## 

## 1 项目背景

传统的动作生成方法往往依赖于预先录制的动作数据与手工设计的动作规则，虽然计算机图形学与人工智能的发展，使这些问题逐渐得到解决；但动作生成研究领域依然面临着数据获取困难，动作多样性、自然性不足等问题。虽然动作生成已然满足了电影、游戏及虚拟角色动画等虚拟现实方面的部分需求；但它在医疗康复、人机交互及机器人控制等高安全、高精度的需求场景下，依然因为动作非平滑过渡，环境适应性弱等问题而存在广泛应用的掣肘。

本小组基于现有动作生成算法与模型环境适应性弱，广泛依赖于特定数据集训练的问题展开研究，希望提出一类不需要多样性数据集训练，而天然可以获得较好的环境适应性模型。为简化模型需求，我们暂时认为该模型的动作生成仅基于平滑地面，只需要处理中位及高位空间中的几何障碍即可。这一模型的成功实现，可以为全场景应用动作生成模型减少特定环境数据集的训练依赖。

## 2 目标与任务

#### 目标：

我们提出了一种区别于现有的3D场景运动合成方法。该方法基于在标准坐标系下原点处迭代收敛的思想，使得我们提出的方案在接受运动种子序列与动作关键点标定或简单的语言命令动作指令后，可以完成3D场景中行走与各种动作之间的长距离运动过渡合成。

在3D场景运动的生成过程中，该方案以动作目标点为原点建立标椎坐标系，使用逆运动学和姿态先验的策略生成相应运动过程。且该方案使运动在平滑地面生成时不再受到3D场景中几何图形的限制，拥有更佳的场景应用收缩性。

#### 任务：

1. 学习自然语言处理（Natural Language Processing，下称“NLP”）方面的基础知识，理解其基础模型与作用机制；可以考虑学习预训练生成式转换（Generative Pre-trained Transformer，下称“GPT”）大模型对映的API接口，用于处理语言命令的解析与转换。
2. 学习transformer模型的相关知识，特别关注预训练的语言表征模型（Bidirectional Encoder Representation from Transformers，下称“BERT”）的全注意力机制。探究适合模型使用的学习率。

3. 设计并训练“WalkNet”的具体结构，使人物在对应3D场景目标坐标系下，可以基于给定动作关键点（路线），使用迭代算法与路径规划算法完成正常的连续行走动作。

4. 设计并训练“TransNet”的具体结构，使解析与转换后的运动指令可以迅速计算出目标动作的关键点姿态。并介导从行走运动状态到锚定姿态状态的平滑过渡。

5. 进行消融研究设计，探究“TransNet”是否可以被类似方案如SoTA取代；语言关键点的设置效果对运动生成序列的影响等内容。

## 3 可行性分析

#### 技术可行性：

1.在3D场景运动合成领域，过去的研究提供了丰富的基础和经验，这些研究成果可以作为我们项目的参考和借鉴。通过研究过去的文献、论文和项目，我们可以了解到不同的方法和算法在该领域的应用和效果。这些研究成果可以为我们的项目提供技术方案和实施思路，避免重复造轮子的情况，并从前人的经验中获益。

2.研究中所用到的关键技术如多人皮肤线性模型（Skinned Multi-Person Linear，下称“SMPL”）、Transformer等，都已经在学术界和工业界得到广泛应用，并且相关资源可以在网络上获得。我们还可以通过搜索和查找论文、开源代码、文档和教程等资源，深入了解这些关键技术的工作原理、实现方式和最佳实践方案。这些资源的可用性使我们能够更好地理解和应用这些技术，加速我们项目的开发和实施过程。

3.在本项目中，我们需要处理和分析大量的数据，我们可以通过公开可用数据集，数据调研，数据生成等方式获取数据，建立起一个有意义的数据集，从而用于验证和评估我们的研究成果。

#### 环境条件可行性：

1.本小组具有一定的python编程能力和项目开发经验。

2.本小组的srtp项目和3D场景运动生成相关。通过srtp的项目构建，我们已经深入研究了与3D场景运动生成相关的领域，并掌握了大量的相关资料。因此我们对现有的研究成果、技术方法和算法有着深入的了解，并具备一定的在此领域进行研究和开发的能力。

3.同时，导师实验室在3D场景运动生成领域具有相当丰富的研究成果。导师和学长们在这一领域拥有广泛的经验和专业知识，并愿意为本小组提供支持和指导。当我们在项目中遇到困难或需要进一步的指导时，可以随时向他们寻求帮助。他们的经验和指导将为我们提供宝贵的支持，帮助我们克服技术难题和实现项目目标。导师和实验室的支持是环境条件可行性分析中的重要因素，确保我们在项目中能够获得必要的指导和支持，提高项目的成功概率。

综合上述分析，本小组认为该项目可行。

## 4 初步方案及关键技术

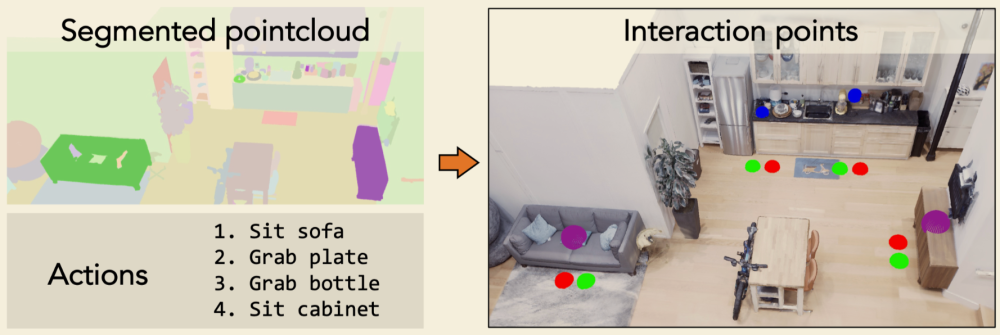
#### 初步方案：

首先，通过用户标注或使用语言模型处理用户输入或自动生成来指定 3D 场景中的交互目标锚点，作为生成动作的目标位置。同时给入一系列初始动作片段作为起始数据。接下来的部分分为两个阶段，即行走阶段与交互过渡阶段。前者我们期望使用“ WalkNet” 实现，通过路径规划算法寻找通往目标点附近的路径并给出路径中的人体关键信息（朝向等），我们计划使用将“到达下一个目标点”转化为“从某个位置向以目标位置为原点的坐标系原点逼近”的优化的 trick；后者使用 “TransNet” 实现，使用 transformer 模型，即使用放置在规范坐标系中的 AMASS 序列来训练，当人物进入目标点附近以后，进行与目标物体的交互动作生成。

4f25c9a1ecbd0500645992f2fafe144在数据表达方面，我们考虑使用 SMPL 人体模型来表达，主要包括四个参数：（**φ**，θ，**t**，β），分别为朝向、姿势、平移（位置）、形状。其中我们假设整个过程中 β 不改变，于是以三元组序列来表示动作序列：

在实际训练过程中，我们会将它表示为潜在向量 **Z** 。

而关于目标锚点的生成，第二种方法可以通过语言模型处理得到目标锚点。而第一种与第三种标定方案需要对3D 场景进行实例分割；在此之上，第三种方案还需要过先验规则生成可能的交互位置（其中包括手脚股的位置）。



从状态机的角度来分析，人物动作可以分为两种关键状态三种转换过程，状态转换图如下：

### 

#### 关键技术：

1.SMPL 模型是基于人体形状和姿势的传统模型，该模型是从大量的3D人体扫描数据集中学习而得到。该模型包括一组线性混合蒙皮权重，将底层关节和骨骼的运动映射到皮肤表面。这允许在身体移动时实现皮肤和肌肉的逼真变形。

2.Transformer 模型的主要优点是能够处理长列数据，同时避免了传统循环神经网络（RNN）中存在的梯度消失和梯度爆炸问题。Transformer 模型由编码器和解码器两部分组成，其中编码器用于将输入序列转换为一组特征向量，而解码器则用于将这些特征向量转换为输出序列。在编码器和解码器中，自注意力机制被用于计算输入序列中每个位置的重要性权重，从而实现对输入序列的有效编码和解码。

## 5 预期工作结果

1.完成自然语言处理的学习，使用相应的API将GPT成功接入对应的3D场景动作生成方案中，可以正确的解析与转换语言指令输入。

2.完成“WalkNet”与“TransNet”网络的搭建。在语言指令输入及动作关键点标定后，3D场景中人物可以先进行行走状态的动作生成，成功的绕过不同环境模型下的障碍物（不基于数据集的训练）。在靠近目标动作标定点时，会自动进行行走状态至锚定状态的过渡，并延续于指定动作的姿态完成。

3.3D场景下运动生成的过程不再依赖于对应动作或对应场景的训练集，各种动作间的连接可以适用于大多数3D场景，且生成的动作有很高的控制性与多样性。

4.相较于其他3D场景运动生成模型，该方案的应用范围更加广泛，且动作生成特别是状态过渡部分更具自然性。对于手动设置运动关键点和语言指令输入两种方法而言，它们在运动生成质量上也没有明显区别。