基于文本控制优化的人体动作生成研究

摘要

随着扩散模型的出现，基于文本提示的人体运动合成取得了显著的进步。在最近的部分工作中，将灵活的空间控制信号合并到基于扩散过程的文本条件人体运动生成模型，实现了在不同时间在不同关节上的灵活控制。本文工作聚焦于复现 Xie 等人设计的 OmniControl 方法，在实现对于人体关节的精确控制上，进一步对模型的 ControlNet 微调架构、训练超参数实现了优化创新，并尝试引入运动学引导优化动作真实性。具体而言，模型的改动主要面向模型微调模块的编码架构。并且，新模型在 HumanML3D 和 KIT-ML 两项数据集上的可视化实验表明，优化后的模型在保留原先特定关节控制的基础上，能够对动作的文本描述有更好的理解能力。

关键词：人体动作生成；扩散模型；文本驱动；多模态

1 引言

人体动作合成旨在通过输入文本、音频、场景等驱动生成人体动作，在机器人控制、电影和动画制作等领域有着广泛的应用。在这些输入信号中，文本，即自然语言，可以提供丰富的语义细节，是最人性化、最方便的运动合成信号。

生成单人动作作为文本转动作（T2M）的一项基本任务，受到了研究界的广泛关注。先前有许多方法 [2, 6] 利用 VAE 结构在同一空间中对齐文本和运动，另一类方法 [13, 16, 17] 则利用以文本为条件的扩散模型，使生成的运动更加多样化和真实。随着应用场景的不断扩大，一些作品开始探索两人或者多人的动作合成。ComMDM [10] 方法通过少量学习在只有 20 个两人文本注释动作的数据集上微调了预训练的单人动作生成模型。InterGen [5] 进一步引入了一个大型带注释的两人运动数据集 InterHuman，并采用共享权重扩散网络和交叉注意力来模拟两人运动的交互。

上述方法在单人或是多人运动生成方面都取得了显着的成就，但仍然有局限性：（1）虽然一些研究已经完成了运动生成，但将空间信号纳入当前的运动生成模型仍然是一个不小的挑战。（2）基于扩散模型的方法可以生成多样化且逼真的人体运动，但无法生成灵活的人体运动。对许多应用至关重要的空间控制信号，例如要合成拿起杯子的动作，模型不仅要在语义上理解“拿起”，还要控制手的位置在特定位置和时间触摸杯子。

Xie 等人 [12] 提出的 OmniControl 模型引入了空间和真实度控制来优化人体运动的生成，实现了任何时刻对任意关节的灵活控制。在本文中，我们首先复现了 Xie 等人 [12] 提出的 OmniControl 模型，其原理是利用 ControlNet [14] 作为条件控制传入和微调模块，对先前的动作生成模型 Motion-Diffusion-Model（MDM）[11] 进行控制信号的传入，以实现在空间位置

上引导某一关节运动的功能。在引入空间引导过程中，采用了 L2 范数去衡量指定的运动关节与所需空间位置的对齐程度，并通过运动学函数将关节的局部位置转换为全局位置。这个空间信号作为 ControlNet [14] 的控制条件输入，经过空间编码器后进入微调模块的自注意力层和零卷积层后，作为控制条件传入 MDM 模型 [11] 的扩散过程中，最后实现对运动序列的预测。

在复现过程中，我们发现在获取空间控制信号要经过一系列复杂的函数运算，并且在三维空间中的全局位置获取难度较大，一定程度上对模型的复杂度有较大影响。因此，我们首先尝试采用文本代替空间信号的方法，再经过 CLIP [7] 等文本编码方式实现对真实度的控制工作。但是我们又注意到仅采用对比学习的文本对齐方式无法很好的实现对关节的精确控制，因此我们进一步采用了运动学控制和 Mask 的方式，结合细粒度文本数据集对控制信号进行了优化，实现了动作的精确控制。

总结来说，本文的工作有以下几点：(1) 使用 MDM 模型和 ControlNet 模块复现了 OmniControl 原本的效果（2）用文本和文本编码器替换原先的空间控制模块作为 ControlNet 部分的输入，实现了仅用文本驱动的人体运动控制生成（3）采用不同规模的细粒度动作文本数据集对模型进行了训练和测试，探究了文本编码器的选用和数据规模对模型性能的影响。

剩余的章节将按以下逻辑展开，在第二章介绍与本文相关的工作，第 3 章详细介绍本文的方法，第 4 章介绍复现的细节以及本文创新点，第五章进行实验并分析结果，最后总结本文的结论并提出展望。

2 相关工作

2.1 人体动作生成任务

人体运动合成可大致分为两类：自回归方法 [9] 和序列级方法 [11]。自回归方法使用过去运动的信息逐帧递归地生成当前运动。这些方法主要是针对实时场景而定制的。相反，序列级方法被设计为生成整个固定长度的运动序列。由于这一固有特征，它们可以与现有的生成模型无缝集成，例如 VAE [1] 和扩散模型 [15]。对于人体运动的提示词，通常源自各种外部模态，例如文本、音频、图像、运动轨迹、3D 场景和特定对象等。

尽管合并空间约束是一个基本特征，但它对于基于文本的人体运动合成方法仍然是一个挑战。理想的方法应该保证产生的运动紧密遵循全局空间控制信号，与文本语义保持一致，并保持保真度。这种方法还应该能够控制任何关节及其组合，以及处理稀疏控制信号。PriorMDM [10] 和 MDM [11] 使用基于修复的方法将空间约束输入到生成的运动中。然而，受到相对人体姿势表示的限制，其中其他关节的位置是通过 w.r.t 定义的。对于骨盆，这些方法很难纳入除骨盆之外的其他关节的全局约束并处理稀疏的空间约束。尽管基于修复的方法 GMD [3] 引入了两阶段引导运动扩散模型来处理稀疏控制信号。它仍然面临着将空间限制纳入任何其他关节的挑战。在本文中，我们专注于序列级运动生成，并提出了一种新颖的方法，即使使用稀疏的控制信号，也可以使用单个模型控制任何关节。

2.2 基于扩散模型的可控生成

最近，基于扩散的生成模型因其在图像生成方面卓越的性能而受到了广泛关注。扩散模型非常适合控制和调节，有多种条件生成方法。例如图像修复任务中，用观察到的数据来填补的缺失部分，使得填充的内容在视觉上与周围区域一致。然而，当观察到的数据与填充部分位于不同的空间时，仅通过语义图生成对应的图像是很困难的。在有分类器指导的方法下，通过利用训练单独的分类器来改进条件扩散生成模型。而在无分类器指导方法中，通常通过联合训练条件和无条件扩散模型，并将它们结合起来，以实现样本质量和多样性之间的权衡。例如 GLIGEN 模型 [4] 在每个 transformer 块上添加了一个可训练的门控自注意力层，以吸收新的接地输入。

基于扩散模型的可控合成的另一个代表模型是 ControlNet [14]。它引入了一种旨在控制大型图像扩散模型的神经网络，能够以最少的数据和训练快速适应特定于任务的控制信号。这些控制方法并不是相互排斥的，单独采用一种方法可能达不到预期的目的。受分类器引导和 ControlNet 的启发，我们设计了由空间引导和现实主义引导组成的混合引导，将空间控制信号纳入人体运动生成中。空间引导应用分析函数来近似分类器，从而能够对生成的运动进行多种有效的扰动。同时，真实感引导使用类似于 ControlNet 的神经网络来调整输出以生成连贯且真实的运动。这两个引导模块都是必不可少的，并且在平衡运动真实性和控制精度方面具有高度互补性。

3 本文方法

3.1 OmniControl 模型回顾

OmniControl [12] 方法根据文本提示和空间控制信号生成人体动作。在去噪扩散步骤中，采用了 MDM [11] 的基本架构，模型将文本提示和带噪声的运动序列 x_t 作为输入，与一般的扩散模型不同的是，MDM 模型 [11] 在每一个时间步后都估计无噪声的运动序列 x_0。同时，OmniControl [12] 为了将灵活的空间控制信号融入到生成过程中，采用了由真实感和空间引导组成的混合引导，以鼓励运动在真实的同时符合控制信号。其模型整体架构如图 1所示：

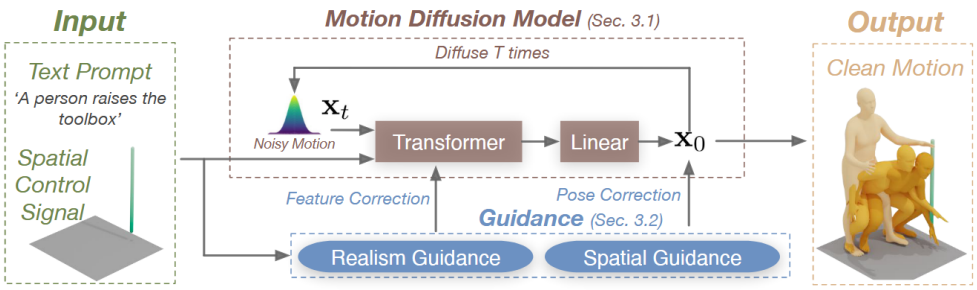


图 1. Xie 等人 [12] 提出的 OmniControl 模型

空间引导的架构如图 2所示。空间引导的核心是一个分析函数 $G(\mu_t, c)$，它评估生成的运动的关节与所需空间位置 c 的对齐程度。参考 Dhariwal 和 Nichol (2021) 的观点，利用解析函数的梯度来引导生成的运动向所需的方向移动。我们使用空间引导来为每个去噪步骤 t 的预测平均值加噪：

$$\mu_t = \mu_t - \tau \nabla_{\mu_t} G(\mu_t,), \tag{1}$$

其中 τ 控制引导的强度。G 测量生成运动的关节位置与空间约束之间的 L2 距离：

$$G(\mu, c) = \frac{\sum_n \sum_j \sigma_{nj} \left\| c_{nj} - \mu^g_{nj} \right\|_2}{\sum_n \sum_j \sigma_{nj}}, \ \mu^g = R(\mu), \tag{2}$$

其中 σ_{nj} 是一个二进制值，表示空间控制信号 c 是否包含帧 n 处关节 j 的有效值。$R(\cdot)$ 将关节的局部位置转换为全局绝对位置。为了简单起见，在这里省略了扩散去噪步骤 t 。在这种情况下，骨盆关节点在特定帧的全局位置可以通过所有先前帧的旋转和平移的累加来确定。通过骨盆位置和其他关节的相对位置的累加，还可以确定其他关节的位置。

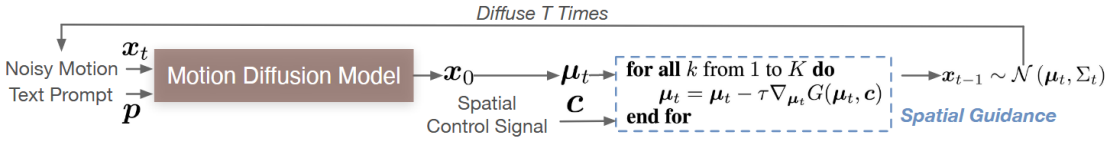


图 2. OmniControl 中空间控制细节示意图

尽管空间引导可以有效地强制受控关节遵守输入控制信号，但它可能会生成不符合人体运动规律的结果，为了解决这个问题，受到 Zhang 等人的启发。(2023b)，OmniControl 还提出真实度控制。

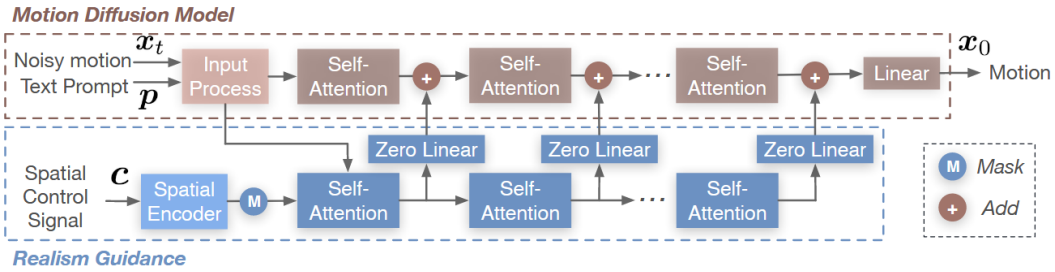


图 3. OmniControl 中真实性控制细节示意图

具体来说，真实度控制是运动扩散模型中 Transformer 编码器的可训练副本，用于学习执行空间约束。真实感引导的架构如图3所示。真实感引导采用与运动扩散模型相同的文本提示 p 以及空间控制信号 c。原始模型和新的可训练副本中的每个 Transformer 层都通过线性层连接，权重和偏差都初始化为零，因此它们在开始时没有控制效果。随着训练的进行，真实感引导模型学习空间约束，并将学习到的特征校正添加到运动扩散模型中的相应层，以隐式修改生成的运动。

3.2 本文方法概述

此部分对本文采取的优化方法进行简述。如图4所示，在图3呈现出的 OmniControl 细节示意图的基础上，将空间控制模块（Spatial Control Signal）用 Vanilla CLIP [7] 文本编码器代替，用于处理输入的细粒度文本，以实现对关节点的控制。

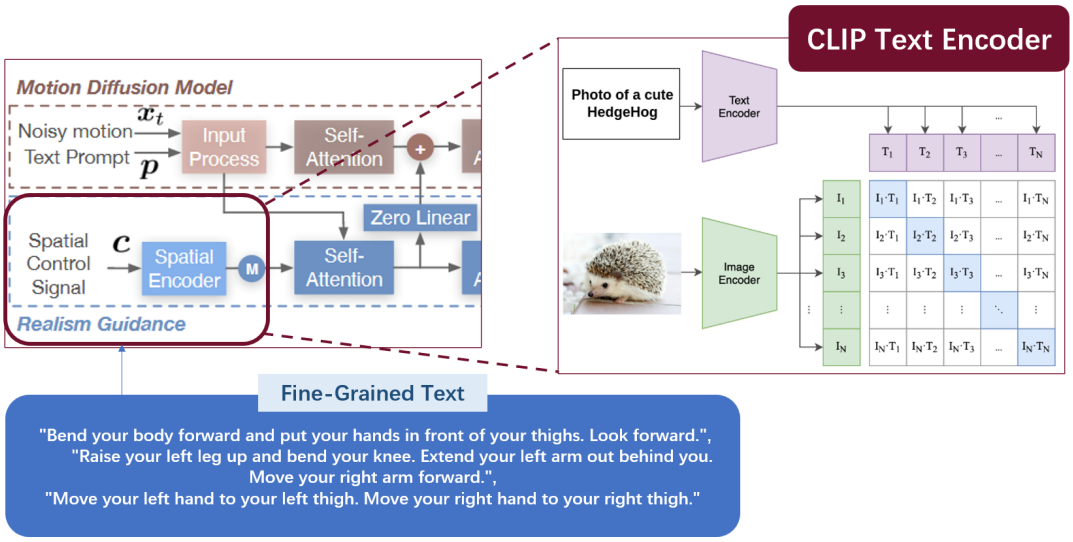


图 4. 基于 CLIP [7] 文本编码器对 OmniControl 控制模块的优化示意图

此外，考虑到 CLIP [7] 预训练模型依赖对图像知识的先验学习，本次实验中最后用于替换的语义编码模型替换为 Text-to-Text Transfer Transformer(T5) [8] 模型。T5 模型中，自然语言处理的每项任务（包括翻译、问答和分类）都被视为将模型文本作为输入并训练它生成一些目标文本，这使得模型能够很好的迁移到本次工作的文本处理中。

具体而言，本文方法实现的过程有以下几个过程：(1) 首先处理好原先用于 OmniControl 的 HumanML3D 数据集，并导入细粒度文本数据集作为控制信号 (2) 训练基于 T5 的自然语言模型，用于代替空间控制编码器的文本编码器 (3) 文本特征 mask 工作，使得模型能理解细粒度文本中涵盖的对特定关节的控制识别。具体的复现细节将在下一部分复现细节中的创新点详细展开说明。

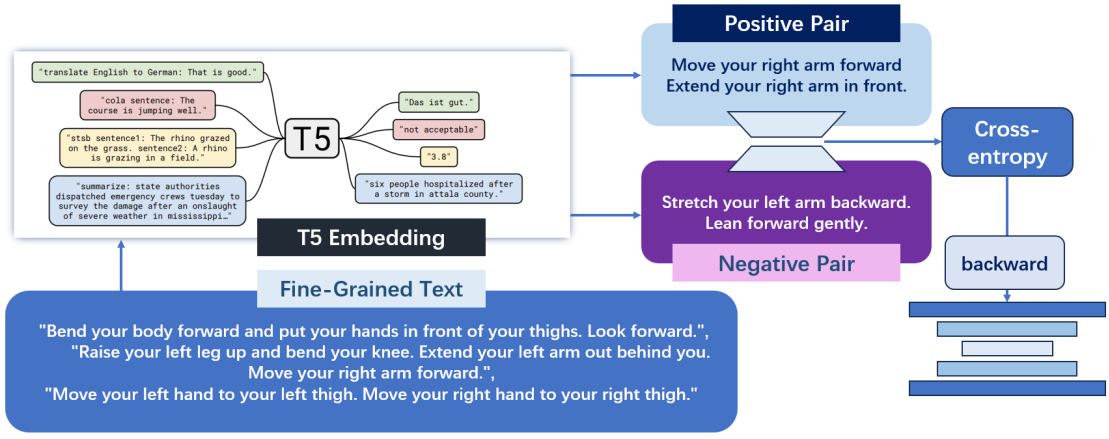


图 5. 基于 T5 语言模型 [8] 对 OmniControl 控制模块的优化示意图

3.3 重定义空间损失

由于用文本编码代替了原先在三维空间中的坐标点来实现对人体关节点的控制工作，因此原先的空间引导函数不再适用。在训练过程中，这一部分的优化目标函数重写为文本的对

齐损失，即

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(p_i\right) + \left(1-y_i\right)\log\left(1-p_i\right)\right] \tag{3}$$

其中，N 为样本对的总数，$y_i \in \{0,1\}$ 是第 i 个样本对的真实标签。p_i 是模型对第 i 个样本对预测为正样本（即对齐）的概率。

4 复现细节

4.1 与已有开源代码对比

本文所复现的内容基于 Xie 等人 [12] 提出的 OmniControl 模型，旨在实现 3D 人体动作生成过程中对关节点的精确控制。作者开源了项目代码在链接 https://neu-vi.github.io/omnicontrol/中。

在复现过程中，主要修改源代码数据集的加载、ControlMDM 模型定义、模型前向传播以及训练主函数中的扩散模型分支，代码的详细信息和改动内容如表1所示

表 1. 代码文件变动说明

代码文件	改动描述
get_data.py	更改原有空间控制信号 hint 为 detailed_text
dataset.py	改变原有的 ___getitem()___ 方法
cmdm.py	调整传入控制信号时张量的合成逻辑
guassian_diffusion.py	修改原先空间引导 spatial guidance 方法
train_mdm.py	调整 batchsize 和学习率退火策略

4.2 实验环境搭建

本文中的所有实验均在 Linux 系统的服务器下运行。所有复现代码均由 Python 3.7 语言实现，模型框架选用 Torch 1.7.1 + CUDA 12.1，其他硬件配置详细信息如表2所示。

表 2. 实验环境信息表

配置名称	详细信息
编译器	python 3.7.13
操作系统	Linux Ubuntu 20.04
CPU	AMD EPYC 7742 64-Core Processor
GPU	A100-SXM4-40GB

实验的环境配置具体过程如下：

(1) 安装 ffmpeg。在终端中输入命令：

sudo apt update

sudo apt install ffmpeg

(2) 初始化 conda 环境，依次输入命令：

conda env create −f environment.yml

conda activate omnicontrol

python −m spacy download en_core_web_sm

pip install git+https://github.com/openai/CLIP.git

(3) 通过 shell 脚本文件安装相关依赖：

bash prepare/download_smpl_files.sh

bash prepare/download_t2m_evaluators.sh

4.3 创新点

本工作在复现 Xie 等人 [12] 的 OmniControl 模型的基础上，针对模型、数据类型、训练策略都进行了创新，有一部分内容已经通过实验证实能在模型性能上实现一定程度的优化，主要如下：

（1）选用细粒度文本代替原先的空间控制信号作为微调条件的输入

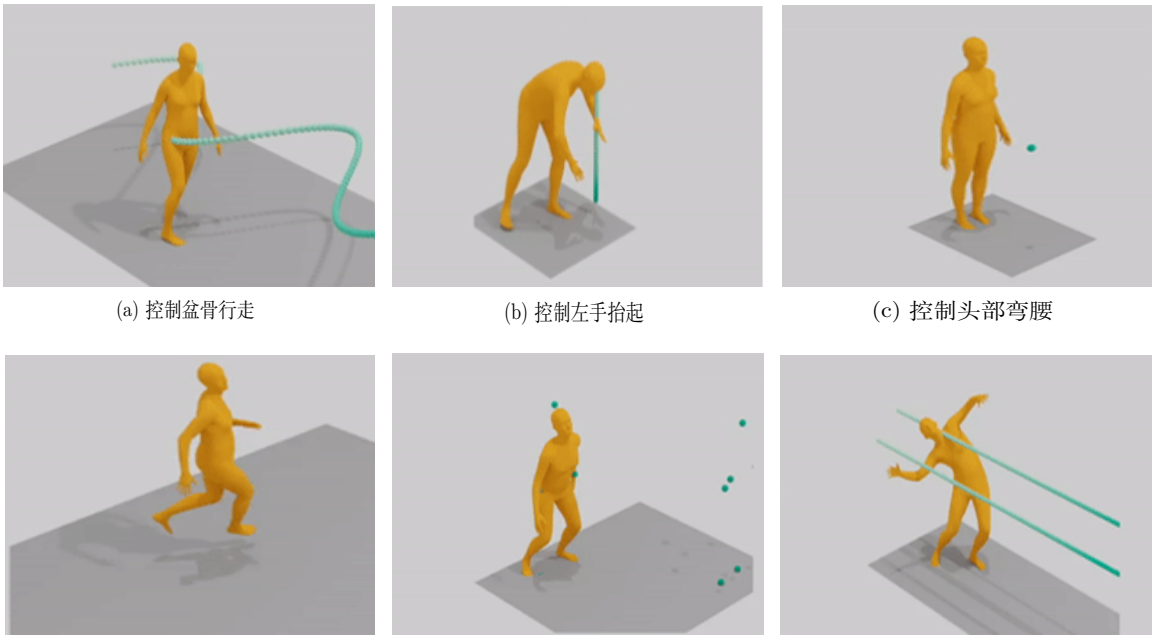
（2）采用对比学习模式重新训练了一个文本编码器，用于对控制文本进行编码和维度对齐

（3）更改了部分超参数和训练策略，一定程度上模型训练的显存消耗降低，提高了速度

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

5.1 定性复现与对比分析



(a) 控制盆骨行走 (b) 控制左手抬起 (c) 控制头部弯腰

(d) 控制骨盆跑步 (e) 空间控制移动 (f) 左手 + 头部控制行走

图 6. OmniControl 模型定性复现可视化结果

在完成对 OmniControl 模型的复现任务后，首先在 HumanML3D 数据集上对模型控制的定性实验。该过程主要经过以下几个关键步骤：

(1) **模型训练**

运行训练代码 train_mdm.py。在终端的运行指令中设置超参数，并且下载预训练模型进行训练，运行命令为:

$python -m\ train.train\_mdm --save\_dir\ save/my\_omnicontrol --dataset\ humanml --num\_steps\ 400000\ --batch_size\ 64\ --resume\_checkpoint\ ./save/model000475000.pt\ --lr\ 1e-5$

(2) **模型推理**

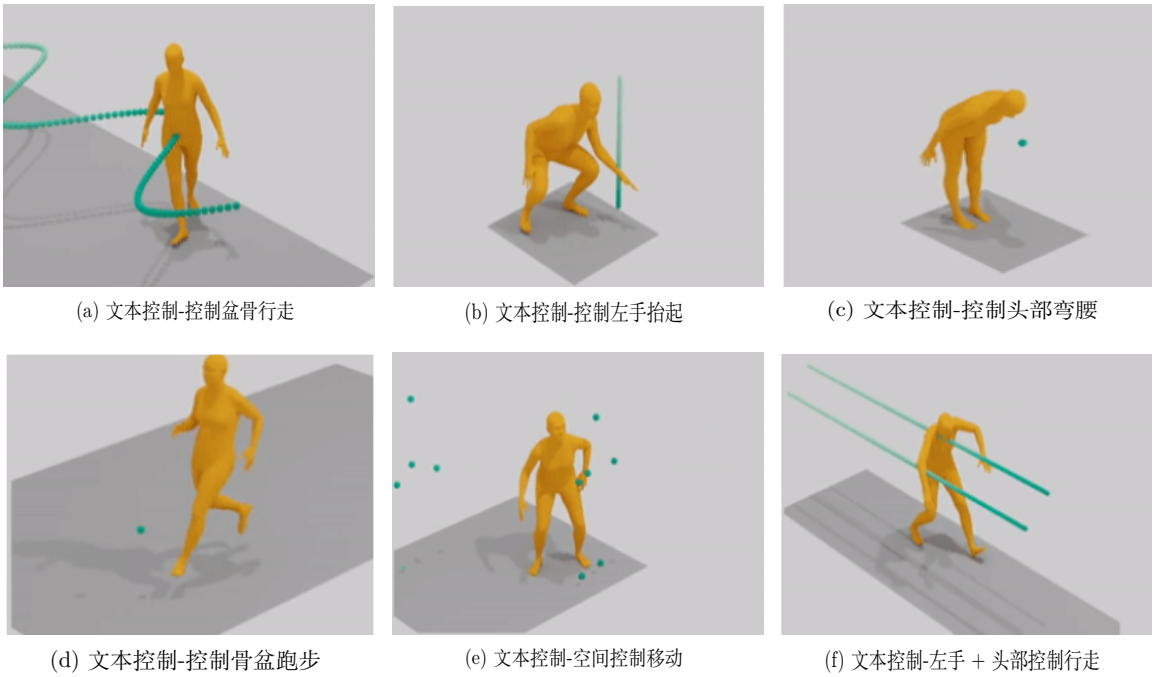
调用 HumanML3D 数据集提供的测试集信息，通过推理程序运行得到预测的含关节控制信息的人体动作序列。运动序列会以 numpy 数组的形式输出到本地文件目录中。

(3) **可视化**

3D 人体动作模型的可视化依赖 SMPL（或 SMPL-X）人物关节点框架和 Blender 等 3D 动作渲染引擎，本文复现后在 Blender 中实现对动作的可视化定性实验。

在完成可视化后，可以进一步将获取 Ground Truth 中涵盖的空间控制信息，并将每一帧控制信号对应地三维坐标点绘制到空间中，得到连续的控制信号（例如图6中的点、曲线、直线等）和 3D 人物模型的可视化结果，如图6所示。同作者在开源平台上提供的 Demo 对比，复现的 OmniControl 同样能实现较好的控制效果，定性分析可以和原论文中实现接近的效果。

在复现原模型的推理基础上，进一步在修改后的模型上，对同样的文本提示词进行动作生成，得到图7所示的结果。与图6的可视化结果对比，可以看到基于文本引导的空间控制信号同样能实现较好的控制效果，特别是针对控制信号较为复杂的情况，例如对多个关节的控制和控制序列为复杂曲线的情况，基于文本的空间控制同样具备较为准确的控制能力。



(a) 文本控制-控制盆骨行走　(b) 文本控制-控制左手抬起　(c) 文本控制-控制头部弯腰

(d) 文本控制-控制骨盆跑步　(e) 文本控制-空间控制移动　(f) 文本控制-左手 + 头部控制行走

图 7. 基于文本的空间控制的模型可视化结果

此外，在实验过程中发现，在特定帧下文本具备比空间坐标点更好的控制效果，因传入的空间序列难以在某个时刻进行改动，文本的灵活性在时序控制方面具备更好的优势。

5.2 定量评估与对比分析

除了定性实验之外，原文还给出一系列用于评估动作生成性能的指标。生成任务经典的评价指标同样适用于该任务中，例如 Frechet 起始距离 (FID) 可以反映生成运动的自然度。R-Precision 可以评估生成的动作与其文本提示的相关性，而 Diversity 则衡量所生成运动的可变性。

Xie 等人 [12] 在论文中指出，为了评估控制性能，定义了足部滑行比率（foot skating ratio）作为轨迹与人体运动和物理合理性之间不连贯性的评价指标。足部滑行比率测量任一脚关节点在保持与地面接触（脚高 < 5 厘米）的同时滑动超过一定距离（2.5 厘米）的帧的比例。此外，还定义了关键帧中受控关节位置的轨迹误差、位置误差和平均误差，以测量控制精度。轨迹误差是预测出现错误的轨迹的比率，定义为任何关键帧位置误差超过阈值（通常为 20 或 50 厘米）的轨迹。位置误差是指在阈值距离内未到达关键帧位置的比率。平均误差测量生成的运动位置与在关键帧运动步骤中测量的关键帧位置之间的平均距离。

表 3. 论文 OmniControl 在 KIT-ML 数据集下的定量评估结果

Keyframe	FID ↓	R-precision ↑	Diversity →	Foot skating ratio ↓	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err. ↓
1	0.272	0.675	9.547	0.0533	0.0000	0.0000	0.0079
2	0.251	0.683	9.498	0.0540	0.0195	0.0103	0.0253
5	0.210	0.691	9.345	0.0541	0.0645	0.0203	0.0507
49 (25%)	0.179	0.689	9.410	0.0558	0.0635	0.0108	0.0452
196 (100%)	0.181	0.698	9.311	0.0564	0.0459	0.0066	0.0398

表 4. 修改后模型在 KIT-ML 数据集下的定量评估结果

Keyframe	FID ↓	R-precision ↑	Diversity →	Foot skating ratio ↓	Traj. err. ↓ (50 cm)	Loc. err. ↓ (50 cm)	Avg. err. ↓
1	0.283	0.681	9.532	0.0588	0.0003	0.0002	0.0081
2	0.255	0.688	9.492	0.0543	0.0207	0.0108	0.0275
5	0.215	0.690	9.347	0.0543	0.0662	0.0217	0.0489
49 (25%)	0.182	0.694	9.419	0.0562	0.0618	0.0123	0.0437
196 (100%)	0.198	0.701	9.285	0.0572	0.0463	0.0074	0.0384

表 3和表 4分别记录了原始 OmniControl 模型和修改后的模型在 1、2、5、49、196 这 5 个关键帧上动作预测的指标。图 8展示了原始 OmniControl 模型和修改后的复现模型在不同指标上随着关键帧变换的变化曲线。

原文中的 FID 值随着关键帧数量的增加而逐渐降低，显示出生成运动的自然度有所提升。修改后的模型的 FID 指标在 0.283 到 0.198 之间波动，与 OmniControl 相比略有偏高，但仍然表明生成的运动在自然度上接近原文结果，表现出了较高的生成质量。

R-Precision 在衡量生成动作与文本提示的相关性方面，修改后的模型表现出色。R-Precision 值从 0.681 到 0.701 不等，表明随着关键帧的增加，生成动作与文本提示之间的相关性得到提升。这个规律与原文保持一致，证明了模型在理解文本提示和生成匹配动作方面的稳定性。对于足部滑行比率，该指标衡量生成运动的物理合理性。

原文中，足部滑行比率较低，而修改后的模型的结果在 0.0588 到 0.0572 之间波动，与原模型相比略有增加，但整体上生成的运动依然保持了较高的合理性，表明模型在真实性上表现良好。关于轨迹误差、位置误差与平均误差这些控制精度的评估指标，修改后的模型在控制精度上略高于原文，特别是在轨迹误差和位置误差上。尽管存在小幅度的增加，但整体误差依然处于较低水平，显示出模型在控制运动轨迹方面的有效性。

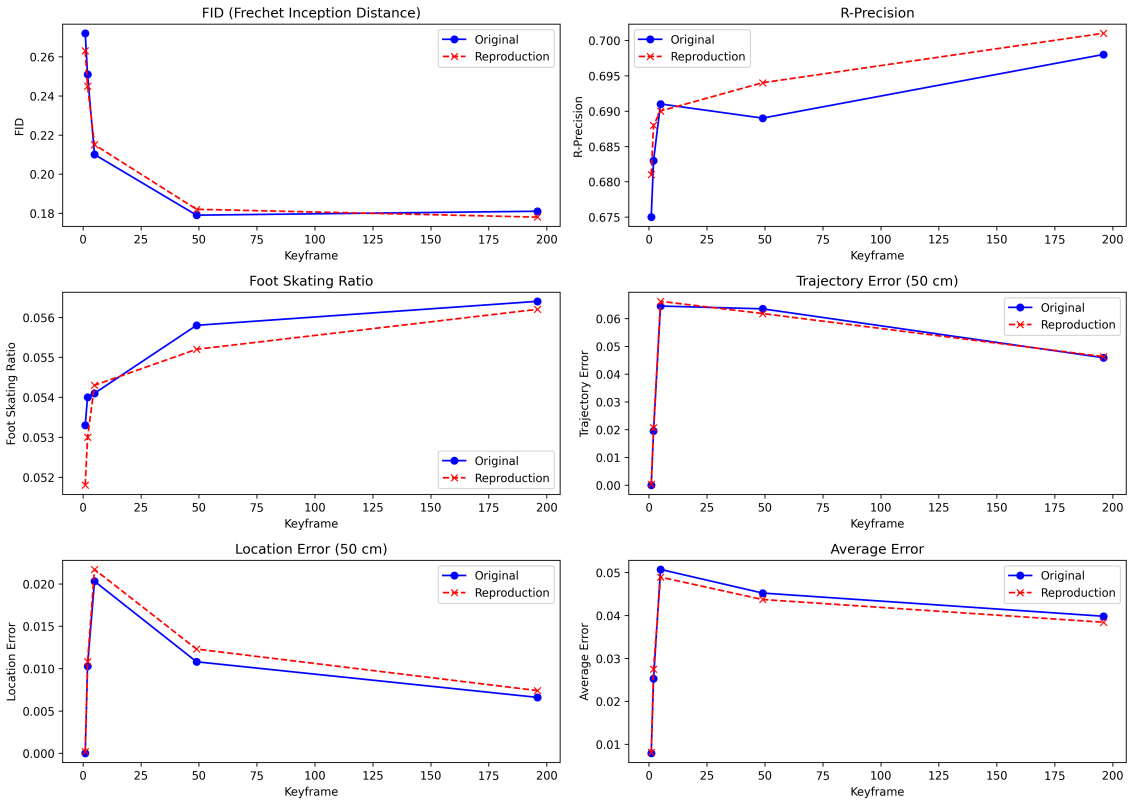


图 8. 复现修改模型与原模型定量对比统计折线图

6 总结与展望

本次工作聚焦于复现 Xie 等人 [12] 提出的 OmniControl 模型，旨在利用 MDM 基本架构和 ControlNet [14] 微调方法实现对生成的 3D 人体动作实现关节点上的精确控制。

在复现阶段，首先实现了 MDM 动作预测模型，然后将空间控制信号和真实性控制作为 ControlNet [14] 的微调信号传入模型中进行训练。在创新方面，主要进行了以下尝试：

（1）将空间控制信号替换为细粒度文本数据，通过对文本的编码实现对关节点的控制功能；

（2）调整文本理解的训练策略，通过对控制信号的合理编码实现对控制信号的识别编码，保证对动作的灵活控制和准确性

受时间影响，本次工作的实验内容仍然有限，例如在多人数据集环境下对模型的测试并未开展。同时，模型只对控制信号进行了微调，并没有对冻结的 MDM 模型进行更换，无法证明 MDM 能在控制任务中获得最佳的性能。未来将会结合理论证明，在多种数据集上进行消融对比实验，进一步优化本次工作的相关内容。

参考文献

[1] Rania Briq, Chuhang Zou, Leonid Pishchulin, Chris Broaddus, and Juergen Gall. Recurrent transformer variational autoencoders for multi-action motion synthesis, 2022.

[2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[3] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[4] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023.

[5] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023.

[6] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[9] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[10] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[11] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[12] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. *arXiv preprint arXiv:2312.10960*, 2023.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[15] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[16] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023.

[17] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *arXiv preprint arXiv:2312.15004*, 2023.