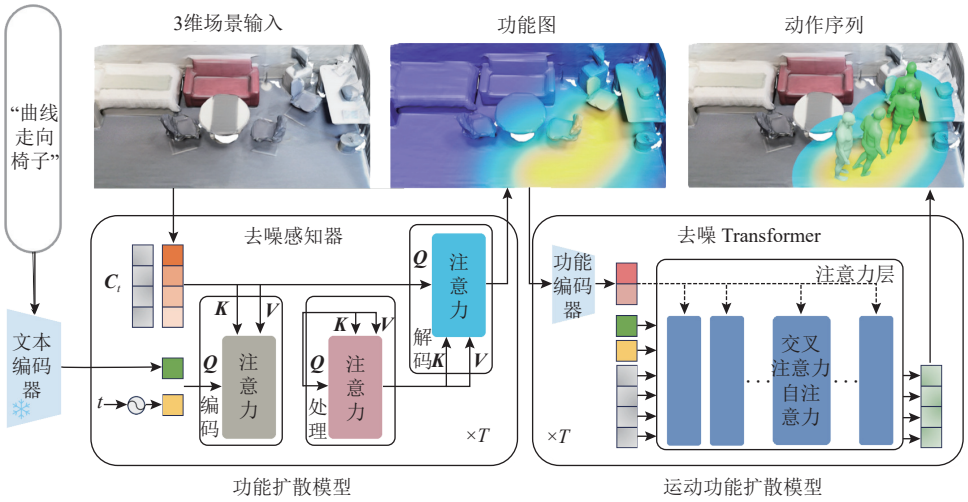


Fig. 5 Network architecture[62]

图 5 网络架构[62]

交互运动与场景间的映射关系。第 2 阶段以交互文本描述和场景功能图作为条件通过另一扩散模型生成交互运动。通过构建场景功能图表征，模型在不同场景下的跨域泛化能力得以提升。

HOIAnimator[33] 基于条件扩散模型构建了以文本描述为条件的交互运动生成模型。该方法首先结合文本描述和接触信息建立交互接触场以预测接触信息，同时提出具有消息传递机制的感知扩散模型(perceptive diffusion model)，以文本描述和接触信息为条件同时生成物体运动和人体运动。TeSMo[63] 使用 2 维地图和 3 维物体表示场景交互数据，对动作生成扩散模型进行场景感知微调训练。该方法基于导航和交互 2 个阶段的扩散模型，给定起始姿态、导航阶段的结束姿态和文本描述，进而生成运动轨迹和目标物体的交互运动。

场景交互运动生成技术经历了从确定性自回归预测到概率生成建模的发展历程。图 6 为基于场景交互的动作生成方法的发展历程，清晰描述了该领域的技术发展脉络。

2.5 实验对比

为客观对比不同交互运动生成模型的性能，表 1 汇总了 3 种生成模型在基于文本描述引导及 3 维场景约束下的人与场景交互运动生成任务中的定量实验结果。所有实验均基于 HUMANISE 数据集[19]，并遵循 Wang 等人[62] 提出的实验设置和数据集划分方式进行。物理合理性的评估通过无碰撞(non-

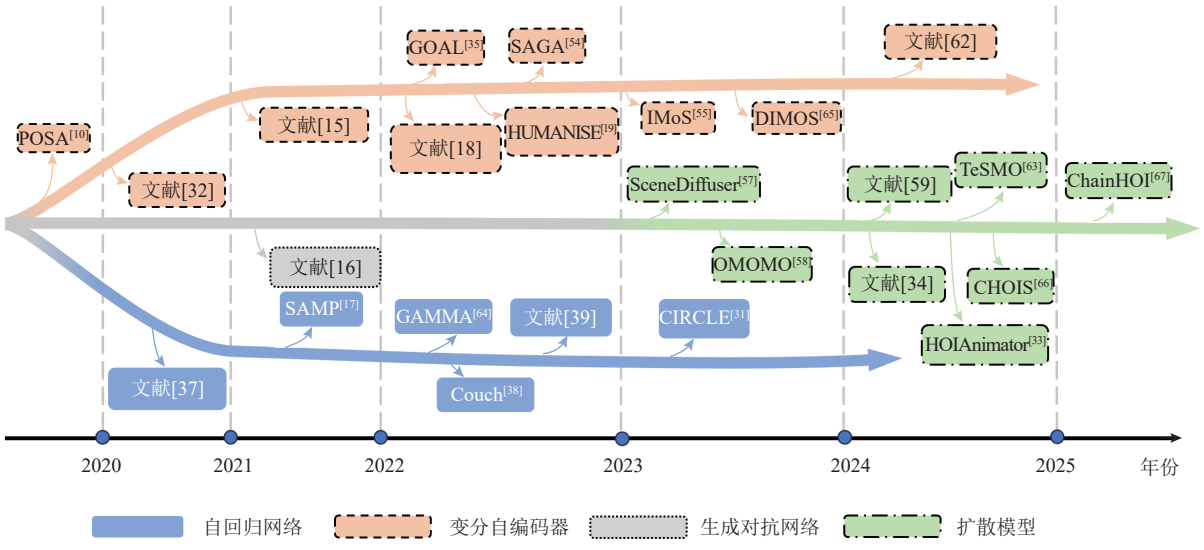


Fig. 6 Development process of action generation technology based on scene interaction

图 6 基于场景交互的动作生成技术发展历程

collision)分数和接触(contact)分数量化，其中无碰撞分数计算生成动作中未发生接触的身体网格顶点的占比，接触分数计算接触距离在一定阈值范围内身体网格顶点的占比，数值越大代表生成动作更合理。

Table 1 Effectiveness Comparison of Different Human-Scene Interaction Motion Generation Models

表 1 不同人与场景交互运动生成模型的效果对比

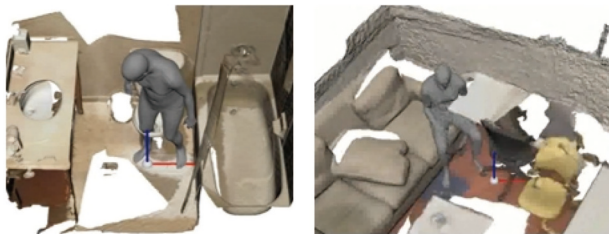
方法	模型框架	无碰撞分数↑	接触分数↑
HUMANISE[19]	CVAE	**99.77**	84.07
文献 [39]	Regression	98.38	92.16
文献 [62]	Diffusion	99.69	**96.02**

注：黑体数值为最优结果，“↑”表示数值越大效果越优。

图 7 展示了 3 个模型在人与场景交互运动生成任务上的定性实验结果。分析表明，基于 CVAE 模型的方法 HUMANISE[19] 实现了较高的无碰撞分数值，但其生成的动作存在交互位置偏移问题，例如图 7(a)示例中“人坐在沙发上”“人走向桌子”的交互指令未生成正确交互位置。文献 [39] 提出的基于自回归模型的方法具有严重的人体网格与场景网格的穿透问题，如图 7(b)所示。最新的基于扩散模型的文献 [62] 方法取得了最优的接触分数，视觉效果上也取得了最好的物理合理性，体现了其在处理复杂交互场景方面的优越性能。



(a) HUMANISE[19]方法效果



(b) 文献[39]方法效果



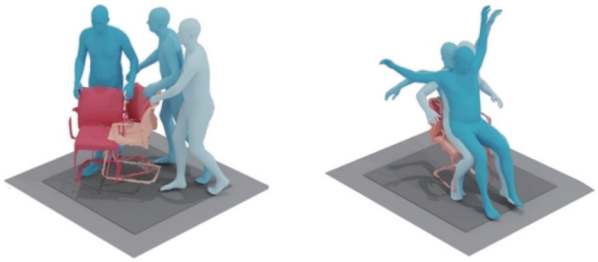
(c) 文献[62]方法效果

Fig. 7 Effectiveness illustration of interaction motion generation models

图 7 交互运动生成模型效果示意图

3 人与场景交互运动生成任务

人类与场景进行交互的本质是人类与场景中的物体进行相互作用[39]，而人与场景中的物体交互可以根据交互时的运动状态分为 2 类：动态物体交互和静态物体交互[59]。具体而言，动态物体交互是指物体的空间位置或方向会因交互行为发生改变，例如物体抓取、搬运和堆叠等；静态物体交互是指人与物体进行交互时，被作用的物体始终保持静止状态，例如躺上床、坐上沙发和上下楼梯等。本节将对人与动态物体和静态物体的交互运动生成任务进行讨论和分析。图 8 展示了不同的交互任务，表 2 针对这些方法进行汇总。



(a) 人拿着椅子 (b) 人坐在椅子上

Fig. 8 Illustration of dynamic-static interaction tasks

图 8 动静交互任务示意图

3.1 动态物体的交互

本文讨论的任务为全身抓取生成(whole-body grasping generation)，该任务不同于传统的手部物体抓取生成(hand-object grasping generation)[68-69]。有关后者，请读者参考相关综述文献 [70] 以获取详尽讨论。传统的手部物体抓取动作旨在单手稳定的抓握、举起或者简单使用物体[71]。而全身抓取[72] 区别于传统手部物体抓取，其不仅包含手部的抓取动作，还包括身体其他部位在操作过程中的协同配合，例如“使用碗喝水”这一动作，人需要首先调整头部姿态确定碗的空间位置，接着通过手臂和手部的协同运动稳定地将碗拿起并放到嘴边，最后适当倾斜头部以完成喝水动作。此类全身抓取运动中，物体通常会与人体的多个部位发生接触，而传统物体抓取运动中，接触只会发生在手部。全身抓取运动数据集 GRAB[73]、人-物交互数据集 BEHAVE[74] 和 OMOMO[58] 的提出为学习并建模这种复杂的全身抓取动作提供了训练

Table 2 Summary of Human-Scene Interaction Motion Methods

表 2 人与场景交互运动方法汇总

方法	数据表示	模型	数据集	开源情况
文献 [37]	Kpts.(3D)	Regression	CMU-Mocap, WBHM	√
文献 [32]	Kpts.(3D)	VAE	GTA-IM, PROX	×
文献 [15]	Rot.	VAE	Matterport3D, PROX	√
文献 [16]	Kpts.(3D)	GAN	GTA-IM, PROX	×
文献 [17]	Rot.	VAE	SAMP	√
文献 [18]	Rot.	VAE	Matterport3D, PROX	×
文献 [64]	Kpts.(3D)	RL	AMASS	√
文献 [35]	Rot.	VAE	GRAB	√
文献 [54]	Kpts.(3D)	VAE	GRAB	√
文献 [38]	Rot.	Regression	Couch	√
文献 [39]	Kpts.(3D)	Regression	GTA-IM, PROX	√
文献 [19]	Rot.	VAE	HUMANISE, PROX	√
文献 [55]	Rot.	VAE	GRAB	√
文献 [57]	Rot.	Diffusion	PROX	√
文献 [31]	Rot.	Regression	CIRCLE	×
文献 [65]	Kpts.(3D)	RL, VAE	AMASS, SAMP	√
文献 [58]	Kpts.(3D)	Diffusion	OMOMO	√
文献 [59]	Kpts.(3D)	Diffusion	TRUMANS	√
文献 [34]	Rot.	Diffusion	AMASS, HUMANISE	仅推理
文献 [62]	Rot.	Diffusion	HumanML3D, HUMANISE	√
文献 [33]	Rot.	Diffusion	Behave, InterCap	×
文献 [63]	Kpts.(3D)	Diffusion	SAMP, 3D-FRONT	仅推理
文献 [66]	Rot.	Diffusion	OMOMO, 3D-FUTURE	√
文献 [67]	Kpts.(3D)	Diffusion	BEHAVE, OMOMO	×

注：Rot.为基于旋转的表示，Kpts.(3D) 为基于关键点坐标的表示。

资源。在全身抓取生成任务中，多数现有方法[35,54,58]采用多阶段生成策略，GOAL[35]和 SAGA[54]采用 2 阶段交互运动生成策略。在第 1 阶段，模型需要明确目标交互物体的位置和方向信息，以生成针对目标物体的全身抓取姿态。在第 2 阶段则需提供起始姿态，并结合第 1 阶段生成的目标物体抓取姿态生成中间连续运动序列。

InterDiff[75]基于历史人物交互运动序列和物体形状信息，预测未来一定时间内多样化的物体交互运动。OMOMO[58]针对物体搬运任务，收集了大规模大型物体交互运动数据集，并据此学习了一个 2 阶段交互运动生成模型，该模型首先根据物体运动序列预测手部交互运动序列，然后根据手部的运动生成全身运动序列。

相比之下，CHOIS[66]使用单阶段模型解决物体搬运任务，通过统一的去噪网络将所需条件拼接作为输入，并在生成采样过程中应用接触约束以提升生成动作的合理性。

近期的研究工作[33,55,67]使用动作指令或者文本描述并结合历史运动或起始姿态来引导交互运动生成，通过自然语言指导人与环境中物体交互运动的生成。该方法在许多领域具有广泛的应用前景，包括机器人技术、计算机图形学[76-78]、游戏设计[79]和电影制作[80]等，例如在电影虚拟场景或数字创建的角色扮演游戏中，根据一组指令控制虚拟角色与场景物体进行交互不仅能优化操作环节，而且能够节省大量成本。IMoS[55]提出了基于文本指令进行全身交互运动合成模型。文本指令由动作和目标交互物体组成，例如“拍照+相机”，模型以初始姿态、目标物体的位置以及文本指令作为输入，生成全身抓取运动序列。HOIAnimator[33]提出感知扩散模型，仅使用一段文本描述引导生成物体搬运交互运动序列。ChainHOI[67]则引入时空生成模块和基于运动学的交互模块，分别在关节层面和运动交互语义层面建模人与物体的交互关系。

3.2 静态物体的交互

近年来针对人与场景静态物体交互运动生成研究主要聚焦于 2 个核心方向：其一为复杂场景中的人体避障漫游[16,32,39,64-65]算法研究；其二为人与场景中具有特定功能语义的静态物体(例如椅子、床等)的人物交互运动生成技术[15,17-19,31,34,38,62]。

避障漫游任务中通常以场景信息、运动的开始和结束位置作为条件生成人与物两者之间完整的运动序列，同时要求运动过程避免与场景中的物体发生碰撞。该过程生成运动的合理性依赖于模型对场景信息的理解能力。现有利用场景信息进行约束的方法，包括但不限于从场景图像[16,32]或点云数据[15,19,31,57]中提取特征进行隐式约束、基于接触关系预测的显式约束[33,39,62]、使用强化学习[64-65]或路径规划算法 A*[17-18]等。Cao 等人[32]和 Wang 等人[16]采用多阶段运动生成框架，先利用场景 RGB 图像特征结合历史人体姿态进行运动轨迹预测，再基于运动轨迹和场景特征生成完整运动序列。一些研究采用点云表示场景[15,19,31,57]，利用 Point-Net[81]或 Point Transformer[82]等预训练的编码器从场景点云中提取语义特征作为场景约束进行运动生成。Mao 等人[39]提出基于人体关节点与场景点云间距离场的方法，显式地建模人与场景交互关系，预测每个关节点与场景点云的接触情况并生成人体运动姿态。为更准确地建模接触

情况，Wang 等人[62]对上述距离进行归一化计算，赋予最近点更高权重，进一步提升了模型的训练稳定性。HOIAnimator[33]通过计算采样点到人体表面的距离引入交互接触场进行场景约束。TeMSo[63]使用 Transformer 编码器对场景的 2 维平面图编码，并对路径生成扩散模型进行微调训练以实现模型对场景信息的感知。GAMMA[64]和 DIMOS[65]将避障漫游运动生成任务表示为马尔可夫决策过程，利用奖励机制实现场景的几何约束，使用策略网络指导运动生成。SAMP[17]使用显式路径规划算法 A*[83]计算从起始位置到目标位置的无碰撞移动轨迹，便于在后一阶段根据轨迹生成运动序列。Wang 等人[18]对标准 A*路径搜索算法进行改进，通过在代价函数添加可控噪声项，实现基于交互锚点的运动路径的多样化生成。

在现实场景中，人体运动往往与周围的环境和所接收的指令紧密相关。为了生成更加自然和符合逻辑的运动，交互运动生成模型必须能够理解并响应文本指令，同时考虑场景的上下文信息。人体运动生成领域已实现令人满意的文生动作工作[84-86]效果，基于文本和场景信息引导的运动生成模型需要同时具备场景和文本等多个模态的理解能力，以生成与文本描述一致的物理合理的交互。HUMANISE[19]将大规模动作捕捉数据集 AMASS 中的动作序列与真实世界场景数据集 ScanNet 进行对齐，对每段交互运动序列添加详细文本描述作为引导运动生成的文本条件，使用预训练 BERT 文本编码器[87]和 Point Transformer 编码器[82]对文本条件和场景点云进行特征提取以实现交互运动生成。由于点云噪声可能导致物体定位不准确，Cen 等人[34]提出使用大模型的先验能力辅助模型进行目标物体的定位和交互，只考虑目标对象周围的场景约束，从而降低场景表示的复杂性，提升模型对于场景和文本信息的理解能力。Wang 等人[62]通过计算场景点与人体关节点之间的距离建模场景点云与交互运动间的关系，构建场景功能图，结合此功能图和文本描述对文本和场景交互运动进行联合建模。

综上所述，针对当前的人与场景交互运动生成任务，无论是动态交互还是静态交互，当前的研究普遍采用自然语言文本信息作为辅助信息，引导模型生成具有语义一致性的交互运动。此外，近期的工作逐渐引入多模态大语言模型（multimodal large language models, MLLMs），借助其在文本与视觉理解方面的强大能力，提升模型对于复杂交互运动的建模能力，同时在一定程度上降低了训练成本。然而，相关方法仍处于初步探索阶段，尚未形成成熟的技术体系。未来如何有效地融合多模态信息和大模型的通用先验知识，以实现更高质量、更具泛化能力的交互运动生成，将是一个具有重要研究价值的方向。

4 人与场景交互运动数据集

为了支持人与场景交互运动生成任务的研究和测评，学界提出一系列交互运动和非交互运动数据集，本节将对这些数据集进行讨论。表 3 所示为数据集的详细信息。

4.1 人与场景交互数据集

WBHM 数据集[88]包含由 Vicon-MX 运动捕捉系统采集的多组 3 维全身人体运动数据和物体运动数据，每组数据都具有详细的运动描述，并提供了环境元素和物体的信息。运动数据不仅包括人体运动，还包括人与物的交互动作，例如开门、抓取、推动等。该数据集共包含 41 个不同的 3 维物体例如杯子、楼梯、食物等。

PiGraph 数据集[89]使用体积融合框架[90]对真实世界的 3 维场景进行重建，在场景中使用 Kinect.v2 深度相机捕捉人与场景进行交互时的运动数据，并对 3 维场景中的所有物体和交互动作添加注释。该数据集共包含 30 个场景和 19 个物体。

PROX 数据集[91]使用商用传感器 Structure sensor 和配套的 3 维重建方案 Skanect 重建了 12 个室内场景，包括 3 间卧室、5 间客厅、2 个座位和 2 个办公室。使用 Kinect-One 深度摄像机捕捉 20 名受试者与场景的互动，并使用 SMPL-X 参数化模型表示人体运动参数。

i3DB 数据集[92]对采集到的 RGB 视频利用半自动标注的方式估计场景中物体的位置和人的动作。该数据集共采集 15 个场景和带有标签的物体，例如椅子和桌子等，并且包含人与物体的交互运动序列以及相应的注释。

GTA-IM[32]是一个大规模合成数据集，利用侠盗猎车手游戏引擎进行人物动作的捕捉，同时配备虚拟场景的 RGB-D 视频和相应的标签。该数据集包含 50 个人物角色以及 10 个不同的室内场景，每个场景包括客厅、卧室、厨房和阳台等，可以进行多样化的交互活动。

GRAB 数据集[73]使用 Vicon 运动捕捉系统捕捉真实世界的人体全身运动、抓取物体的手部操作、物体运动以及接触情况。该数据集包含 10 位受试者

Table 3 Summary of Interactive Dataset

表 3 交互数据集汇总

数据集	收集方式	数据表示	受试者人数	序列数	帧数	时间/h	其他信息
WBHM	Marker-based	Rot.	43	3 704	0.7×10^6	7.68	41 个物体
PiGraph	Markerless	Kpts.(3D)	5	63	0.1×10^6	2	49 个场景和物体
PROX	Markerless	Rot.	20	60	0.1×10^6	1	12 个场景
i3DB	Pseudo-labeling	Kpts.(3D)	1				15 个场景
GTA-IM	Marker-based	Kpts.(3D)	50	119	1×10^6		10 个场景
GRAB	Marker-based	Rot.	10	1 334	1.6×10^6	3.8	51 个物体
HPS	Marker-based	Rot.	7		0.3×10^6		8 个大场景
SAMP	Marker-based	Rot.	1		0.185×10^6	0.83	7 个物体
COUCH	Marker-based	Rot.	6	>500		3	3 个椅子
HUMANISE	Marker-based	Rot.		19 600	1.2×10^6		643 个场景
BEHAVE	Markerless	Rot.	8	321	0.015×10^6	4.2	20 个物体
CIRCLE	Marker-based	Rot.	5	>7 000	4.3×10^6	10	9 个场景
OMOMO	Marker-based	Rot.	17	16 900	4.3×10^6	10.1	15 个物体
CHAIRS	Marker-based	Rot.	46	1 390	1.8×10^6	17.3	81 个物体
InterCap	Markerless	Rot.	10	223	0.067×10^6	0.62	10 个物体
TRUMANS	Markerless	Rot.	7		1.6×10^6	15	120 个场景和物体

注：Rot.为基于旋转的表示，Kpts.(3D) 为基于关键点坐标的表示。

与 51 个日常物体的交互动作，例如吃香蕉、用碗喝水、用锤子等，人体动作数据采用 SMPL-X 格式，包含丰富的手部和面部动作。

HPS 数据集[93]使用可穿戴传感器和佩戴在受试者头部的头戴式摄像机进行动作捕捉和 3 维场景的扫描和重建。该数据集包含 7 名受试者和 8 个大型场景，其中一些场景面积超过 1 000 m^2，包含受试者在这些场景中进行各种活动（如阅读、散步、吃饭等）的运动数据。

SAMP 数据集[17]使用配备 54 个 Vicon 相机的光学动作捕捉系统收集了人与场景交互数据集。在动作捕捉过程中使用了 7 个不同的物体，例如桌子和椅子等。该数据集包含人与物体多种风格的交互动作以及物体的位置变化，人体动作使用 SMPL-X 模型来表示。

COUCH 数据集[38]使用 17 个可穿戴惯性测量设备 XSens 收集数据，通过虚拟引擎 Unity[94]从运动数据中获取 SMPL 格式的高质量人体动作序列，该数据集包含 6 个不同人物与椅子进行交互的数据，总计超过 500 个动作序列，总捕获时长为 3 h。

HUMANISE[19]是一个大规模合成数据集，通过将 AMASS[95]数据集中捕获的人体动作序列与 ScanNet[96]数据集中扫描的室内场景进行对齐，每个运动片段都有详细的文本描述，描述主要包含动作标签、交互对象、空间关系和周围锚点对象 4 个部分。该数据集一共包含 643 个 3 维场景，最具代表性的动作是坐、站、躺和行走。

CIRCLE 数据集[31]包含来自 9 个场景时长约 10 h 的交互动作，这些动作包括爬行、蹲下、躺下和弯腰等，数据由 Vicon 动作捕捉系统采集。

OMOMO[58]收集了一个由 3 维物体几何、人和物体的运动组成的大规模高质量数据集。其中包括 15 个常见的物体的网格，例如衣架、拖把、桌子等。使用 Vicon 动作捕捉系统捕捉来自 17 名受试者与不同物体的互动，使用每秒 120 帧的速度录制，总时长共计约 10 h。

BEHAVE 数据集[74]收集了人体和物体的交互运动以及接触注释。该数据集使用 4 台 Kinect RGB-D 摄像机记录了 8 名受试者在 5 个自然环境中与 20 个物体的交互运动的共计 321 个视频序列，每帧都包含人和物体的分割点云以及掩码。

CHAIRS 数据集[97]是一个大型的捕捉人和物体交互的动作数据集。使用运动捕捉系统捕捉人和物体的交互运动，提供准确的 3 维网格，涉及 46 名受试者与 81 个物体长达 17.3 h 的交互动作。

InterCap 数据集[76]从 223 个 RGB-D 视频，共计 67 357 个多视图帧中重建出 10 个受试者与 10 个物体进行日常交互运动的 3 维网格，并使用 SMPL-X 参

数化模型表示人体参数。

TRUMANS[59]是一个大型的人与3维场景物体交互的运动数据集，该数据集使用最先进的Vicon运动捕捉系统，以30 Hz的速率收集共计1.6×10^{6}帧运动数据。该数据集的特点是多样化的交互场景，涵盖了100个不同的室内环境（如客厅、餐厅、卧室和厨房等），参与者共7人（包括4名男性和3名女性）。在这些场景中与20种常见物体进行交互，包括静态物体（如椅子、沙发等）、动态物体（如水杯、瓶子等）和铰接物体（如笔记本电脑、橱柜等）。

4.2 非交互数据集

除了上述将场景与运动对齐的交互运动数据集之外，一些工作也会将只包含场景信息的场景数据集[96,98-99]作为测试集，测试模型的泛化能力。此外，只包含运动数据的纯运动数据集[95,100]也被用于运动生成器的预训练，以提升生成运动的质量。

ScanNet数据集[96]是一个室内场景的RGB-D视频数据集，包含1 513个场景的250万张视图，此外还包含了丰富的标注信息，例如3维相机姿态、表面重建和分割等标注，每个场景都有详细的实例级对象类别标签。

Matterport3D数据集[98]包含90个室内场景的10 800个全景视图，总共涵盖了194 400个RGB-D图像。这些图像包括高动态范围（HDR）的彩色图像和对应的深度图像，为3维场景的分析和理解提供了高质量的数据。同时该数据集提供了详细的标注信息，包括表面重建、相机姿态以及2维和3维的语义分割信息。

Replica数据集[99]是一个3维场景重建数据集，包含18个房间和建筑物的高清纹理场景重建，每个场景都包含密集网格、高分辨率的纹理和图形元素的语义类别以及实例信息。

Human3.6M数据集[100]由11位受试者（5位女性和6位男性）在4个不同视角下表演的360万个3维人体姿态组成。这些数据通过高精度的运动捕捉系统收集，涵盖了谈话、吃饭、问候和打电话在内的多种日常活动场景。数据集提供了同步的2维图像、3维运动捕捉和深度数据，以及所有受试者的精确3维身体扫描信息。

AMASS[95]数据集包含超过40 h的动作数据，涵盖300多个受试者和超过11 000种动作，该数据集的关键特点是使用MoSh++方法，将动作捕捉数据转换成SMPL-X格式的3维人体网格数据。

HuMMan数据集[101]包含1 000名受试者的400 000段运动序列共计60×10^{6}帧。该数据集包括彩色图像、点云、关键点、SMPL参数和带纹理的网格等多种数据格式和注释。此外该数据集包含大约500个不同的动作种类，这些动作覆盖了步行、跳跃、坐下等基础的人体运动。

Motion-X数据集[102]包含了从多个场景中收集的81 100个动作序列，共计15.6×10^{6}帧，并且包含符合SMPL-X格式的全身姿态参数和81 100个序列的语义标签，是当前最新的大型运动估计数据集。

5 评价指标

为了评估交互运动生成模型的效果和定量分析模型的性能，学术界提出了各种衡量生成结果的评价指标，这些指标主要考虑生成运动质量、多样性和自然程度等方面。此外，随着文本条件的引入，评价指标也拓展到对于文本和运动2个模态的一致性、相关性的衡量。

5.1 运动质量度量

5.1.1 距离误差

评估生成运动质量的一种直接做法是将生成运动与真值之间的距离作为度量，例如人体根节点位移和旋转参数的l_1范数误差、关节平均位置误差（*APE*）和网格顶点的平均位置误差[15,19,31,55,58]，以及计算关节旋转的Frobenius范数误差[58]。生成运动的第j个关节与真值之间的*APE*计算公式为：

$$APE[j]=\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left\|P_t[j]-\hat{P}_t[j]\right\|_2,$$

其中T表示总帧数，N表示测试集的运动序列个数。$P_t[j]$表示第t帧第j个关节的位置，这种直接计算距离的做法比较简单，往往不能全面地反映生成动作的质量。因此一些工作[17,55,58]提出使用运动数据分布的几何统计量距离作为一种衡量运动质量的度量，例如平均方差误差（*AVE*）、*FD*等指标。其中*AVE*具体计算公式为：

$$AVE[j]=\frac{1}{N}\sum_{n=1}^{N}\left\|\sigma[j]-\hat{\sigma}[j]\right\|_2,$$

其中σ表示生成运动的方差，$\hat{\sigma}$表示真值的方差。此外还有一些工作[16,33-34,55,59,62]通过计算生成运动和真值在特征空间中的距离来衡量运动质量的好坏例如FID，该指标的计算需要额外训练一个运动分类器以提取运动特征。

5.1.2 用户主观评价

由于人类自身对运动质量和自然程度的感知十

分敏感, 能够捕捉到客观指标无法评估的细微方面, 因此主观评价在衡量生成运动的自然程度方面发挥了重要作用, 许多工作[16,18,55,64]使用主观评价作为补充以进一步评估生成运动的质量。以文献[16]为例, 该研究采用了场景匹配度分数和运动匹配度分数2个主观评价指标, 分别要求用户评估生成的人体运动与给定场景之间的协调程度和人体运动轨迹与运动姿态序列的匹配度。

5.2 条件一致性度量

5.2.1 场景条件一致性

在人与场景交互运动生成任务中, Zhu等人[21]将场景信息看作是一种引导模型生成交互运动的条件信号, 为了衡量生成运动与场景条件的一致性或人与场景交互运动的物理合理性, 现有工作大多采用接触分数和无碰撞分数[15,18,58,62]进行一致性评估, 其中无碰撞和接触的判断一般基于身体网格和场景物体点的符号距离计算, 若距离为正则判断为无碰撞, 否则判断为接触, 无碰撞分数表示没有发生碰撞的运动数量与所有采样运动数量的比例, 接触分数表示具有接触网格数量和所有生成网格数量的比值, 更高的无碰撞分数和接触分数代表交互运动具有更高的物理合理性。Wang等人[16]对生成的人体运动和场景物体构建圆柱体模型来计算无碰撞分数和接触分数, 其中圆柱体的半径是预定义的, 如图9所示。

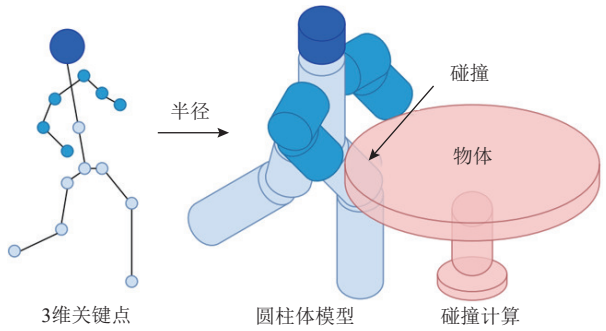


Fig. 9 Collision model of cylinder between human and object[16]

图9 人与物体的圆柱体碰撞模型[16]

一些工作[31,35,54]直接使用人体与场景物体的穿透体积和深度进行场景一致性评估, 穿透体积和深度越小表示生成的运动具有更高的物理合理性。这些指标常用于物体抓取和漫游避障等交互任务的评估, 具有较强通用性。还有一些指标通过评估身体到达目标位置的程度或者与正确物体的交互程度来衡量场景运动一致性, 例如HUMANISE[19]提出使用身体到目标位置的最短距离来评估交互完成情况, GAMMA[64]和Scene-Diffuser[57]使用角色多次尝试成功到达目标位置的成功率和路径规划所需要的规划步骤, DIMOS[65]使用完成交互动作的时间来衡量交互运动的场景一致性。此外Cao等人[32]也将主观评价作为一个衡量生成交互运动与场景的一致性的补充指标。

5.2.2 文本条件一致性

本文使用文本描述引导交互运动生成的工作[33-34,55,62]来衡量文本条件与交互运动的一致性, 评估两者之间的匹配程度, 通常使用识别准确率、召回率、多模态距离等指标。具体来说, 识别准确率使用预训练的动作分类器对生成运动进行分类并计算分类正确的比例。召回率计算每个生成运动与描述池中每个文本描述的特征向量之间的欧几里得距离, 并对这些距离进行排序, 计算前3个距离最近的文本特征与运动特征的匹配准确率。多模态距离计算的是每个生成动作与其对应的文本描述特征向量之间的平均欧几里得距离, 它衡量生成运动与其文本描述之间的相关性。多模态距离数值越小说明生成运动与文本描述条件匹配程度越高、相关性越强。

5.3 运动多样性度量

在场景交互运动生成领域, 交互运动的多样性问题长期以来备受关注, 为了更合理地评估生成运动的多样性, 现有的工作通常从运动序列内和运动序列间2个角度出发来全面评估多样性。对于运动序列内多样性评估, 已有方法往往使用平均成对距离(average pairwise distance, APD)[17-18,54,62]; 对于运动间多样性评估方法, 已有工作[33-34,55]使用多样性(*Diversity*)分数和多模态(*Multimodality*)分数[103]。值得注意的是, 这2个指标均是在运动特征空间中进行多样性评估。

5.3.1 平均成对距离

平均成对距离通过计算来量化单个生成运动序列的多样化程度[17-18,54,62], 公式为:

$$APD=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=0,j\neq i}^{N}\left\|\boldsymbol{X}_i-\boldsymbol{X}_j\right\|_2^2,$$

其中N是运动序列的总帧数, $\boldsymbol{X}_i$表示序列第i帧的局部姿态, 其由位移、旋转和速度构成, APD值越大表示运动内的多样化程度越高。

5.3.2 多样性

多样性分数计算所有运动类型之间特征的方差, 从包含各种运动类型的生成运动中随机抽取2个大小为S的子集, 分别提取这2个子集中所有运动的特征向量表示为$\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_S\}$和$\{\boldsymbol{v}'_1,\boldsymbol{v}'_2,\cdots,\boldsymbol{v}'_S\}$, 具体计算

公式为：

$$Diversity=\frac{1}{S}\sum_{i=1}^{S}\left\|\boldsymbol{v}_i-\boldsymbol{v}'_i\right\|_2。$$

多样性分数越接近真值，表明生成运动的多样化程度越强。

5.3.3 多模态分数

多模态分数衡量同一种动作类型中的多样化程度，假设一共有H个动作类型，对于第h种动作类型随机抽取2个大小为T的运动子集，分别提取其特征向量表示为$\{\boldsymbol{v}_{h,1},\boldsymbol{v}_{h,2},\cdots,\boldsymbol{v}_{h,T}\}$和$\{\boldsymbol{v}'_{h,1},\boldsymbol{v}'_{h,2},\cdots,\boldsymbol{v}'_{h,T}\}$。具体计算公式为：

$$Multimodality=\frac{1}{H\times T}\sum_{h=1}^{H}\sum_{i=1}^{T}\left\|\boldsymbol{v}_{h,i}-\boldsymbol{v}'_{h,i}\right\|_2,$$

其中$\boldsymbol{v}_{h,i}$和$\boldsymbol{v}'_{h,i}$分别表示第h种动作类型的2个运动子集中第i个运动序列的特征，多模态性分数越高表明每个运动类型内的多样性越明显。基于运动特征计算的指标的可靠性往往取决于运动特征提取器的有效性，并且特征提取器的训练高度依赖数据集分布，因此这些指标通常难以进行跨数据集的评估。

6 未来工作展望

综上所述，人与场景交互运动生成技术正迅速发展，并且受到计算机视觉、计算机图形学和机器人技术等领域众多学者的广泛关注，在游戏制作、动画生成和元宇宙等领域具有广阔的应用前景，但仍面临以下3个方面的关键挑战。

1)环境感知能力增强。现有方法[33,35,39,54,62]通常使用简单的几何特征(例如人与场景中物体点的距离或者接触信息)进行场景理解，亟需构建多模态感知框架来融合场景语义理解与实时空间拓扑更新等信息，以实现对交互意图的精准解析。

2)风格可控的场景交互运动生成。当前基于场景条件的交互运动生成输出动作风格单一、缺乏个性化表达，例如在避障漫游任务[17-18]中生成交互锚点之间的运动无法指定生成特定风格的运动(如谨慎避障和随意漫步)，其根源在于训练数据缺乏风格标注(情感标签、运动能量等级)及对应的条件编码机制。可借鉴文本/音频驱动运动生成领域的风格控制技术[104-106]，实现风格可控的交互运动生成，将有助于提升交互运动的多样性。

3)基于场景条件的协同语音运动生成。复杂应用场景(虚拟发布会、直播带货)要求数字人同时满足场景约束、语音同步与自然手势。现有共语运动研究[107-108]尚未有效整合场景几何与语义信息，导致生成动作常违背物理规律(如手持虚拟产品时穿透展台)。未来需研究如何实现语音韵律-场景语义-生物力学约束的联合建模，通过跨模态注意力机制保障运动合规性。

7 结 论

本文主要对数据驱动的场景交互运动生成任务的研究现状进行全面梳理和总结，首先从场景数据和人体数据2个角度对3维数据表示进行简述，为深入理解相应的模型和算法提供预备知识，随后对该领域任务类型和模型方法进行归纳总结，并对常用的数据集和评价指标进行全面汇总，希望本文帮助对交互运动生成领域感兴趣的读者和领域内新的学者快速了解当前的进展和方向。

作者贡献声明：景宇宸负责文献调研和数据集整理，以及论文撰写和实验实现；温玉辉参与文献调研，统筹研究框架设计及修改；李晨光参与论文撰写及修改；于剑提供领域专家意见和关键理论指导。

参 考 文 献

[1] Cervantes P, Sekikawa Y, Sato I, et al. Implicit neural representations for variable length human motion generation[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 356−372

[2] Petrovich M, Black M J, Varol G. Action-conditioned 3D human motion synthesis with Transformer VAE[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10965−10975

[3] Barsoum E, Kender J, Liu Z. HP-GAN: Probabilistic 3D human motion prediction via GAN[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2018: 1418−1427

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

[5] Park S I, Shin H J, Shin S Y. On-line locomotion generation based on motion blending[C]//Proc of the ACM SIGGRAPH/Eurographics Symp on Computer Animation. New York: ACM, 2002: 105−111

[6] Shum H P H, Komura T, Shiraishi M, et al. Interaction patches for multi-character animation[J]. ACM Transactions on Graphics, 2008, 27(5): 1−8

[7] Agrawal S, van de Panne M. Task-based locomotion[J]. ACM Transactions on Graphics, 2016. https://dl.acm.org/doi/10.1145/

2897824.2925893

[8] Starke S, Zhang H, Komura T, et al. Neural state machine for character-scene interactions[J]. ACM Transactions on Graphics, 2019, 38(6): 1−14

[9] Holden D, Komura T, Saito J. Phase-functioned neural networks for character control[J]. ACM Transactions on Graphics, 2017, 36(4): 1−13

[10] Hassan M, Ghosh P, Tesch J, et al. Populating 3D scenes by learning human-scene interaction[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14703−14713

[11] Li Xueting, Liu Sifei, Kim K, et al. Putting humans in a scene: Learning affordance in 3D indoor environments[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 12360−12368

[12] Zhang Siwei, Zhang Yan, Ma Qianli, et al. PLACE: Proximity learning of articulation and contact in 3D environments[C]//Proc of the Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2020: 642−651

[13] Zhang Yan, Hassan M, Neumann H, et al. Generating 3D people in scenes without people[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6194−6204

[14] Zhao Kaifeng, Wang Shaofei, Zhang Yan, et al. Compositional human-scene interaction synthesis with semantic control[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 311−327

[15] Wang Jiashun, Xu Huazhe, Xu Jingwei, et al. Synthesizing long-term 3D human motion and interaction in 3D scenes[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9396−9406

[16] Wang Jingbo, Yan Sijie, Dai Bo, et al. Scene-aware generative network for human motion synthesis[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12206−12215

[17] Hassan M, Ceylan D, Villegas R, et al. Stochastic scene-aware motion prediction[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 11374−11384

[18] Wang Jingbo, Rong Yu, Liu Jingyuan, et al. Towards diverse and natural scene-aware 3D human motion synthesis[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20428−20437

[19] Wang Zan, Chen Yixin, Liu Tengyu, et al. HUMANISE: Language-conditioned human motion generation in 3D scenes[C]//Proc of the 36th Int Conf on Neural Information Processing Systems Advances in Neural Information Processing Systems. New York: Curran Associate, 2022: 14959−14971

[20] Du Tao, Hu Ruizhen, Liu Libin, et al. Research progress in human-like indoor scene interaction[J]. Journal of Image and Graphics, 2024, 29(6): 1575−1606 (in Chinese)
(杜韬, 胡瑞珍, 刘利斌, 等. 室内场景拟人交互研究进展[J]. 中国图象图形学报, 2024, 29(6): 1575−1606)

[21] Zhu Wentao Ma Xiaoxuan, Ro D, et al. Human motion generation: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2430−2449

[22] Badler N I, Phillips C B, Webber B L. Simulating Humans: Computer Graphics Animation and Control[M]. New York: Oxford University Press, 1993

[23] Kucuk S, Bingul Z. Robot kinematics: Forward and Inverse kinematics[M]. London, UK: INTECH Open Access Publisher, 2006

[24] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 851−866

[25] Romero J, Tzionas D, Black M J. Embodied hands: Modeling and capturing hands and bodies together[J]. ACM Transactions on Graphics, 2017, 36(6): 1−17

[26] Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3D hands, face, and body from a single image[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10967−10977

[27] Prokudin S, Lassner C, Romero J. Efficient learning on point clouds with basis point sets[C]//Proc of the IEEE/CVF Int Conf on Computer Vision Workshop. Piscataway, NJ: IEEE, 2019: 3072−3081

[28] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 165−174

[29] Mescheder L, Oechsle M, Niemeyer M, et al. Occupancy networks: Learning 3D reconstruction in function space[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4455−4465

[30] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99−106

[31] Araújo J P, Li J, Vetrivel K, et al. CIRCLE: Capture in rich contextual environments[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 21211−21221

[32] Cao Zhe, Gao Hang, Mangalam K, et al. Long-term human motion prediction with scene context[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 387−404

[33] Song Wenfeng, Zhang Xinyu, Li Shuai, et al. HOIAnimator: Generating text-prompt human-object animations using novel perceptive diffusion models[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognitio. Piscataway, NJ: IEEE, 2024: 811−820

[34] Cen Zhi, Pi Huaijin, Peng Sida, et al. Generating human motion in 3D scenes from text descriptions[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1855−1866

[35] Taheri O, Choutas V, Black M J, et al. GOAL: Generating 4D whole-body motion for hand-object grasping[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13253−13263

[36] Peng Shujuan, Zhou Bing, Liu Xin, et al. Recent advances in deep learning model for human motion generation[J]. Journal of Computer-Aided Design & Computer Graphic, 2018, 30(6): 1166−1176 (in Chinese)
(彭淑娟, 周兵, 柳欣, 等. 人体运动生成中的深度学习模型综述[J]. 计算机辅助设计与图形学学报, 2018, 30(6): 1166−1176)

[37] Corona E, Pumarola A, Alenya G, et al. Context-aware human motion prediction[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6990−7999

[38] Zhang Xiaohan, Bhatnagar B L, Starke S, et al. Couch: Towards controllable human-chair interactions[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 518−535

[39] Mao Wei, Liu miaomiao, Hartley R I, et al. Contact-aware human motion forecasting[C]//Proc of the 36th Int Conf on Neural Information Processing Systems Advances in Neural Information Processing Systems. New York: Curran Associate, 2022: 7356−7367

[40] Elman J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179−211

[41] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 135−1780

[42] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1724−1734

[43] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proc of the 28th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2672−2680

[44] Kingma D P, Welling M. Auto-encoding variational Bayes[J]. arXiv preprint, arXiv: 1312.6114, 2013

[45] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2020: 6840−6851

[46] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint, arXiv: 1710.10903, 2017

[47] Martinez J, Black M J, Romero J. On human motion prediction using recurrent neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4674−4683

[48] Eigen D, Ranzato M A, Sutskever I. Learning factored representations in a deep mixture of experts[J]. arXiv preprint, arXiv: 1312.4313, 2013

[49] Vaswani A. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2017: 6000−6010

[50] Yan Sijie, Li Zhizhong, Xiong Yuanjun, et al. Convolutional sequence generation for skeleton-based action synthesis[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4393−4401

[51] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models[C]//Proc of the 29th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 3483−3491

[52] Xu Tianhan, Takano W. Graph stacked hourglass networks for 3D human pose estimation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE , 2021: 16100−16109

[53] Hochreiter S. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735−1780

[54] Wu Yan, Wang Jiahao, Zhang Yan, et al. SAGA: Stochastic whole-body grasping with contact[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 257−274

[55] Ghosh A, Dabral R, Golyanik V, et al. IMoS: Intent-driven full-body motion synthesis for human-object interactions[J]. Computer Graphics Forum, 2023, 42(2): 1−12

[56] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[57] Huang Siyuan, Wang Zan, Li Puhao, et al. Diffusion-based generation, optimization, and planning in 3D scenes[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 13253−13263

[58] Li Jiaman, Wu Jiajun, Liu C K. Object motion guided human motion synthesis[J]. ACM Transactions on Graphics, 2023, 42(6): 1−11

[59] Jiang Nan, Zhang Zhiyuan, Li Hongjie, et al. Scaling up dynamic human-scene interaction modeling[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1737−1747

[60] Liu Ze, Zhang Zheng, Cao Yue, et al. Group-free 3D object detection via transformers[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 2929−2938

[61] Achlioptas P, Abdelreheem A, Xia Fei, et al. Referit3d: Neural listeners for fine-grained 3D object identification in real world scenes[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 422−440

[62] Wang Zan, Chen Yixin, Jia Baoxiong, et al. Move as you say interact as you can: Language-guided human motion generation with scene affordance[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 433−444

[63] Yi Hongwei, Thies J, Black M J, et al. Generating human interaction motions in scenes with text control[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 246−263

[64] Zhang Yan, Tang Siyu. The wanderings of odysseus in 3D scenes[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 20449−20459

[65] Zhao Kaifeng, Zhang Yan, Wang Shaofei, et al. Synthesizing diverse human motions in 3D indoor scenes[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 14692−14703

[66] Li Jiaman, Clegg A, Mottaghi R, et al. Controllable human-object interaction synthesis[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 54−72

[67] Zeng Lingan, Huang Guohong, Wei Yilin, et al. ChainHOI: Joint-

based kinematic chain modeling for human-object interaction generation[J]. arXiv preprint, arXiv: 2503.13130, 2025

[68] Karunratanakul K, Yang Jinlong, Zhang Yan, et al. Grasping field: Learning implicit representations for human grasps[C]//Proc of the Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2020: 333−344

[69] Christen S, Hampali S, Sener F, et al. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions[C]//Proc of the SIGGRAPH Asia Conf. New York: ACM, 2024: 1−11

[70] Bohg J, Morales A, Asfour T, et al. Data-driven grasp synthesis−A survey[J]. IEEE Transactions on Robotics, 2013, 30(2): 289−309

[71] Napier J R. The prehensile movements of the human hand[C]//Proc of the Classic Papers in Orthopaedics. Berlin: Springer, 2014: 343−345

[72] Hsiao K, Lozano-Perez T. Imitation learning of whole-body grasps[C]//Proc of the IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2006: 5657−5662

[73] Taheri O, Ghorbani N, Black M J, et al. GRAB: A dataset of whole-body human grasping of objects[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 581−600

[74] Bhatnagar B L, Xie X, Petrov I A, et al. BEHAVE: Dataset and method for tracking human object interactions[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 15914−15925

[75] Xu Sirui, Li Zhengyuan, Wang Yuxiong, et al. InterDiff: Generating 3D human-object interactions with physics-informed diffusion[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 14928−14940

[76] Huang Yinghao, Taheri O, Black M J, et al. InterCap: Joint markerless 3D tracking of humans and objects in interaction[C]//Proc of the 44th DAGM German Conf. Berlin: Springer, 2024, 132: 2551−2566

[77] Ahuja C, Lee D W, Nakano Y I, et al. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springe, 2020: 248−265

[78] Wang Xi, Li Gen, Kuo Yenling, et al. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors[C]//Proc of the Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2022: 353−362

[79] Sung K, Shirley P, Rosenberg B R. Experiencing aspects of games programming in an introductory computer graphics class[J]. ACM SIGCSE Bulletin, 2007, 39(1): 249−253

[80] Hanser E, Mc Kevitt P, Lunney T, et al. Scenemaker: Intelligent multimodal visualisation of natural language scripts[C]//Proc of the 20th Irish Conf on Artificial Intelligence and Cognitive Science. Berlin: Springer, 2009: 144−153

[81] Charles R Q, Su Hao, Kaichun M, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 77−85

[82] Zhao Hengshuang, Jiang Li, Jia Jiaya, et al. Point Transformer[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 16239−16248

[83] Hart P E, Nilsson N J, Raphael B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE Transactions on Systems Science and Cybernetics, 1968, 4(2): 100−107

[84] Tevet G, Gordon B, Hertz A, et al. Motionclip: Exposing human motion generation to clip space[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 358−374

[85] Jiang Biao, Chen Xin, Liu Wen, et al. Motiongpt: Human motion as a foreign language[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates 2023: 20067−20079

[86] Guo Chuan, Mu Yuxuan, Javed M G, et al. Momask: Generative masked modeling of 3D human motions[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 1900−1910

[87] Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the NAACL-HLT. Stroudsburg, PA: ACL, 2019: 4171−4186

[88] Mandery C, Terlemez Ö, Do M, et al. The KIT whole-body human motion database[C]//Proc of the Int Conf on Advanced Robotics. Piscataway, NJ: IEEE, 2015: 329−336

[89] Savva M, Chang A X, Hanrahan P, et al. PiGraphs: Learning interaction snapshots from observations[J]. ACM Transactions on Graphics, 2016. https://dl.acm.org/doi/10.1145/2897824.2925867

[90] Nießner M, Zollhöfer M, Izadi S, et al. Real-time 3D reconstruction at scale using voxel hashing[J]. ACM Transactions on Graphics, 2013, 32(6): 1−11

[91] Hassan M, Choutas V, Tzionas D, et al. Resolving 3D human pose ambiguities with 3D scene constraints[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 2282−2292

[92] Monszpart A, Guerrero P, Ceylan D, et al. iMapper: Interaction-guided scene mapping from monocular videos[J]. ACM Transactions on Graphics, 2019, 38(4): 1−15

[93] Guzov V, Mir A, Sattler T, et al. Human poseitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4318−4329

[94] Juliani A, Berges V P, Teng E, et al. Unity: A general platform for intelligent agents[J]. arXiv preprint, arXiv: 1809.02627, 2018

[95] Mahmood N, Ghorbani N, Troje N F, et al. AMASS: Archive of motion capture as surface shapes[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5441−5450

[96] Dai A, Chang A X, Savva M, et al. ScanNet: Richly-annotated 3D reconstructions of indoor scenes[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5828−5839

[97] Jiang Nan, Liu Tengyu, Cao Zhexuan, et al. Full-body articulated human-object interaction[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 9331−9342

[98] Chang A, Dai A, Funkhouser T, et al. Matterport3D: Learning from RGB-D data in indoor environments[C]//Proc of Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2017: 667−676

[99] Straub J, Whelan T, Ma Lingni, et al. The Replica dataset: A digital replica of indoor spaces[J]. arXiv preprint, arXiv: 1906.05797, 2019

[100] Ionescu C, Papava D, Olaru V, et al. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325−1339

[101] Cai Zhongang, Ren Daxuan, Zeng Ailing, et al. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2022: 557−577

[102] Lin Jing, Zeng Ailing, Lu Shunlin, et al. Motion-x: A large-scale 3D expressive whole-body human motion dataset[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associate, 2023: 25268−25280

[103] Guo Chuan, Zuo Xinxin, Wang Sen, et al. Action2motion: Conditioned generation of 3D human motions[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 2021−2029

[104] Ao Tenglong, Zhang Zeyi, Liu Libin. GestureDiffuCLIP: Gesture diffusion model with clip latents[J]. ACM Transactions on Graphics, 2023, 42(4): 1−18

[105] Yang Sichen, Wu Zhiyong, Li Minglei, et al. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models[C]//Proc of the 3rd Int Joint Conf on Artificial Intelligence. Berlin: Springer, 2023 5860−5868

[106] Li Chenguang, Wen Yuhui, Jing Yuchen, et al. Shape-aware stylized dance motion generation driven by music[J/OL]. Journal of Computer-Aided Design & Computer Graphics, 2024[2025-03-02]. https://www.jcad.cn/cn/article/pdf/preview/10.3724/SP.J.1089.2024-00492.pdf (in Chinese)
(李晨光, 温玉辉, 景宇宸, 等. 体型感知的音乐驱动舞蹈动作风格化生成[J/OL]. 计算机辅助设计与图形学学报, 2024[2025-03-02]. https://www.jcad.cn/cn/article/pdf/preview/10.3724/SP.J.1089.2024-00492.pdf)

[107] Yang Sichen, Xu Zunnan, Xue Haiwei, et al. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness[C]//Proc of the IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2024: 7945−7949

[108] Chen Bohong Li Yumeng, Ding Yaoxiang, et al. Enabling synergistic full-body control in prompt-based co-speech motion generation[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 6774−6783



Jing Yuchen, born in 2000. PhD candidate. Student member of CCF. His research interests include computer vision, 3D motion representation and generation, continuous learning, AI, and machine learning.

景宇宸, 2000 年生。博士研究生。CCF 学生会员。主要研究方向为计算机视觉、3 维动作表示与生成、持续学习、人工智能、机器学习。



Wen Yuhui, was born in 1990. PhD, associate professor. Member of CCF. Her research interests include computer vision, computer graphics, and machine learning.

温玉辉, 1990 年生。博士, 副教授。CCF 会员。主要研究方向为计算机视觉、计算机图形学、机器学习。



Li Chenguang, born in 2002. PhD candidate. Student member of CCF. His research interests include computer vision, 3D motion representation and generation, and machine learning.

李晨光, 2002 年生。博士研究生。CCF 学生会员。主要研究方向为计算机视觉、3 维动作表示与生成、机器学习。



Yu Jian, born in 1969. PhD, professor, PhD supervisor. Fellow of CCF. His main research interests include artificial intelligence and machine learning.

于　剑, 1969 年生。博士, 教授, 博士生导师。CCF 会士。主要研究方向为人工智能、机器学习。