参赛队员姓名：李一昕

中学：北京市第一 0 一中学

省份：北京

国家/地区：中国

指导教师姓名：王威 周宇辰

指导教师单位：北京大学 北京市第一 0 一中学

论文题目：融合运动常识的自然语言动作序列自动生成

摘要：

自然语言动作序列自动生成技术广泛应用于元宇宙虚拟交互、机器人运动规划与控制、电影脚本可视化等领域。该技术面对的核心挑战是如何将语言中的概念精确映射到动作序列。现有端到端训练神经网络模型的方法在较为简单的测试环境下取得了不错的效果，但无法解决不同场景下动作多样性和准确性问题，在训练分布与实际分布区别较大时生成效果较差。本文提出了一种将常识引入神经网络的方法,通过 Prompting、引入 CLIP 特征空间和注意力机制等方法提升模型精度和泛化性。我们利用 Prompting 从自然语言处理大型预训练模型中获取关于动作的详细描述，通过不同的方法获取这些描述的表征，并将这些表征引入动作生成模型中。进一步地，我们通过大型多模态预训练模型 CLIP 的 Text Encoder 将 CLIP 的特征空间引入我们的模型。我们将所提出的方法与 TEMOS、JL2P、Seq2Seq 等现有方法进行比较实验，并且由测试人员对基于文本复杂度和动作复杂度进行分类的生成结果进行主观评估。结果显示本文所提出的方法对于训练和测试集内出现过的动作有着更好的平均表现，对于数据分布外生成、复杂动词短语描述、多关节参与的复杂动作有明显的生成效果提升，表现出较强的 Zero-Shot 生成能力。

关键词：文本动作生成，常识提取，Prompting，CLIP 特征空间，注意力机制

目录

1. 引言

通过自然语言自动生成动作序列是当前人工智能的研究热点之一，广泛应用于元宇宙虚拟交互、机器人运动规划与控制、电影脚本可视化等领域。自然语言可以显式描述或隐式蕴含不同动作的种类、速度和方向或目的地。该技术面对的核心挑战是如何将语言中的概念精确映射到动作序列。现有自然语言动作生成方法主要为通过神经网络进行端到端模型训练，在较为简单的测试环境下取得了不错的效果。但是该方法无法解决不同场景下的动作多样性问题，在训练分布与实际分布区别较大时生成效果较差。并且，该方法 Zero-Shot 生成能力较差，对于训练中未出现过的动作几乎无法生成或者生成结果错误。

针对这些问题，**本文提出了一种将常识引入神经网络的方法，通过 Prompting 和引入 CLIP 特征空间的方法去提升模型的泛化性**。利用 Prompting 从大型语言模型中获取关于动作的定义，并通过不同的方法获取这些描述的表征。将这些表征输入 Text Encoder，从而将从大型语言模型中获取的常识引入文本动作生成模型中。进一步地，我们通过大型多模态预训练模型 CLIP Text Encoder，将 CLIP 的特征空间引入模型。CLIP 原本设计目标是通过将文本和图像对应的特征分享一个特征空间建立起文本和图像间的映射[2]。由于 CLIP 训练数据的丰富性和 CLIP 特征空间良好的泛化性，我们引入 CLIP Text Encoder 的模型可以根据文本得到训练数据集以外的人体动作从而提高模型的 Zero-Shot 生成能力。

2. 相关研究工作

在基于文本的动作序列生成领域，已经有很多学者尝试将语言指令转化为智能体的动作。Takano 等人利用隐马尔可夫模型建立了人类动作和文字标签间的联系，使得人们可以通过输入文字标签生成动作，但是该工作并不能提取复杂语句中的动作信息[3, 4]。Ahn 等人通过 RNN 编解码器和注意力机制完成了任务，但是该工作只关注了人体上半身的动作[5]。Plappert 等人通过双向 RNN 网络，得到了文本到运动序列的映射[6]。Yamada 等人开创了将文字描述和运动序列嵌入同一个向量空间中的方法[7]，Ahuja、Morency 和 Ghosh 在此方法的基础上通过不同的改进得到了更好的效果。Ahuja 和 Morency 通过课程式学习的办法使得

模型达到了更好的效果，而 Ghosh 等人则将文字与运动序列都分为了上身和下身两个部分，分别进行训练从而得到了更好的效果[1, 8]。还有部分工作通过 VAE 来完成生成流程。Petrovich 等人通过训练 VAE 学习到了运动的潜在表示，通过在隐空间中进行采样和位置编码查询合成生成序列[9]。Guo 等人除了使用 VAE 还通过使用李代数去代表人的自然运动，从而使生成序列符合物理规则[10]。除此以外，许多工作将 CLIP 引入生成过程。Tevet 等人通过将隐空间与 CLIP 对齐将 CLIP 模型中丰富的语义知识引入生成过程[11]。Hong 等人利用在运动 VAE 中学到的先验，提出了一种基于 CLIP 引导的合成方法[12]。

如何在人工智能中引入常识是人工智能中非常重要的一个研究领域。虽然现在很多领域中，大规模预训练模型都有着很好地效果，但他们对常识却只有极其有限的理解[13]。目前学者们主要将知识引入在在网络的四个不同位置，训练数据[14]，网络机构与超参[15]，损失函数[16]，最终输出[17]。目前知识的表达也是多样的，主要方法有，代数公式[14]，微分方程[15]，模拟结果[16]，空间表示[18]，逻辑规则[19]，知识图谱[17]，概率关系[20]与人类反馈[21]。虽然这些方法在特定情况下能达到一定的效果，但是我们没有一个标准的框架将逻辑限制引入神经网络。虽然我们可以修改损失函数，但是对于一些逻辑表示，构造一个损失函数内的惩罚项并不直接。更详细的描述可以参考 Rueden 等人的综述[22]。

3.基于常识的文本动作生成优化

3.1 现状分析与研究动机

由给定的文本生成对应的运动序列是一个极具挑战性的任务，最早该任务仅仅关注较为简单的动作，由句子生成较为复杂的全身动作的任务由 Yamada 等人开创，他们通过 Encoder-Decoder 结构将文字描述与动作编码到同一个特征空间中。然而，**这些传统的由端到端的神经网络模型在动作描述较为复杂或者生成语句与训练集语句差距较大时生成结果很差。**

（1）复杂动作描述

图 1 展示了传统模型对“A man walks forward while playing guitar.”

的生成结果。可以看出，关于“playing guitar”的部分模型完全没有给出对应的生成结果，这可能是因为传统模型在训练时所使用的句子较为简单，并没有从句结构，或者是传统模型对于 playing guitar 这种需要同时考虑动词和宾语的多义性动词无法完成较好的生成。

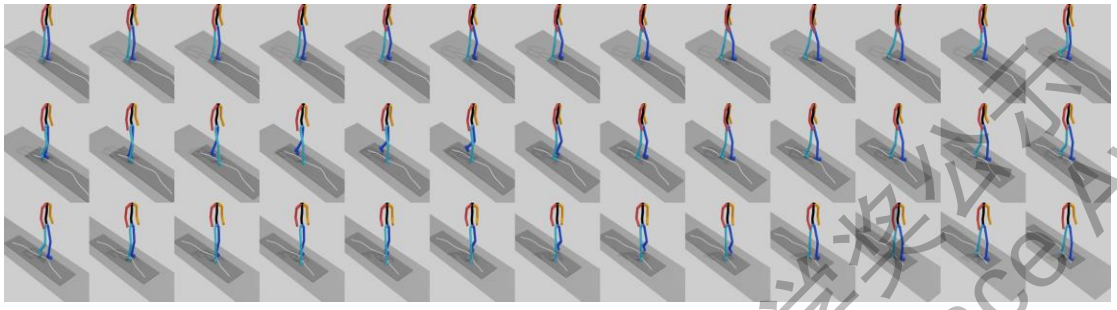

图 1. 传统模型针对输入文本“A man walks forward while playing guitar.”的动作序列生成结果，几乎没有生成关于“playing guitar”的动作序列。

因此，我们希望通过使用更加强大的语言模型来增强对动作描述的理解，同时建立各动作描述与对应的动作序列在隐空间的对齐。

（2）数据分布外（Out-of-Distribution）生成

实际测试中用到的动作描述可能由不同背景的人提供，与训练集中的句式、用词往往存在较大差异，甚至测试集中存在训练集没有涉及到的动作，这给当前的模型带来了巨大的挑战。

图 2 展示了当前模型对“A person plays handstands.”的生成结果。可以看出，该模型完全没有给出“handstands”对应的生成结果，这是因为在训练集中没有出现过“handstands”的任何相关描述，相对其他动作，“handstands”显得很小众很少见，当前模型对于这种 Few-Shot/Zero-Shot 动作生成无能为力。

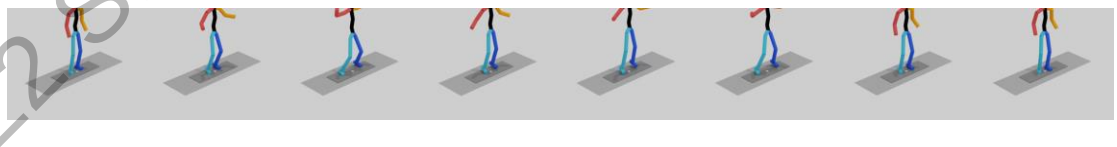

图 2. 当前模型对输入文本“A person plays handstands.”的动作序列生成结果，完全没有生成关于“handstands”的动作序列。

因此，可以考虑通过外源知识库引入运动常识，即特定动作的通用描述，譬

如“handstands”是倒立双手撑地、双脚向上等，通过增加额外信息引导模型生成对应的动作序列，使模型在面对未见过的分布以及复杂动作时能有更好的表现。

针对当前模型存在的以上两类问题，本项目提出采用 CLIP 模型来增强对文本描述的理解，大规模多模态数据预训练的 CLIP 模型也提供了一个更准确的多模态隐含表示空间，同时便于数据分布内文本描述、数据分布外文本描述与对应运动序列在该隐含空间的对齐。此外，本项目显式地通过 Prompting 技术从 GPT-NeoX 中获取关于特定动作的文本描述，作为一种外源知识或者运动常识，并进一步通过注意力机制进行增强，融合到生成模型中去，能够有效地解决 Few-Shot/Zero-Shot 生成能力。图 3 给出了一个示意图去描述我们的研究思路。

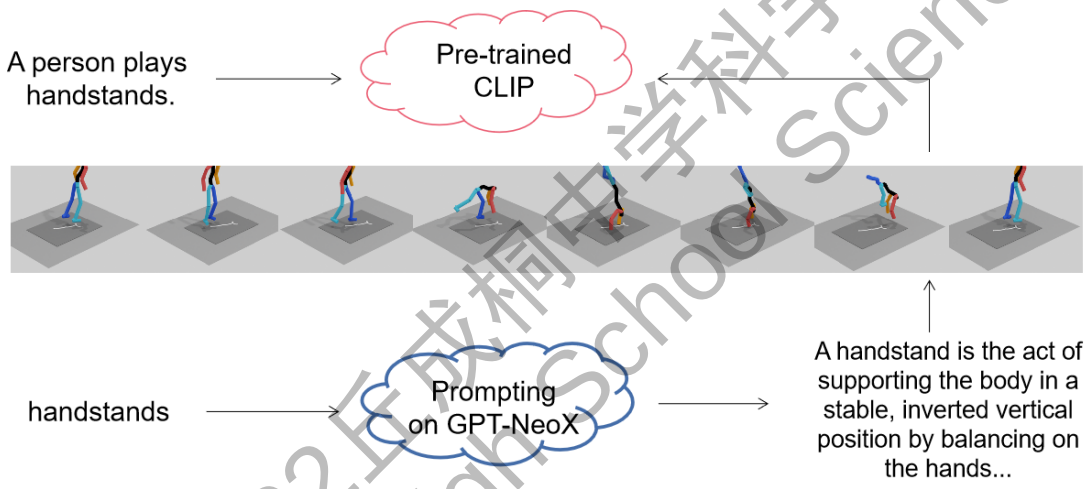

图 3. 我们的研究思路示意图：CLIP 模型提供了一个更准确的多模态隐含表示空间，实现了文本与动作的精准对齐，Prompting 技术帮忙提取更加详细的运动常识。

3.2 融合运动常识的文本动作生成

我们选取了典型的 TEMOS 方法做为文本动作生成的 Baseline 模型，通过 Prompting 引入常识进行扩展和优化，并将 CLIP 作为模型增强的方法之一[2, 23]。Prompting 是指在不修改大模型的基础上，通过针对下游任务设计合适的 Prompt 并将其输入大模型中，将大模型的输出作为任务输出的方法。比如对于一个简单的 Q and A 问题，我们可以设计一个“The answer of the question that [Sq] is”。将这个 Prompt 输入大模型（如 GPT），我们便可以得到一个对应的输出。然而，由于这些输出是语言模型直接输出，其中可能包含一些不重要的信息或重复的信息。因此，我们进一步引入了注意力机制，用于筛选其中有价

值的信息，并将这些信息用于强化 Text Encoder，从而增强模型泛化性。

3.2.1 Baseline 模型

作为 Baseline 模型的 TEMOS 使用了一种变分方法，利用可变自动编码器（VAE）训练的模型编码人体运动数据，结合文本编码器，产生与 VAE 潜在空间兼容的分布参数，从而产生多种多样的人体运动。与其他方法相比，TEMOS 考虑了本质上不明确的语言性质，并产生关于文本的多种运动描述[23]。TEMOS 模型结构如图 4 所示。

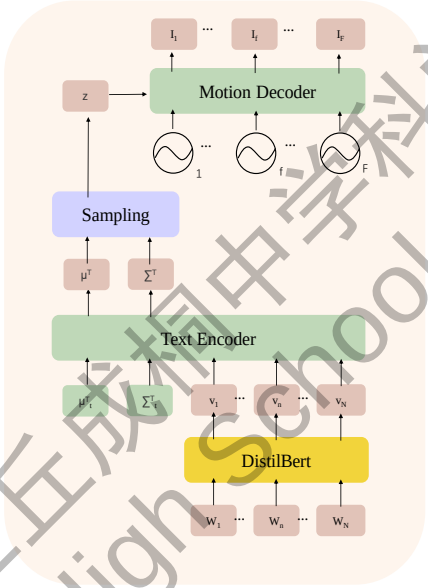

图 4. TEMOS 模型结构

3.2.2 基于 Prompting 的常识提取

Prompting 被称为是 Fine-Tune 后 NLP 领域的一个新范式。预训练模型+Fine-Tune 模式通过引入额外参数和目标函数从而对预训练模型进行微调，使其适应不同的下游任务[24]。不同于该模式，在预训练模型+Prompting 的模式里，我们通过调整下游任务，使其接近于预训练模型在训练时所遇到的任务。我们给下游任务设立合适的 Prompt，并将 Prompt 输入预训练模型，而预训练所给出的预测则是我们需要的结果。

我们通过 Prompting 的方法从 GPT-NeoX 预训练模型中获取关于运动的常识

[25]。考虑到 GPT 模型在训练时的任务为从左至右顺序生成，为了使 Prompt 与 GPT 在训练时的任务相似，我们构造如下 Prompt，“The definition of [motion] in kinesiology”。其中 motion 为表示具体动作的简单动词。我们利用 GPT-NeoX 预训练模型在之前训练过程中隐性积累的运动学常识，将 Prompt 中得 motion 扩展为包含更多细节的动作细节描述。

我们通过 Promoting 所获得的动作细节描述来增强 TEMOS 模型，多个简单动作通过 GPT-NeoX 生成的运动细节描述输入 Word2Vec 模型[26]。其输出运动细节描述表征向量通过 Average Pooling 操作输出为 1024 维向量。该向量与样本经 TEMOS 模型中 DistilBert 处理后的输出结合，用于增强 TEMOS 中的 Text Encoder。模型结构如图 5 所示。

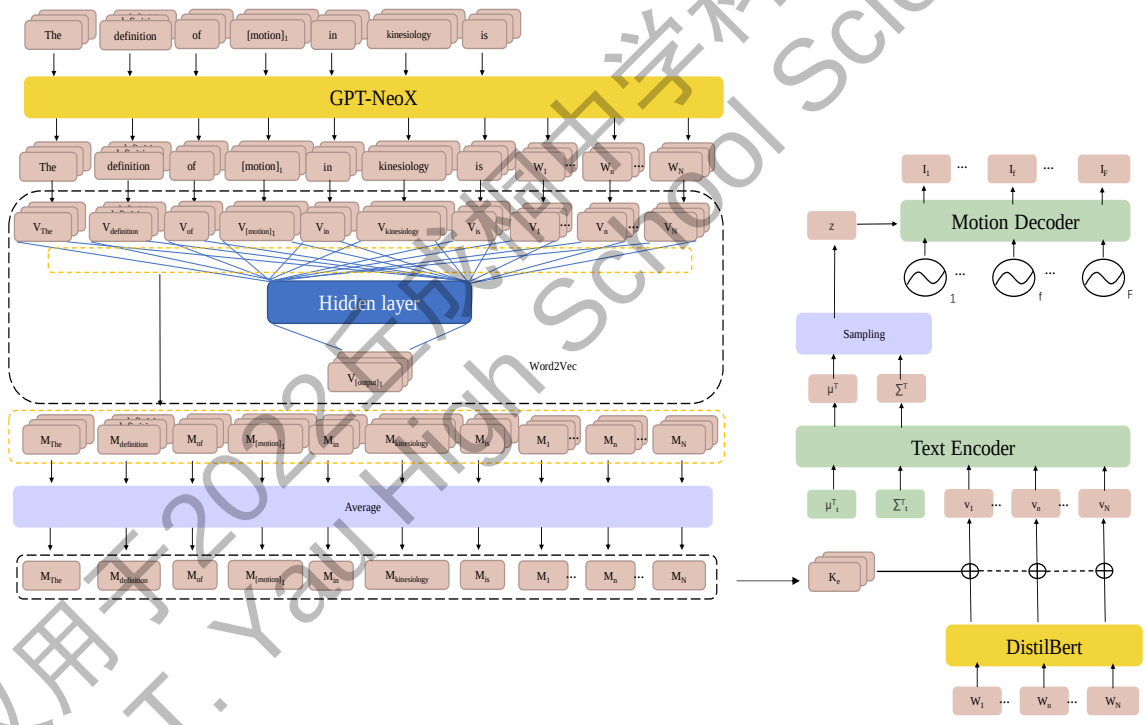

图 5. 通过 Promoting 增强 TEMOS 模型

3.2.3 基于注意力机制的常识编码

在 3.2.2 中，我们通过 Prompting 得到了一些类似这样的生成结果：

“The definition of handstand is the motion that supporting the body in a stable, inverted vertical position by balancing on the hands...”

不难注意到，其中可能会出现一些重复和一些相对来说并不重要的词语。因此，如果我们直接将这些描述的表征输入 Text Encoder，这些问题也会被带入模型中。所以我们通过注意力机制将 Prompting 所生成的结果进行筛选加权，从而获得更为有效的表征。通过将 Prompting 结果输入 Sentence-BERT（SBERT）模型来实现这一点[27]。BERT 究其根本是一个带有注意力机制的多层双向的 Transformer 编码器[28]。如图 6 所示，我们通过 Sentence-BERT 来计算我们所输入 Prompting 的向量表示，通过 BERT 的注意力机制，我们将 Prompting 所生成的结果进行了加权计算，加强了其中对我们后续生成有用的常识相关信息。我们将通过注意力机制筛选加权后的表征加入网络，从而提升网络的表现。

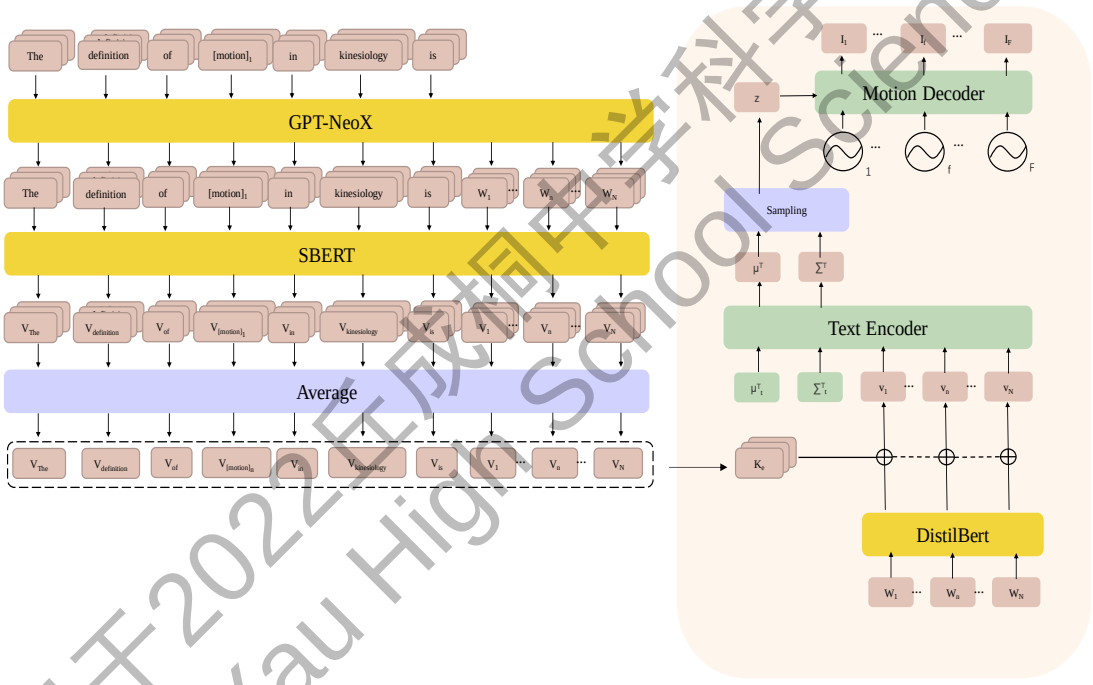

图 6. 通过注意力机制进一步增强 TEMOS 模型

3.2.4 基于 CLIP Text Encoder 的文本编码增强

CLIP 作为一个成功的多模态预训练模型，将文本和图像两个不同模态的特征映射到同一个特征空间，从而建立了两模态之间的映射关系[2]。CLIP 的大量训练使其积累了大量隐藏的多模态常识以及运动学尝试。并且 CLIP 具有丰富的文本表达以及泛化性良好的特征空间良好。因此，我们将通过替换上述模型中的 Text Encoder 的方法，将 CLIP 中的大量常识引入到我们的网络中。采用 CLIP Text encoder 的 Baseline 模型结构如图 7 所示。

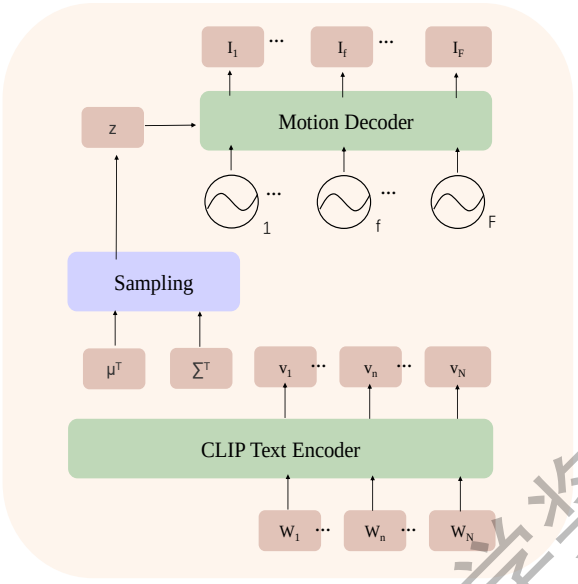

图 7. 采用 CLIP Text Encoder 的 Baseline 模型结构

在 3.2.2 和 3.2.3 中，我们论证了使用 Prompting 和注意力机制加强模型的有效性和合理性。因此，我们将采用上述方法对采用 CLIP Text Encoder 的 baseline 模型进行增强。类似的将多个简单动作或动词短语通过 GPT-NeoX 生成的运动细节描述输入 SBERT 模型。将 SBERT 加权筛选后的运动学表征与 CLIP text Encoder 的表征结合，从而增强 CLIP text Encoder。其模型结构如图 8

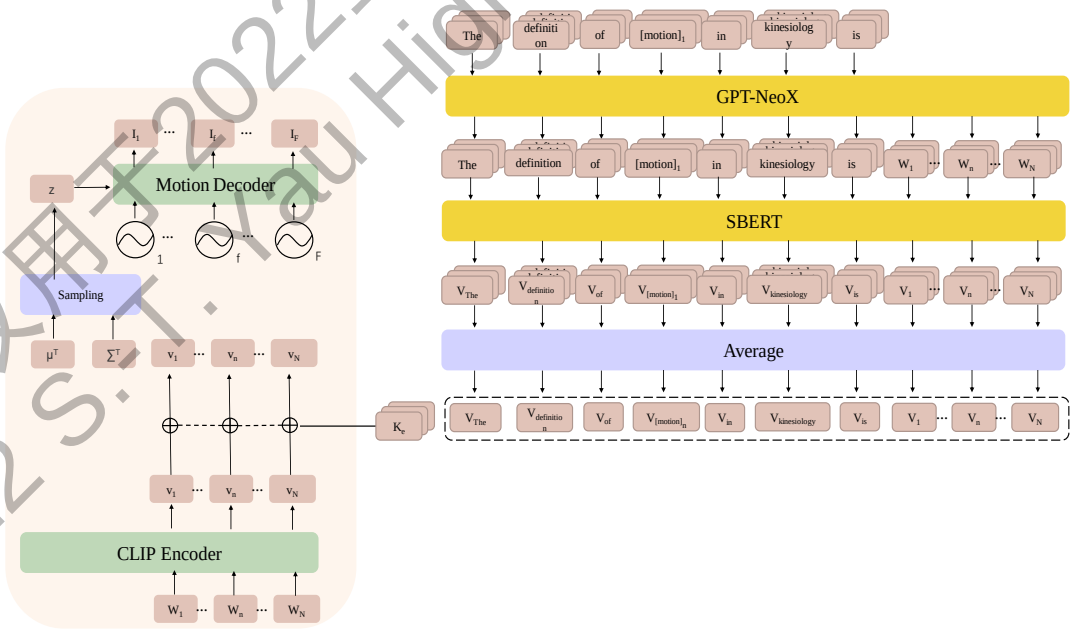

图 8. 通过 Promoting、注意力机制和 CLIP Text Encoder 增强的模型

所示。由于 CLIP text Encoder 的特征空间具有良好的泛化性，且在大量数据上进行过训练，我们认为通过引入 CLIP text Encoder 并对其采用 Prompting 和注意力机制的方法去进行加强可以提升模型的 Zero-Shot 能力。

4.实验结果与分析

4.1 实验数据集

4.1.1 KIT

KIT 是一个大型可扩展运动语言数据集。其中包含了大量语言描述以及其对应的运动序列。该数据一共包含 3911 个总时长 11.23 个小时的运动序列以及 6353 个总计 50000 多词的自然语言注释。其中数据来源于多个运动捕捉数据库，并使用了一套独立于捕捉系统和标记集的统一表示法将这些数据重新标注[29]。本文模型的训练与评估在该数据集上完成。

4.1.2 BABEL

BABEL 是一个大型数据集，其中的语言标签描述了 MoCap 序列中正在进行的动作。BABEL 包括来自 AMASS 的约 43 小时的 mocap 序列的动作标签。动作标签有两个抽象层次--序列标签描述序列中的整体动作，而帧标签描述序列中每一帧的所有动作。每一帧标签都与 mocap 序列中相应动作的持续时间精确对齐，而且多个动作可以重叠。在 BABEL 中，有超过 28000 个序列标签和 63000 个帧标签，它们属于 250 多个独特的动作类别[30]。在本文中，我们采用了 BABEL 中所提供的动作标签将其用作 Prompting 所使用的[motion]。我们将这 250 多个动作输入 GPT-NeoX 模型，将其生成的知识通过处理后用来强化基础模型的 Encoder。

4.2 评估方法

我们从定量和定性两个角度分别对实验结果进行评估。

（1）定量评估指标

首先，从定量的角度我们使用的 Average Positional Error (APE) and

Average Variance Error (AVE)[1, 23]作为指标进行评估，其定义如下：

关节 j 的 APE 是关于帧（F）和测试样本（N）的生成动作和基本事实之间关节位置的 L2 距离平均值。

$$\mathrm{APE}[j] = \frac{1}{NF}\sum_{n\in N}\sum_{f\in F}\left\|H_f[j] - \hat{H}_f[j]\right\|_2$$

关节 j 的 AVE 是关于帧（F）和测试样本（N）的生成动作和基本事实之间方差的 L2 距离平均值。

$$\mathrm{AVE}[j] = \frac{1}{N}\sum_{n\in N}\|\sigma[j] - \hat{\sigma}[j]\|_2$$

其中：

$$\sigma[j] = \frac{1}{F-1}\sum_{f\in F}\left(H_f[j] - \tilde{H}_f[j]\right)^2 \in \mathbb{R}^3$$

在实验结果报告中我们使用了如下四个关键指标：

APE_root：基于足关节三维坐标的足关节位置误差

APE mean pose：姿态位置平均误差

AVE_root：基于足关节三维坐标的足关节方差

AVE pose：姿态均方差

（2）定性评估方法

我们通过测试人员观察，进一步评估生成动作序列的语义差异。通过两个维度的划分，将动作分类成如图 9 所示的 4 种类型，并根据不同动作类型评估算法的有效性。

- 首先，在动作词语语义维度将动作分为多意和单意。例如动作 throw 在涉及宾语的不同语境下会产生动作序列差异，如 throw a bunch（出拳）和 throw the javelin (投标枪)的整体动作序列差异；

- 其次，在动作所涉及的关节维度将动作分为部分肢体简单动作和全身性复杂动作。如动作 arm movement 只涉及上肢的部分肢体简单动作，而动作 walk 为涉及上下肢和躯干等关节的全身性复杂动作。

从时序角度，我们也通过测试人员观察动作序列所包含的分段语义映射，以克服 APE 和 AVE 中以对应帧为基础评估的局限性。

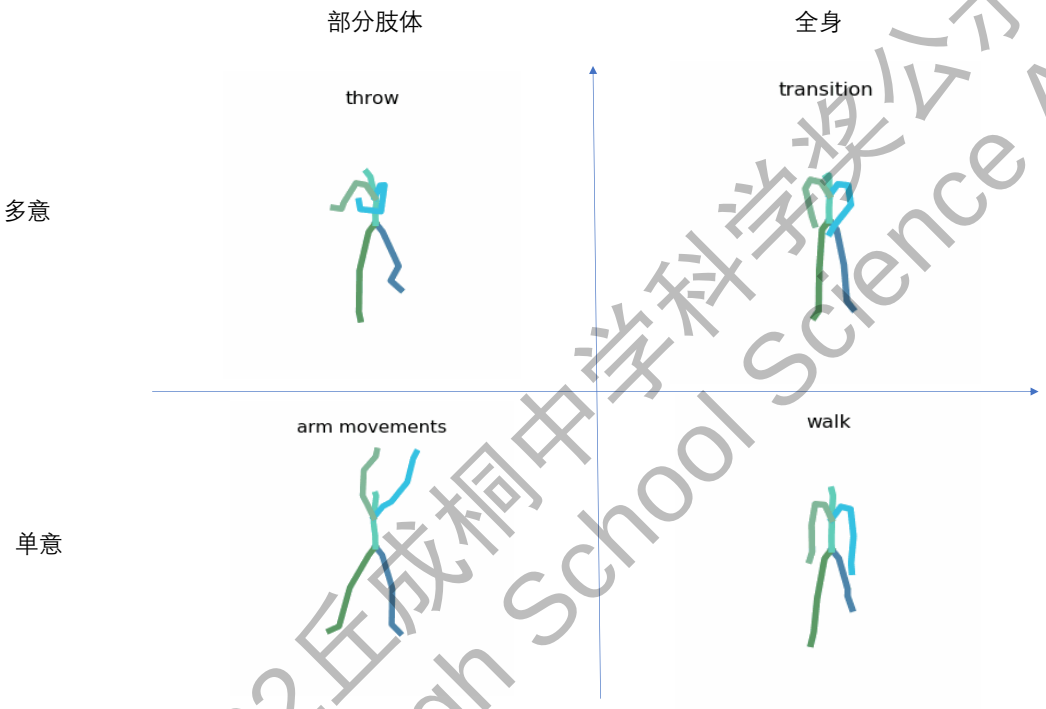

图 9. 基于语义和复杂度的动作分类

4.3 比较、消融和主观评估实验

4.3.1 对比方法 1 - Seq2Seq

Seq2Seq 实验采用最基本的 Many-to-Many Seq2Seq 模型。该模型采用了 Encoder 与 Decoder 结构。对于 Encoder 和 Decoder，我们都采用了 LSTM 模型。Encoder 任务为读取输入的文本序列并获得隐藏状态向量，而 Decoder 则将获取 Encoder 所得到的隐藏状态向量进行解码，成为解码网络的第一个单元。其结构如图 10 所示。

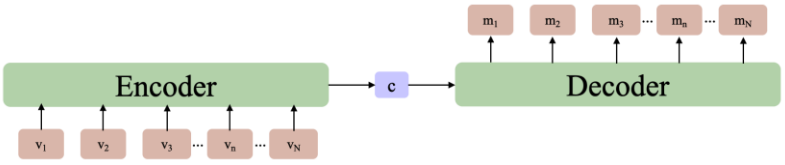

图 10 Seq2Seq 模型结构

4.3.2 对比方法 2 - Joint Language-to-Pose (JL2P)

Joint Language-to-Pose(JL2P)通过基于课程学习(curriculum learning)的方法端到端地学习文本和人体姿态的联合嵌入。文本和姿态通过 Pose Encoder 和 Language Encoder 被映射到一个联合嵌入空间。该嵌入空间可以被用于训练生成姿态序列的 Pose Decoder。在训练过程中，Pose Encoder 和 Language Encoder 都会被使用一建立联合嵌入空间。在动作姿态序列生成过程中，只使用 Language Encoder 和 Pose Decoder 为自由模式的动作描述生成相应的姿态序列。[1] 原文中的模型结构如图 11 所示。

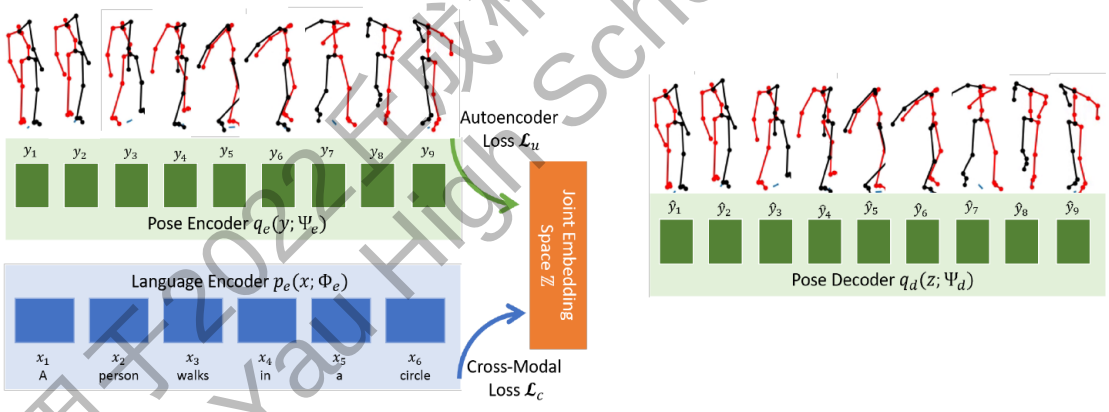

图 11. Joint Language-to-Pose (JL2P)模型结构。

4.3.3 消融实验

我们分别基于基本 TEMOS 和 CLIP 增强模型进行了两组消融实验，分别去除了增强模型中的 Prompting 和注意力机制，保持模型其余部分、数据预处理方法、训练超参数等均保持不变，以验证上述机制的有效性。

4.3.4 主观评估实验

我们通过测试人员观察，进一步评估生成动作序列的语义差异。通过两个维

度的划分，将动作分类成如图 10 所示的 4 种类型，并通过数据分布外生成实验评估算法有效应和 Zero-Shot。

4.4 实验结果

4.4.1 对比和消融实验结果

除了本文所设计的各种模型，我们还使用了其他研究人员的相关算法进行实现，包括 JL2P 和 Seq2Seq。各种算法实验的 APE 和 AVE 指标如表 1 所示。

表 1. 对比和消融实验结果

评估指标	Baseline	Prompting	Prompting + Attention	CLIP	CLIP+ Prompting	JL2P	Seq2Seq
APE_root	1.0387	1.0531	1.0271	1.0581	**1.0143**	1.1976	1.3689
APE_mean_pose	**0.1002**	0.1012	0.1007	0.1037	0.1041	0.1236	0.2403
AVE_root	0.5571	0.4677	**0.4597**	0.4782	0.4603	0.6313	0.6238
AVE_mean_pose	0.0061	0.0059	**0.0057**	0.0058	**0.0057**	0.0081	0.0121

在表 1 中我们可以看到，本文提出的 CLIP+Prompting+Attention 方法在和 AVE_root 和 AVE_mean_pose 两项指标中取得了最好的效果，本文提出的 Baseline+Prompting+Attention 方法在 AVE_root 指标方面取得最好效果，在 AVE_mean_pose 指标方面取得并列最好效果。这说明我们的方法能够切实有效地通过常识引入提高动作生成的精确度。

同时我们也可以看到，去除 Prompting 和注意力机制的情况会对 4 个指标都产生负面影响。这说明我们添加的这些结构对提升精度有着关键的作用。

4.4.2 主观评估结果

对于训练集和测试中存在且只涉及部分肢体的简单动作，常识增强模型和 Baseline 模型生成的动作序列略有差异，但都可以体现基本语义。如图 12 所示，输入文本为“A person jumps with right foot”。Baseline 模型生成的动作序列为原地起跳，运动轨迹范围较小，上肢关节运动不明显。常识增强模型生成的动作序列为向前跳跃，相应运动轨迹接近直线运动，上肢关节和躯干有明显运动。

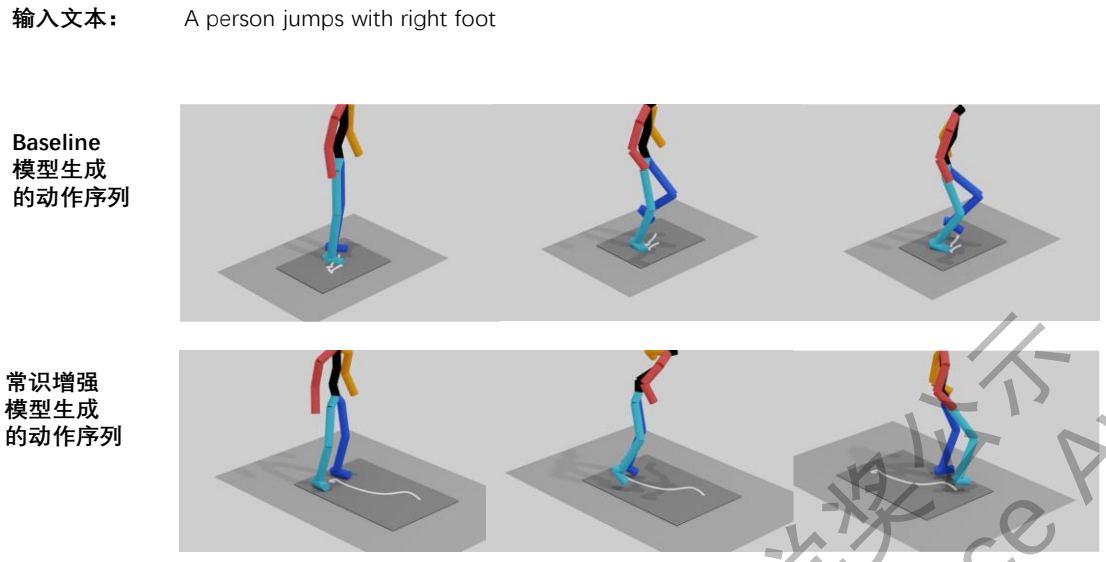

图 12.常识增强模型和 Baseline 模型未输入文本“A person jumps with left foot”生成的动作序列对比

对于由多个动词描述的以及复合全身性动作，常识增强模型产生了更好的效果，能够体现文本输入的完整语义。而 Baseline 模型则只产生部分动作而造成动作序列语义缺失和语义理解错误。如图 13 所示，当输入为“A person walks while playing handstand”时，Baseline 模型只生成了倒立的动作序列，而没有生成行走相关动作，造成动作序列中“walk”语义缺失。而常识增强模型则将倒立和行走动作复合在同一动作，较好的体现了输入文本中的包含“walk”和“handstand”的完整关键语义信息。

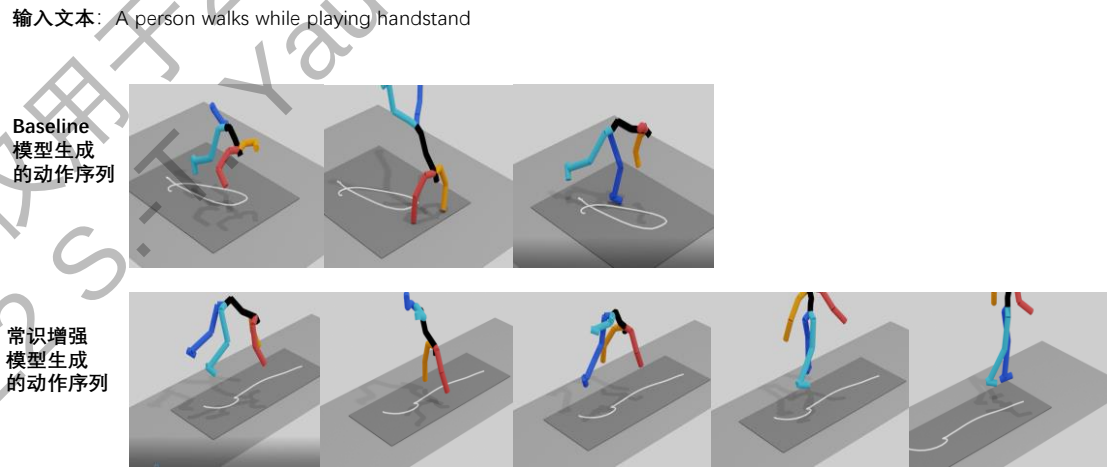

图 13.常识增强模型和 Baseline 模型为输入文本“A person walks while playing handstand”生成的动作序列对比

对于训练和测试集中未出现的数据分布外生成、由多个动词描述的、复合全身性动作，常识增强模型产生了更好的效果，能够体现文本输入正确语义，体现了较强的 Zero-Shot 生成能力。而 Baseline 模型则只产生部分动作而造成语义理解错误，Zero-Shot 能力较弱。如图 14 所示，当输入为“A person squats on the ground and then does a forward somersault”时，Baseline 模型只生成了行走的动作序列，而没有蹲前和滚翻的相关动作，造成文本中“squats”和“somersault”语义理解错误。而常识增强模型则蹲和前滚翻动作复合在同一动作，较好的体现了输入文本中的包含“squats”和“somersault”语义的正确语义信息。

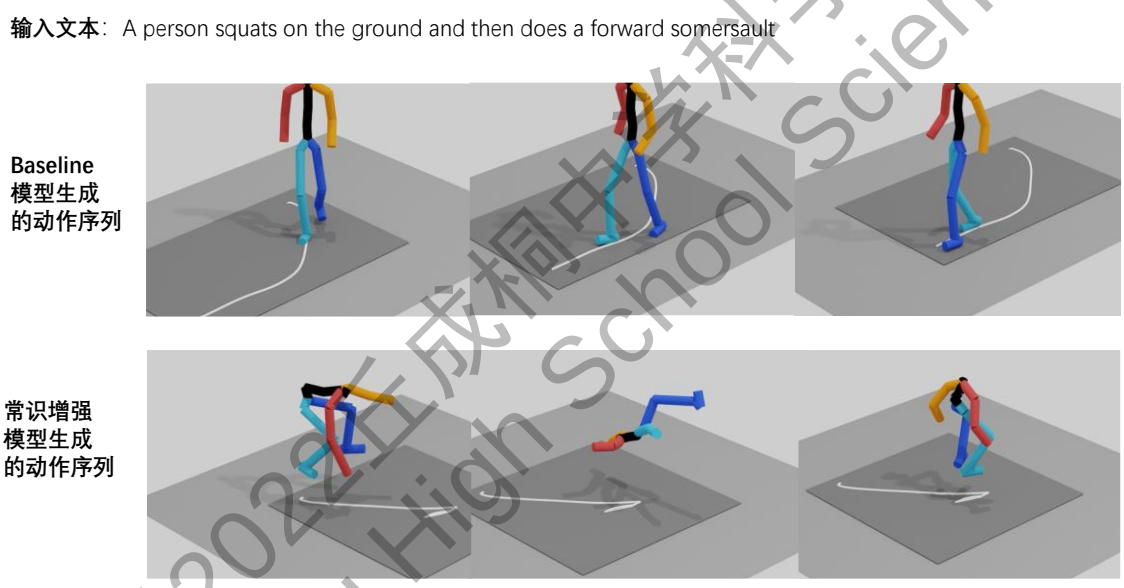

图 14. 常识增强模型和 Baseline 模型为输入文本“A person squats on the ground and then does a forward somersault”生成的动作序列对比

5. 结论

本文提出了一种将常识引入自然语言动作序列自动生成的方法，通过 Prompting、引入 CLIP 特征空间以及注意力机制等方法提升模型精度和泛化性。我们利用 Prompting 从自然语言处理大型预训练模型中获取关于简单动作的详细描述，通过不同的方法获取这些描述的表征，并将这些表征引入动作生成模型中。进一步地，我们通过大型多模态预训练模型 CLIP Text Encoder 的方法，将 CLIP 的特征空间引入我们模型。我们将所提出的方法与 TEMOS、JL2P、Seq2Seq

等方法进行比较实验，并且由测试人员对基于文本复杂度和动作复杂度进行分类的生成结果进行主观评估。结果显示本文所提出的方法对于训练和测试集内出现过的简单动作有着更好的平均表现，对数据分布外生成、多个动词短语描述、多关节参与的复杂动作有明显的生成效果提升，体现出较强的 Zero-Shot 生成能力。本文所提出的方法仍然由进一步提升的空间，我们会进一步通过其他技术增强常识提取和动作生成机制，争取在生成动作精度和泛化性能方面进一步提升。

参考资料

1. Ahuja, C. and L.-P. Morency. *Language2pose: Natural language grounded pose forecasting*. in *2019 International Conference on 3D Vision (3DV)*. 2019. IEEE.
2. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International Conference on Machine Learning*. 2021. PMLR.
3. Takano, W., D. Kulic, and Y. Nakamura. *Interactive topology formation of linguistic space and motion space*. in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2007. IEEE.
4. Takano, W. and Y. Nakamura, *Symbolically structured database for human whole body motions based on association between motion symbols and motion words.* Robotics and Autonomous Systems, 2015. **66**: p. 75-85.
5. Ahn, H., et al. *Text2action: Generative adversarial synthesis from language to action*. in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018. IEEE.
6. Plappert, M., C. Mandery, and T. Asfour, *Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks.* Robotics and Autonomous Systems, 2018. **109**: p. 13-26.
7. Yamada, T., H. Matsunaga, and T. Ogata, *Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions.* IEEE Robotics and Automation Letters, 2018. **3**(4): p. 3441-3448.
8. Ghosh, A., et al. *Synthesis of compositional animations from textual descriptions*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
9. Petrovich, M., M.J. Black, and G. Varol. *Action-conditioned 3d human motion synthesis with transformer vae*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
10. Guo, C., et al. *Action2motion: Conditioned generation of 3d human motions*. in *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
11. Tevet, G., et al., *MotionCLIP: Exposing Human Motion Generation to CLIP Space.* arXiv preprint arXiv:2203.08063, 2022.
12. Hong, F., et al., *AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars.* arXiv preprint arXiv:2205.08535, 2022.
13. Talmor, A., et al., *Commonsenseqa 2.0: Exposing the limits of ai through gamification.* arXiv preprint arXiv:2201.05320, 2022.
14. Ladický, L.u., et al., *Data-driven fluid simulations using regression forests.* ACM Transactions on Graphics (TOG), 2015. **34**(6): p. 1-9.
15. De Bézenac, E., A. Pajot, and P. Gallinari, *Deep learning for physical processes: Incorporating prior scientific knowledge.* Journal of Statistical Mechanics: Theory and Experiment, 2019. **2019**(12): p. 124009.
16. Du, Y., et al., *Learning to exploit stability for 3d scene parsing.* Advances in

Neural Information Processing Systems, 2018. **31**.
17. Mrkšić, N., et al., *Counter-fitting word vectors to linguistic constraints.* arXiv preprint arXiv:1603.00892, 2016.
18. Worrall, D.E., et al. *Harmonic networks: Deep translation and rotation equivariance*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
19. Hu, Z., et al., *Harnessing deep neural networks with logic rules.* arXiv preprint arXiv:1603.06318, 2016.
20. Yet, B., et al., *Combining data and meta-analysis to build Bayesian networks for clinical decision support.* Journal of biomedical informatics, 2014. **52**: p. 373-385.
21. Kaplan, R., C. Sauer, and A. Sosa, *Beating atari with natural language guided reinforcement learning.* arXiv preprint arXiv:1704.05539, 2017.
22. Von Rueden, L., et al., *Informed Machine Learning--A Taxonomy and Survey of Integrating Knowledge into Learning Systems.* arXiv preprint arXiv:1903.12394, 2019.
23. Petrovich, M., M.J. Black, and G. Varol, *TEMOS: Generating diverse human motions from textual descriptions.* arXiv preprint arXiv:2204.14109, 2022.
24. Liu, P., et al., *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.* arXiv preprint arXiv:2107.13586, 2021.
25. Black, S., et al., *Gpt-neox-20b: An open-source autoregressive language model.* arXiv preprint arXiv:2204.06745, 2022.
26. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.
27. Reimers, N. and I.S.-B. Gurevych, *Sentence Embeddings using Siamese BERT-Networks. arXiv 2019.* arXiv preprint arXiv:1908.10084, 1908.
28. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
29. Plappert, M., C. Mandery, and T. Asfour, *The KIT motion-language dataset.* Big data, 2016. **4**(4): p. 236-252.
30. Punnakkal, A.R., et al. *BABEL: Bodies, action and behavior with english labels*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

致谢

本项目得到北京通用人工智能研究院王威老师、北京大学王亦洲和鲁东祐老师、北京市第一 0 一中学周宇辰老师的悉心指导和帮助，在此表示衷心的感谢！

1. 论文选题来源、研究背景

论文的选题主要来源于对本人兴趣和当前人工智能研究热点的结合，是本人自然语言处理领域长期学习积累在计算机视觉领域的进一地步深入和扩展。

2. 每一个队员在论文撰写中承担的工作以及贡献

本项目为本人在导师指导下独立完成。

3. 指导老师与学生的关系，在论文写作过程中所起的作用

王威老师是北京市第一 0 一中学北大前沿计算研究中心 AI 实验室校外指导老师。周宇辰老师是北京市第一 0 一中学北大前沿计算研究中心 AI 实验室负责人。本人为该实验室成员。

4. 他人协助完成的研究成果

本文工作均为本人在导师指导下独立完成。

团队成员介绍：

北京市第一 0 一中学高三国际英才班。课内成绩优异，对人工智能和数学有着浓厚的兴趣。自学多门大学数学课程，如数学分析、组合数学等。曾为本年级同学讲授 AI 选修课中的 NLP 部分。

曾参与和完成多个人工智能相关项目，如基于 GPT2 与知识图谱的中文可控生成模型、基于 GPT3 的机器翻译预训练模型调优、Diagnose Parkinson's Disease: Utilizing 1D-inception Network to Analyze Typing Data、Predict the water level of the Lake Mead for the next 30 years based on ARIMA、Characterizing Spectral Properties of Bridge Graphs 等。

在信息学和数学竞赛中成绩优异，如美国计算机奥赛公开赛满分并列第一(2021.4)、美国计算机奥赛二月赛满分并列第一 (2021.2)、美国高中数学邀请

赛全美第 7 名（2021.3）、美国数学奥赛国家集训队分数入围 (2021.7)、全美数学竞赛 10 Top 1% (2021.2)、全美数学竞赛 12 Top 1% (2021.2)等。

指导教师介绍：

王威，博士，北京通用人工智能研究院研究员，中国图像图形学会类脑视觉专委会秘书长。深入系统地开展视觉认知计算研究，取得了多项原创性进展。相关工作发表领域国际期刊和国际会议论文 48 篇，获得 1 个最佳论文奖（Yann LeCun 组织的 CVPR DeepVision Workshop2014）和 1 个最佳学生论文奖（ICPR2014），研究工作被南加州大学、加州理工学院、微软亚洲研究院等国际知名机构学者的引用。研究成果近 5 年 Google Scholar 累计引用 5000 多次，单篇引用最高 1700 多次，相关关键技术获 14 项专利授权。

周宇辰，博士，研究员，北京市第一 0 一中学英才学院人工智能导师。ACM 和 IEEE 高级会员，曾任中国计算机学会嵌入式系统专委会委员。IBM 二十年技术创新经验，曾任 IBM 中国研究院人工智能感知研究主管、IBM 科学院成员、IBM 发明大师等，3 次获杰出技术成就奖。学术专著 1 部，国际标准 2 项，国际专利近 50 项，学术论文 30 余篇。