CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement

Chengwen Zhang*1,2 , Yun Liu*1,3,4 , Ruofan Xing1 , Bingda Tang1 , Li Yi

摘要

了解如何生成人与物交互在 VR/AR 和人机交互领域至关重要用。然而，对建模的深入研究由于缺乏相关数据集，使得这些方面的研究不足。本文提出一个新的数据集 CORE4D，这是一种新型的大规模 4D 人体物体交互数据集侧重于协作对象重新排列，包括各种对象几何形状、协作模式的不同组合，以及 3D 场景。在 1K 人体对象中捕获人体运动序列在现实世界中，通过提供迭代协作重定向策略来丰富 CORE4D 数据集，以增强各种新对象的运动。利用这一点 CORE4D 方法包括总共 11K 个协作序列，跨越 3K 真实和虚拟对象形状。本文针对文章的 baseline 进行复现，并提出一个新的任务进行实验。数据集和代码链接为 https://github.com/leolyliu/CORE4D-Instructions

关键词：扩散模型；人体模型

1 引言

人类经常通过多人协作重新安排家居用品，例如移动或者一起捡起一把翻倒的椅子。分析和综合这些不同的协作行为可以广泛应用于 VR/AR、人机交互，以及灵巧和类人操纵。然而，理解和由于缺乏大规模、丰富的建模方法，对这些交互式运动的建模研究不足带注释的数据集。大多数现有的人类对象和手对象交互数据集都集中在个人行为和两人交接上。但是这些数据集通常包含有限数量的对象实例，因此难以支持跨不同对象几何的通用交互理解。放大精确的人体物体交互数据具有挑战性。而基于视觉的人体运动跟踪方法尽管已经取得了重大进展，但在严重闭塞的情况下，他们仍然难以保持低保真度，这是常见的在多人协作场景中。然而，光学动捕设备 (mocap) 价格昂贵，难以扩大规模以覆盖大量要重新排列的对象。本文章 [2] 希望策划一个大规模的类别级别以高效的方式提供具有高运动质量的人-物-人（HOH）交互数据集，并且基于此数据集提出人与物体交互 baseline。

2 相关工作

2.1 人与物体交互生成

人机交互生成是一个新兴的研究课题，旨在综合现实受周围 3D 场景、已知物体轨迹或动作影响的人体物体运动类型。为了生成与静态 3D 场景交互的人类，有使用条件自编码器

(CVAE) 进行静态人体姿势生成。为了生成动态的人体运动，通过自回归方式、扩散模型或两阶段设计，首先生成开始和结束姿势，然后在两者之间插值生成运动。

3 本文方法

3.1 本文方法概述

此部分对本文将要复现的工作进行概述，本文针对数据集提出了人体动作生成 baseline。baseline 算法主要使用 MDM [1]，其网络结构如图 1所示：，基于条件扩散模型来生成动作序列。主要任务的输入为给定物体的运动序列，由于 MDM 的网络输入条件为文本描述，所以将其条件改为物体，经过物体的 BPS-Encoder 编码后输入到 diffusion 扩散模型生成人体的姿态参数，通过 SMPL-X 模型生成逼真的 3D 人体运动，最终生成多人与物体交互的运动序列。

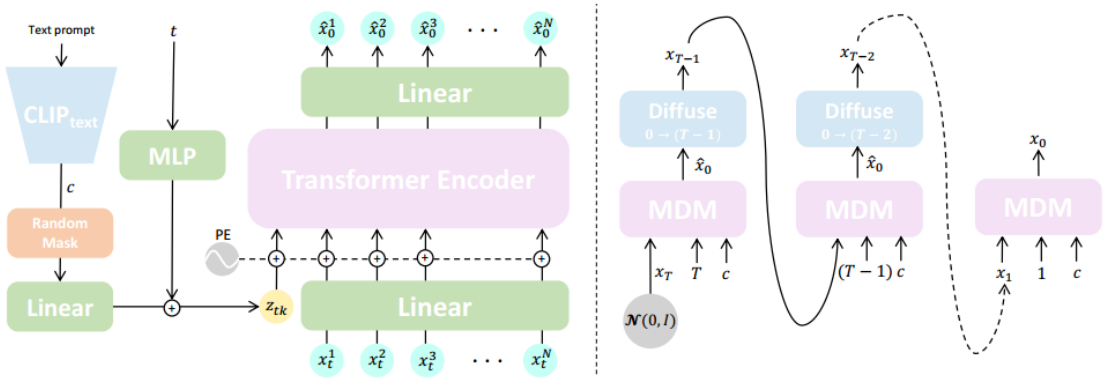

图 1. MDM 模型示意图

4 复现细节

4.1 与已有开源代码对比

复现的代码链接为：https://github.com/leolyliu/CORE4D-Instructions。首先对本文提出的 baseline 多人交互生成代码进行复现，随后在此复现工作基础上，本工作提出一个新的任务，由于多人协同物体交互任务较为复杂，为了实现更加复杂的交互情况，提出对于给定特定人与物体交互的情况，生成另外一个人同给定人与物体交互的任务。该任务相比之前只给定物体的运动序列更加复杂，主要研究在于如何将人体运动表征与物体作为条件输入到扩散模型中。

4.2 代码复现

首先，对于原本代码的实验环境，通过 conda 创建环境，安装 requirments 文件所需的 python 库。并且下载复现工作提出的数据集，总共大小为 243GB，对数据集进行预处理，随后对代码进行复现，包括数据集训练，模型权重保存，评价指标测试复现。在训练上，代码在搭载 Nvidia RTX-3090 显卡的服务器上进行部署训练，经过约一天的时间训练完毕。设置每一万步保存权重，总共训练步数为 40 万步。训练 Loss 从 1.05 下降到 0.010 左右。在评价指

标上有 MPVPE(Mean per-vertex position error)，MPJPE(Mean per-joint position error) 和 Contact Accuracy。文中的 MDM 生成指标测试结果如图 2中红框所示。

Table 3: **Quantitative results on interaction synthesis.**

Test Set	Method	$RR.J_e$ (mm, ↓)	$RR.V_e$ (mm, ↓)	C_{acc} (%, ↑)	FID (↓)
S1	MDM [73]	138.0 (± 0.3)	194.6 (± 0.2)	76.9 (± 0.5)	7.7 (± 0.2)
	OMOMO [40]	137.8 (± 0.2)	196.7 (± 0.3)	78.2 (± 0.5)	8.3 (± 0.6)
S2	MDM [73]	145.9 (± 0.2)	208.2 (± 0.2)	76.7 (± 0.1)	7.7 (± 0.2)
	OMOMO [40]	145.2 (± 0.6)	209.9 (± 1.0)	77.8 (± 0.3)	8.3 (± 1.0)

图 2. MDM 性能测试结果

接下来载入训练好的模型权重，对 Test Set 数据集进行测试，其分为 S1 和 S2, 分别对应有训练过的物体场景和无训练过的物体场景，最终的测试结果如图 3所示。可以看到复现的结果都在文中的指标测试结果范围内。

```
###### overall score ######
[seen]
mean MPVPE (mm) = 197.026123046875
mean MPJPE (mm) = 139.54905700683594
mean contact accuracy (%) = 76.94047689437866
[unseen]
mean MPVPE (mm) = 207.9473114013672
mean MPJPE (mm) = 145.74102783203125
mean contact accuracy (%) = 77.02258825302124
```

图 3. 复现结果

4.3 新的任务实验

接下来对新的任务进行实验，主要研究如何加入已知人体的运动表征，SMPL-X 的人体顶点总数是 10475 个，显然将整个人体顶点作为输入对于需要学习的参数过于庞大冗余，不是很好的解决方案。SMPL-X 还有需要关于人体的参数，而在运动中，主要能代表人体运动的参数为人体关节，SMPL-X 的人体关节总共为 24，可以作为模型的输入，最终将人体关节位置参数 24*3 和旋转矩阵 22*3*3 作为模型的输入，人体的参数维度不大，直接作为扩散模型的条件输入即可，物体则是转为 BPS 编码最后输入到 BPS-Encoder 后作为扩散模型的条件输入，最终将两者拼接共同作为扩散模型的条件输入。在训练过程每一万步保存一个模型权重，训练到 40 万步停止。接下来对生成的人体运动进行测试，以下为一段截取关键帧的人体运动可视化展示，总共生成的运动帧数量为 120 帧，截取第 0 帧，第 30 帧，第 60 帧，第 120 帧作为可视化展示，如图 4所示。图中蓝色的人和物体为已知的一部分，并作为条件扩散模型的输入，绿色的人为扩散模型生成的。

图 4. 可视化结果

最终同样对模型的性能进行测试，评价指标为 MPVPE(Mean per-vertex position error)，MPJPE(Mean per-joint position error) 和 Contact Accuracy。最终的测试结果如图 5所示

```
###### overall score ######
[seen]
mean MPVPE (mm) = 184.15085
mean MPJPE (mm) = 127.84504
mean contact accuracy (%) = 77.66919136047363
[unseen]
mean MPVPE (mm) = 193.94376
mean MPJPE (mm) = 134.64333
mean contact accuracy (%) = 79.0231704711914
```

图 5. 复现结果

可以看到，相比于双人的生成任务，单人的生成任务会更加精准。

5 总结与展望

本部分对整个文档的内容进行归纳并分析目前实现过程中的不足以及未来可进一步进行研究的方向。首先对本文的评价指标性能进行复现确认，随后提出一个新的生成任务，可以较好地生成人体的运动，也可以看到误差仍较大，未来可以继续研究如何提升生成的人体运动的准确性。

参考文献

[1] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353*, 2024.