

轻量级健身 Transformer: 一种用于远程监控身体训练的语言视觉模型

Alexander Postlmayr^a, Pamela Cosman^a, Sujit Dey^a

^a*Dept. of Electrical and Computer Engineering University of California, San Diego,*

Abstract

我们引入了一种健身跟踪系统，只使用 RGB 智能手机摄像头即可进行远程监控，使得健身跟踪更加私密、可扩展且具成本效益。虽然之前的工作探索了自动化的锻炼监督，但现有模型在锻炼种类上要么过于有限，要么过于复杂而难以在现实世界中部署。先前的方法通常专注于一些小范围的锻炼，而无法在不同的动作中实现泛化。相比之下，我们开发了一个强大且多任务的运动分析模型，能够在数百种锻炼中执行锻炼检测和次数计数，这个规模远超以往的方法。我们通过组建一个大规模的健身数据集 奥林匹亚，涵盖 1900 多种锻炼，克服了以前的数据限制。据我们所知，我们的视觉-语言模型是第一个可以在骨骼健身数据上执行多任务的模型。在 奥林匹亚 上，我们的模型仅使用 RGB 视频即可以 76.5% 的准确度检测锻炼，并以 85.3% 的离一误差计数重复次数。通过为锻炼识别和计数呈现单一的视觉-语言变压器模型，我们在普及由人工智能驱动的健身跟踪方面迈出了一大步。

1. 介绍

虽然肌肉骨骼疾病是美国医疗成本的主要驱动因素之一 [1]，而体育锻炼 [2, 3] 是管理这些病症的行之有效的干预措施，但正确的技术对于有效的康复和预防损伤至关重要。然而，专业健身训练的获取始终昂贵，通常需要面对面的监督，限制了其可及性。研究表明，力量训练可以减少炎症 [4]，甚至可以减少某些癌症的风险 [5]。此外，肌肉质量与死亡风险呈负相关 [6]，这突显了保持身体健康的重要性。随着现代生活方式越来越久坐不动，将结构化的锻炼重新融入日常生活至关重要。我们的研究旨在通过人工智能驱动的可扩展解决方案来民主化个性化健身训练的获取，从而加速这一转变。借助 AI 支持的远程监控进行身体训练，可以让人们独自锻炼，同时享有来自专业健身教练监督的部分好处。通过计算机，用户可以看到锻炼动作，以便他们可以随时随地锻炼。虚拟体能训练需要提供个性化反馈 [7]

，因为当前的在线程序不够完善，尤其是在跟踪用户的锻炼动作、重复次数和正确姿势方面 [8]。通过跟踪提高责任感可以提升动机 [9, 10]，设定目标、提醒和进度跟踪的健身计划更为成功 [11]。目前，运动反馈的技术方法显示了多样化的方法。许多现有工作 [12, 13, 14, 15, 16] 依赖于昂贵的专用硬件，例如 RGB-D 相机，限制了它们的使用。纯软件解决方案提供了一种可扩展的替代方案，使用标准 RGB 相机（如智能手机和笔记本电脑中的摄像头）的单目视频。在这里，运动分析要么直接在视频像素上进行，要么在使用 3D 人体姿态估计（3D HPE）模型从视频中提取的骨架关键点上进行。仅使用视频的方法已显示出足够的结果 [17, 18, 19, 20, 21, 22]，然而，它们对于移动设备而言效率低下。将视频流传输到云端会极大地降低隐私性，消耗大量带宽，并产生网络延迟。为了解决这些缺点，近期在 3D HPE 模型 [23, 24] 方面的研究表明可以在移动设备上实时从视频中提取人体姿态。关键的是，这种方法提供了受物理约束的人体关节角度，这是运动分析中的主要指标。无关的背景信息会被舍弃。最后，使用 3D HPE 可以通过允许高效存储运动信息，从而实现大规模保护隐私的深度学习方法。

在本文中，我们提出了一个用于多任务动作分析的深度学习方法。LiFT 是一个轻量级健身 Transformer，能够对多种锻炼进行检测和重复计数。模型概述如图 1 所示。我们的方法适合移动设备，并可用于家庭个人健身追踪。与现有解决方案不同，LiFT 在多个包含实际应用中多种锻炼的健身数据集上进行训练。表 1 显示了现有健身数据集以及包含的科目和练习数量。我们的训练数据集 Olympia 在章节 3.5 中进行了描述。我们的贡献如下：

- LiFT- the first 3D HPE skeleton based language-vision model capable of performing exercise detection and repetition counting across hundreds of exercises.
- We demonstrate that our approach effectively captures fine-grained relationships between motion segments and textual descriptions, unlike methods that rely on global feature alignment. Through evaluations on four publicly available datasets and few-shot fine-tuning experiments, we show that our model generalizes to previously unseen exercise classes, addressing a key challenge in real-world fitness applications.
- We assemble the *Olympia* dataset, which contains 7,618 annotated videos for exercise detection and repetition counting.

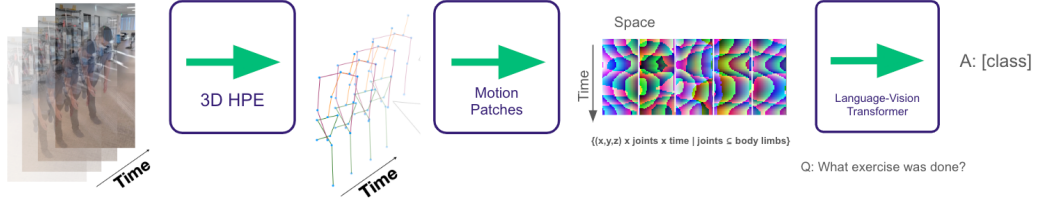


Figure 1: LiFT 的概述，其中 RGB 视频通过 3D HPE 转换为骨架关键点。骨架运动被转换为运动片段，随后与一个问题任务一起输入到语言-视觉模型中。模型的输出是一系列可能属于该运动的类别。

2. 相关工作

如下文将详细讨论，现有的健身分析模型往往尝试处理多项任务，但通常局限于少于 30 种运动，并依赖于人工设计的特征或模板化方法，从而限制了它们泛化到多样化运动的能力。此外，大多数动作-语言模型构建于在广泛、通用的动作数据集上训练的复杂架构，使得它们难以适应像运动追踪这样的专门应用。相比之下，我们的动作-语言方法专门为运动设计，提供了一个更有效、可扩展且现实世界准备的解决方案，同时避免了过度的模型复杂性。

我们将评估健身锻炼所需的任务定义为：1) 锻炼检测（分类），2) 重复动作计数，以及 3) 锻炼重复正确性评估。

Table 1: 比较现有的数据集，以检测运动和计数重复次数。

Dataset Name	Exercises
PHYTMO [25]	6
NOL-18 [26]	18
KIMORE [27]	5
UI-PRMD [28]	10
Fit3D [29]	47
Repcount [22]	6
Olympia	1306

对于仅依赖软件的方法，存在几种基于 3D HPE 的模型，这些模型仅专注于以下其中一项任务：动作检测 [30, 31]、重复计数 [22, 32, 33] 或质量评估 [34]。

几个现有模型尝试执行多种健身分析任务，但它们的范围明显有限，通常仅在少于 30 个动作上进行评估，并依赖手工制作的特征或模板化方法，

这些方法在不同的动作中泛化效果不佳。[35] 的作者使用训练好的多层感知器来对三个动作进行正确或错误的分类。该方法依赖于预先计算的欧几里得关节角度，使其难以扩展到更广泛的动作范围。在 [36] 中，使用随机森林分类器对四个动作进行分类；他们基于傅里叶变换的重复计数和每项运动的模板形式评估需要手动创建模板，这无法扩展到预定义运动之外。同样，在 [37] 中，预定义的运动模板用于七个动作的识别和重复计数。[29] 的工作使用自相似性度量评估重复质量，将用户动作与教练模板进行比较。然而，他们的方法不执行动作识别并限制于 30 个动作。[26] 的作者通过使用关节角度阈值将每个动作分为单位动作（收缩、放松、暂停、噪声）来评估 18 个动作的识别和重复计数。他们的方法需要为每个动作手动定义启发式方法。与这些方法相比，我们的模型利用运动-语言学习，在无需手工制作模板或每项运动调整的情况下，在数百个动作上进行泛化。

2.1. 基于 *Transformer* 的方法

监督的大型语言模型 (LLMs) 已经被证明能够学习自然语言和人类动作之间的关系 [38, 39, 40]。然而，所使用的 LLMs (Flan-T5-Base [41], Vicuna/LLaMA 7B-v1.5 [42, 43]) 需要大量的计算资源。它们采用基于标记的方法，将动作骨架序列使用向量量化变分自编码器编码为一系列动作标记。相关的自然语言标签被转换为文本标记 [44, 45, 46]。然后使用 LLM 进行序列到序列的建模，以学习动作标记和文本标记之间的关系。模型经过因果推理训练，使其能够执行动作到文本和文本到动作的生成。这意味着，给定一系列动作输入，模型可以生成对应的执行该练习的文本描述（动作到文本），反之亦然。这些工作 [38, 39, 40] 使用 HumanML3D [47]，这是一个基于骨架的数据集，包括通过动作捕捉系统记录的一般人类运动（行走、坐着等）。每个动作都被注释了描述该运动的字幕。然而，该数据集不包括各种健身练习。此外，HumanML3D 中的骨架进行了广泛的预处理。所有骨骼关节的角速度通过每个视频的逆运动学提取，然后使用正运动学应用于一个单一的预定骨架，以创建不同运动的统一骨架数据集。最后，骨架在 $t=0$ 时从几何上对齐至面向 $x,y,z=(0,0,1)$ 。总的来说，现有的动作-语言方法依赖于在精心策划的一般运动数据集上训练的复杂架构。相较之下，我们的方法更为高效，专为健身练习量身定制，并针对实际应用进行了优化，而没有过多的模型复杂性。

3. 方法

3.1. 概述

如图 1 所示，我们首先使用 MotionBERT [24] 从 RGB 视频中提取 3D 关键点，并将每个视频获得的所有骨架组合成一个单一的“图像”。该模型

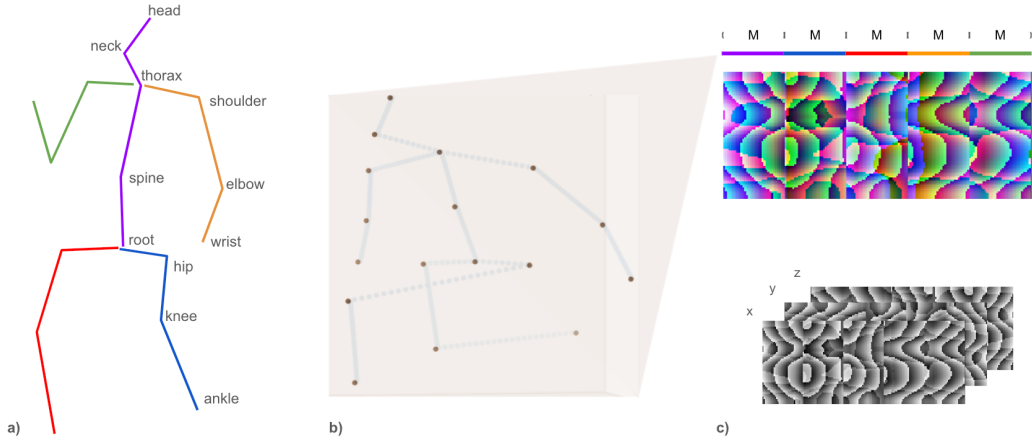


Figure 2: (a) 由 MotionBERT 检测得出的 Human3.6M 格式关键点；(b) 每个运动链插值得出的 $M=64$ 个位置的骨架结果；(c) 将所有运动链拼接在一起后的结果图像（上方），同时保持 x, y, z 位置（下方）。

通过最小化方程 (1) 的多类交叉熵损失函数来学习骨骼运动与自然语言标签之间的关系。在训练过程中，每个批次包括重复计数和锻炼检测这两个任务。可训练的模型包括 (a) 线性投影层，(b) 视觉变换器，以及 (c) 分类头。

3.2. 动态图像

类似于 [48]，我们将骨架运动转换为运动图像。设 $\mathbf{S}$ 为骨架数据，一个形状为 $(T, N, 3)$ 的张量，其中 T 是时间步的数量， N 是关节的数量，每个关节有一个 3D 位置 (x, y, z) 。我们的工作使用 Human3.6M [49] 格式，它有 $N = 17$ 个关键点，如图 2 a 所示。为了减少人类运动中未观察到的高频噪声，我们使用窗口大小 $w = 3$ 对时间进行平均，因此对于每个关节 n 和维度 $d \in \{x, y, z\}$ ，数据在时间上与之前和之后的值进行平均。这产生了平滑的骨架数据 $\mathbf{S}'$ 。这分为 5 个运动链，每个链涉及 17 个关节的子集。我们将运动链定义为：[根，脊柱，胸腔，颈部，头部]，[根，左髋，左膝，左踝]，[根，右髋，右膝，右踝]，[胸腔，左肩，左肘，左腕]，[胸腔，右肩，右肘，右腕]。一个运动链被定义为 $\mathbf{C}(t) = [\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_P]$ ，用于表示时间 t 的关节位置，其中 P 是链中的关节数量，每个为 $\mathbf{n}_i \in \mathbb{R}^3$ 。对于每个运动链，我们对每个空间维度 (x, y, z) 执行 1D 线性插值，以创建 $\mathbf{C}'(t) = [\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_M]$ ，包含 $M = 64$ 个均匀间隔的点，如图 2 b 所示。我们将 5 个插值链 $\mathbf{C}'(t)$ 连接起来以创建一个骨架运动图像，定义为 $\mathbf{I} \in \mathbb{R}^{(M \times 5) \times T \times 3}$ 。

Table 2: 词类和相关权重。

Category Name	Weight	Examples
Core Action	1.0	curl, squat, press, hinge
Spatial Adjective Variation	1.0	alternating, wide, supine
Noun Variation	1.0	burpee, clamshell, skater, diamond
Body Part Noun	1.0	bicep, hamstring, shoulder
Temporal Adverb Variation	0.4	slow, hold, eccentric, power
Equipment Noun	0.1	dumbbell, chair, wall, heavy
Combination Conjunction	0.5	and, plus, with
Uncategorized	0.0	wacky, truffle, vogue
Integers	1.0	1, 2, 3... 30

3.3. 多类别标记

锻炼类型的标签可能会有很大的不同。有些锻炼被描述性地标记，例如“右腿向前的反向弓步”，而其他锻炼则获得了口语化的描述或独立的名称，如“阿诺德推举”或“鸟狗”。一个核心动作（例如，弓步）可以通过副词（慢弓步）、定语名词（滑冰者弓步）、空间相关描述符（右腿弓步）或形容词（反向弓步）来修饰。此外，还有带有连接词的复合锻炼（例如，“挺举推举”）和特定于设备的锻炼（墙坐）。

我们将单个单词以及整数（从 1 到 30）视为预测类别。这允许进行练习检测，其中一个练习至少属于一个类别，以及重复计数。总的来说，这些预测类别形成一个词汇集合 $\mathcal{V}$ ，然而并非所有的单词都同等重要。我们将单词分为 9 类之一，如表 2 所示，其中每个类别都被分配了一个重要性权重。

考虑一个包含文本标签的数据集。每个标签 L_i 由一组单词组成： $L_i = \{w_1, w_2, \dots, w_m\}$ ，其中 m 是标签 L_i 中不同单词的数量。定义词汇表 $\mathcal{V}$ 为数据集中所有唯一单词的集合： $\mathcal{V} = \bigcup_i L_i$ 。每个类别都被分配了一个权重；我们使用权重集合 $\mathcal{W} = \{0, 0.1, 0.4, 0.5, 1\}$ 。根据所属类别，每个词 $w \in \mathcal{V}$ 被分配一个权重 $f: w \rightarrow \mathcal{W}$ 。所有类别及其对应的词汇可以在补充信息部分 5 中找到。每个词 $w \in \mathcal{V}$ 被视为一个独特的预测类别。对于多类别预测，我们定义了大小为 $|\mathcal{V}|$ 的多热编码向量 $\mathbf{y}_i$ ，用于标签 L_i ： $\mathbf{y}_i \in \{0, 0.1, 0.4, 0.5, 1\}^{|\mathcal{V}|}$ ，其中：

$$\mathbf{y}_{i,j} = \begin{cases} f(w_j), & \text{if } w_j \in L_i \\ 0, & \text{otherwise} \end{cases}$$

模型输出一个预测向量 $\hat{\mathbf{y}}_i$ ，多分类损失函数计算如下：

$$\mathcal{L}_i = - \sum_{j=1}^{|\mathcal{V}|} y_{i,j} \log \hat{y}_{i,j} \quad (1)$$

这表示预测和真实多热编码向量之间的分类交叉熵损失。

3.4. 训练和模型拟合

我们使用迁移学习，通过使用预训练权重的现有架构 ViLT [50] 来学习运动和语言之间的关系。ViLT 是在图像和它们的文本描述上进行预训练的。我们选择 ViLT 而不是 [48] 中使用的 CLIP 样式对比对齐，因为基于 CLIP 的模型主要学习图像/运动和文本之间的全局关系。这可以在图 3 b 中看到。由于 CLIP 在图像-文本级别合并交叉熵损失，它没有明确建模单个运动块和特定词语之间的细粒度对应。相较之下，我们修改后的 ViLT 在共享转化器内直接处理运动块和文本，使其能够学习子运动（运动链）和语言成分之间的详细关系，如图 3 a 中所示。这使得 ViLT 更适合多任务运动分析，其中局部运动-语言对齐至关重要。

在 ViLT 架构中，文本和图像嵌入在计算几层自注意力之前被连接在一起，这些自注意力层学习这两种模态之间的交互。为了遵循 ViLT 最大图像尺寸 640×384 ，我们以其原始帧速率的一半对运动块进行时间下采样（从 25 FPS 到 12.5 FPS）。这允许运动的持续时间最长可达 51.2 秒。比 640×384 小的运动图像 I 用零填充； I_p 表示填充后的图像。

我们从头开始训练多类分类器头。我们用词汇表 $\mathcal{V}$ 初始化 471 个类，该词汇表由代表 1749 个练习标签的 441 个独特单词以及表示可能重复次数的 1 到 30 之间的整数组成。在训练期间，模型的输入是一个填充的动作图像 I_p 和相应的任务相关问题，如在第 3.5 节定义。可学习的输出是一个多热编码的向量，如在第 3.3 节定义。

数据集被随机分为 80% 个训练视频、10% 个验证视频和 10% 个测试视频。使用完全注意力掩码来构建 Transformer 模型。采用 Adam 优化器并使用早停法（耐心值为 5 个周期）。学习率设定为 5×10^{-5} ，批量大小为 32。图 4 展示了多类交叉熵损失的损失函数结果，表明模型在数据集上的学习效果良好。

3.5. 数据集

Olympia 包含几个现有的数据集 [29, 51, 52, 53, 22] 以及开源健身视频。这些开源视频从不同的距离和角度拍摄，包含各种背景和光照场景。总体而言，这些视频来源产生了 7,618 个视频。

Existing Datasets: Fit3D [29] 是一个包括从四个不同摄像机角度记录的 47 个练习的集合。我们使用了 8 个具有对应练习标签和重复次数数据

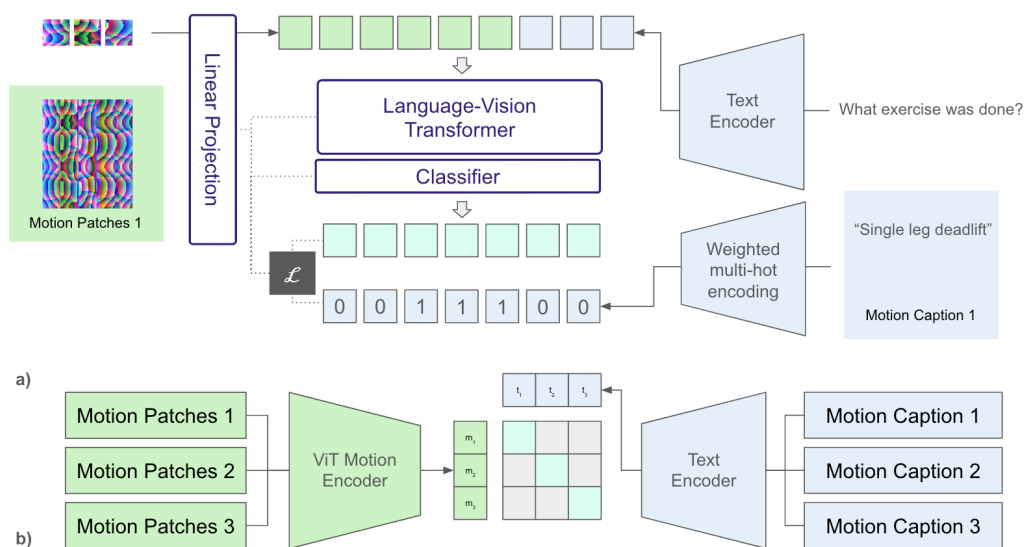


Figure 3: 比较运动语言理解的架构类型。a) 显示了我们的架构，该架构使用多类交叉熵来学习单词和运动块之间的关系。b) 显示了其他作品使用的对比损失方法，以学习运动和语言之间的全局关系。

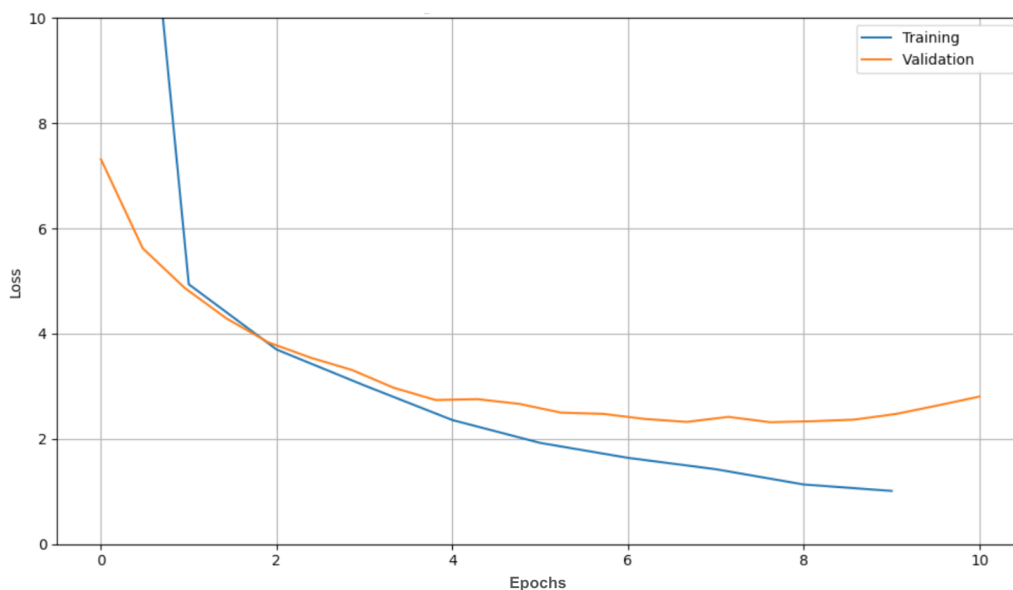


Figure 4: 在 Olympia 数据集上的训练和验证损失。

的受试者。在这 47 个练习中，有 17 个被标记为 *warmupX* 练习，其描述未给出。我们保留了 *warmupX* 的特有名称，用于测试模型学习全球运动特性以及子运动特性的能力。Fit3D 中剩下的 30 个练习每个受试者重复了 5 次。为了增加重复的多样性，我们之前引入了一个变化数据集 [33]，其中 6 种不同练习的重复次数在各受试者间变化了 5-11 次。这些练习包括：膝盖钻石俯卧撑、钻石俯卧撑、串联深蹲、半蹲、椅子深蹲、蹲。InfiniteRep [51] 是一个合成数据集，包括 10 种练习的 100 个视频（鸟狗、举臂、自行车卷腹、卷曲、飞行、抬腿、过头推举、俯卧撑、深蹲、超人）。我们从 [34] 的数据集中包含了 8 个受试者从 5 个摄像机角度进行的 4 个练习（单腿罗马尼亚硬拉、罗马尼亚硬拉、单腿小深蹲、外旋肌袖旋转）。MM-Fit [52] 包含 10 个练习：（深蹲、俯卧撑、哑铃肩推、箭步、站姿哑铃划船、仰卧起坐、哑铃三头肌扩展、二头肌卷曲、坐姿哑铃侧平举、开合跳）。UCO [53] 包含 10 个练习：坐姿屈膝无支撑右下部，水平负重开合站姿左上部，圆形摆动站姿左上部，坐姿屈膝无支撑左下部，坐姿屈膝有支撑右下部，圆形摆动站姿左上部，举起伸展腿仰卧右下部，床支撑屈膝仰卧左下部，水平负重开合站姿右上部，肩部屈曲坐姿右上部）。

Open-source: 我们从 7 个人处收集了锻炼视频，每个视频长度从 10 到 45 分钟不等，总共 2,273 段视频。我们的团队标注了锻炼的时间戳、名称和重复次数。每个视频只包含一个锻炼组。与以往的数据集不同，这个开源健身数据集包括将多个动作组合成的复合锻炼，例如，腿部弓箭步加膝盖驱动。

我们使用以下步骤整理了自然语言标签。所有动词都转化为单数形式（例如，squats 转为 squat）。连词被保留（例如 +, &, 'and'），但除了这些连词外，其他特殊字符被移除（例如，Child's pose）并且连字符被移除（例如，push-up 变为 pushup）。所有字母组成的标签都被转换为小写并纠正拼写。对于解剖类标签（例如，arms 与 arm），我们不会转换为单数，因为它们包含关于运动的重要信息。缩写被扩展（L 变为 left，rdl 变为 romanian deadlift）。表示姿势的字母被保留（例如 w raise, t press, x plank），数字缩写被纠正（例如 1-leg 变为 single leg）。

所有的 Olympia 视频都有一个相应的动作字幕标签 L_i ，表示练习类型和重复次数。对于每个视频，我们生成一对问题回答，分别对应两个任务：i) 运动识别和 ii) 重复计数。答案是视频的练习名称或重复次数。我们以问题的形式预定义了每个任务的几种自然语言变体，例如，“进行的是什么练习？”和“完成了多少次重复？”以创建详细的问题回答配对。这些变体显示在补充信息部分 5 中。最后，我们为 7,618 个独特的视频创建了 152,360 对问题回答配对。

每个视频都经过预处理，使得一个视频包含单个锻炼的重复集合。视频通过 MotionBERT 转换为 3D 关键点，每个视频帧提取 17 个关键点。这

17 个关键点遵循 Human3.6M [49] 格式。随后，这些人体姿态估计被转换为运动图像。

我们的模型执行两个任务：运动识别和重复计数。这里，我们展示用于每个任务的准确性评分方法。对于重复计数，我们采用其他重复动作计数工作中使用的“偏差一个”（OBO）和平均绝对误差（MAE）作为指标。偏差一个指的是模型预测计数 $count_{pred}^i$ 在测试视频的实际重复次数 $count_{gt}^i$ 内相差不超过一个重复数的测试视频的比例：其中 V 为评估的视频数量。

平均绝对误差是预测计数与真实值之间的平均绝对差异。由于模型可以预测多个类别并且有不同的置信度得分，我们仅考虑重复计数中置信度最高的预测类别。

$$MAE = \frac{1}{V} \sum_{i=1}^V |count_{gt}^i - count_{pred}^i| \quad (2)$$

对于练习识别，当对模型的准确性进行评分时，我们采用“部分信用”的概念，使用类似于 BLEU [54] 的指标，其中我们使用类别重叠。给定模型输出的 logits z ，我们计算 softmax 概率： $\hat{y}_i = \text{softmax}(z_i)$ ，其中每个元素 $\hat{y}_{i,j}$ 代表类别 j 和相关词 w_j 的预测置信度。如图 5 b 所示，只有置信度高于阈值 $\tau = 0.05$ 的预测 $\hat{y}_j$ 才会被考虑，这样： $\mathcal{P}_i = \{w_j \mid \hat{y}_{i,j} \geq \tau\}$ 。

如果 G_i 表示样本 i 的真实词集，则匹配词的数量为 $N_i = |\mathcal{P}_i \cap G_i|$ 。练习检测准确率计算为： $\text{Accuracy}_i = \frac{N_i}{|G_i|}$ ，如图 5 c 所示。最终的数据集级别准确率是针对所有样本进行平均计算的。

我们研究了模型在动作检测和重复计数方面的准确性。图 6 描绘了数据标签的分布以及对应的模型准确性。为了评估模型在不同运动描述长度上的表现，我们在图 6 a 中绘制了动作检测准确性与运动描述长度的函数关系。如预期，随着描述长度的增加，模型的准确性下降。图 6 b 中的混淆矩阵显示了预测和真实情况的词数。每一列代表在例子中预测的词数，这些例子只有固定数量的单词，例如，所有包含只有四个真实词的例子。即便在模型可以理论上预测出 20 个单词的 $\tau = 0.05$ 置信阈值情况下，模型仍倾向于低估与真实长度相比的单词数量。由于模型在动作检测和重复计数两个任务上进行了均等训练，至少一半的真实编码向量 y_i 只有一个非零元素。另外，我们在图 6 c & d 中展示了数据集中重复次数的分布。视频中显著集中于 5 和 10 次重复，是由于常见的健身实践是以 5 到 10 次重复为一组的方式进行锻炼。我们检查了模型在动作检测和重复计数上的准确性。图 6 描绘了数据标签的分布以及相应的模型准确性。为了评估模型在不同动作描述长度下的表现，我们在图 6 a 中绘制了动作检测准确性与动作描述长度之间的函数关系。如预期一样，随着描述长度的增加，模型的准确性下降。图 6 b 中的混淆矩阵显示了预测和真实标签中的词数。如果真实标签包含 4 个词，模型倾向于在与真实长度相比时低估词数，即使设有自

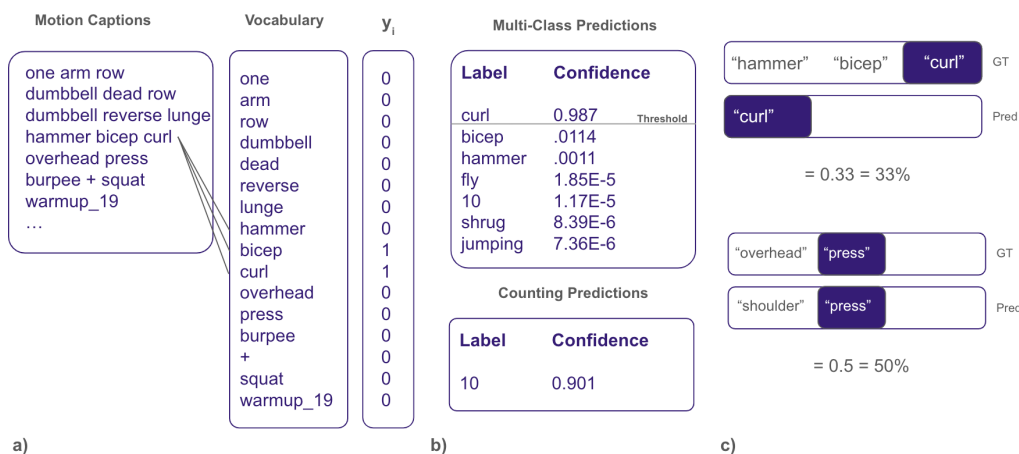


Figure 5: a) 使用共享词汇和主要/次要类别权重进行多类别标注。例如，设备词汇，如“hammer”，被赋予权重 0，因为 3D HPE 不检测设备。b) (上图) 显示多个预测类别的置信度得分，这些类别被考虑用于部分得分 (下图) 重复计数仅考虑最有置信度的预测。c) 提供了两个如何计算多类别运动检测的部分准确度的例子。

信的阈值 $\tau = 0.05$ ，理论上模型可以预测 20 个词。由于模型在动作检测和重复计数这两个任务上均衡训练，至少一半的真实编码向量 y_i 只有一个非零元素。此外，我们在图 6 c & d 中展示了数据集中重复计数的分布。大量集中在 5 次和 10 次重复的视频可归因于常见的健身做法，即以 5 到 10 次重复为一组进行锻炼。在计算绝对误差方面，图 6 c 显示出随着重复次数的增加，绝对误差也在增加。这是可以预料的。当在相同时间内进行更多的重复时，由于模型允许的最大图像尺寸，信号可能会减弱。尽管模型在重复次数增加时得分增加，但 OBO 准确率下降。准确率获得的精确匹配减少，而“偏差为一”的评分增加。这在样本数量多的 5 到 10 次重复区域中显示出来；然而，由于样本较少，重复次为 > 10 时结果变得不太可靠。

对于奥林匹亚中包含的先前存在的数据集，我们详细解析模型对运动检测的准确性。该模型在不同数据集上的运动检测准确性分别为：UCO 数据集上的准确性是 98.77%，MM-Fit 上是 73.66%，Fit3D 上是 73.09%，以及 InfiniteRep 上是 69.75%。这表明该模型在不同的运动数据集上表现出一致性。

我们定性地观察了以下模型性能。对于特定的锻炼名称，观察到较低类别重叠率，例如“相扑深蹲脉冲”或“波比运动加两个弓步跳”。视频数据中观察到的模糊性也可能导致分类不佳。例如，模型将硬拉锻炼标记为“划船”，因为被试者从地上以类似划船的方式抬起杠铃。此外，模型还有提高的空间，因为模型在一些锻炼中也被观察到给出了完全错误的分类。在

Table 3: LiFT 在不同数据集上的性能结果。

Results	Detection	OBO	MAE
Olympia			
UCO	0.9877	1.0	0.0
Open Source	0.7982	0.687	1.50
MM-Fit	0.7366	0.979	0.126
Fit3D	0.7309	0.997	0.007
InfiniteRep	0.6975	0.779	0.772
RepCount	0.6078	0.593	2.167
Experiments			
Fit3D (Fine-Tune)	0.89	.996	0.131
RepCount (Few-Shot)	0.24	.186	4.526
KIMORE (Zero-Shot) ²	0.208	—	—
UI-PRMD (Zero-Shot)	0.167	.647	1.892
UI-PRMD (Few-Shot (40)) ³	0.497	—	—
UI-PRMD (Few-Shot (200)) ³	0.745	—	—

这里，真实值与预测类别显著不同，例如腿弓步右卷曲与深蹲、左弓步手臂推举与深蹲。出乎意料的是，模型能够以 91.27% 的准确性预测 *warmupX* 锻炼，这验证了模型除了使用单个单词类别来学习子动作外，还能够学习运动的全局表示。

为了理解我们模型的泛化能力，我们在 4 个公开可用的数据集上测试该模型。具体来说，我们想展示以下几点：模型的微调，即在评估前特别针对包含大量样本的数据集进行训练；少样本学习，即在评估前只使用有限的数据集样本来训练模型；以及零样本学习，即在训练数据中不包含该数据集。我们在 Fit3D 上展示微调，因为该数据集包含许多种类的锻炼的许多例子。我们在 RepCount 上展示少样本学习，因为该数据集中锻炼种类有限（6 种锻炼），但每种锻炼有许多样本。我们在 UI-PRMD 和 KIMORE 数据集上测试零样本能力，因为这些数据集的锻炼/样本有限，并且包括与 Olympia 训练数据中所见的健身锻炼不同的物理康复锻炼。仅在 Fit3D 上微调我们的模型，我们在 1246 个样本上训练模型并在 160 个样本上测试。该模型实现了运动检测得分为 89% 和次数计算得分为 99.6% OBO，且

²KIMORE dataset has a single recorded motion for each exercise

³UI-PRMD has each exercise performed for 10 repetitions. During fine-tuning on this data, the model learns to always predict 10.

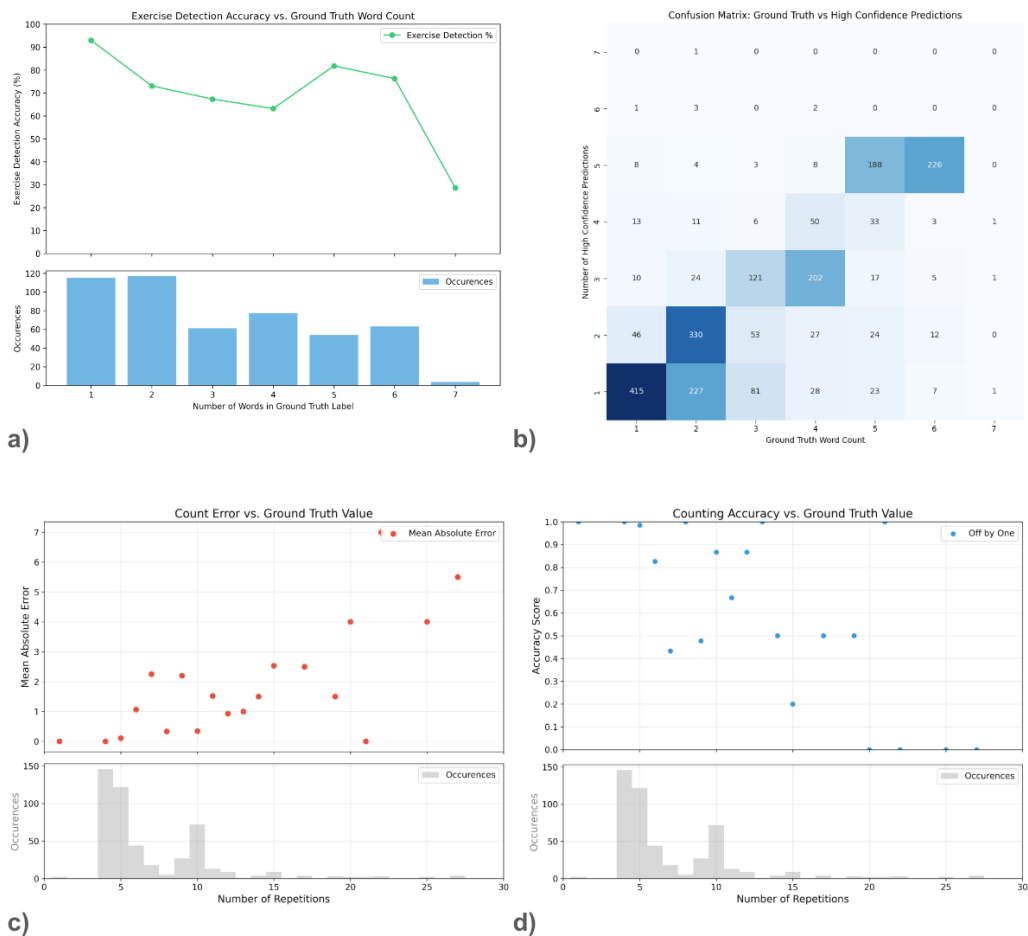


Figure 6: a) 比较模型在 Olympia 中面对不同练习标签词长时的准确性。b) 混淆矩阵突出显示预测标签中的单词数与真实标签中的单词数。c) 说明模型在不同重复次数下重复计数的准确性。d) 说明在不同重复次数下重复计数的绝对误差。

MAE 为 0.13。我们将高次数计算得分归因于所有受试者都进行了所有锻炼的 5 次重复，所以在此数据集上的微调导致了这种单一类别预测。

对于 RepCount 上的小样本学习，模型在 100 个例子上训练，并在 800 个例子上测试。RepCount 数据集包含 7 种不同的锻炼：俯卧撑、引体向上、开合跳、深蹲、卧推、前臂抬起、仰卧起坐。我们排除了鞍马，因为缺乏一致的重复次数。重复计数的准确性与其他报告的成绩 [22] 在 18.9% OBO 处具有竞争力，平均绝对误差 (MAE) 为 4.5，锻炼检测得分为 24%。对 RepCount 数据集进行定性分析后，我们得出结论，这种较低的性能是数据问题。数据集中的视频通常包含编辑过的视频，摄像机角度变化，显示多个主体正在进行锻炼，并且在特定锻炼类别内有许多锻炼修改。

测试模型的零样本能力时，我们更改了预测类别词汇以包括新的词标签。没有任何先前训练的情况下，如果将这些新词添加到预测类别中，模型会有效地忽略它们。我们使用了 KIMORE，它包含 5 项无重复次数的物理康复运动。模型在运动检测中取得了一个得分。我们还测试了 UI-PRMD，它包含 10 项健身运动，所有运动都进行了 10 次重复。模型在运动检测中取得了一个得分，OBO 为 64.7%，MAE 为 1.96。为了定性评估模型，我们报告了 UI-PRMD 的真实值与预测的比较。模型在某些运动方面的泛化能力良好：(1) 深蹲与蹲下，(2) 站-坐与蹲下，(3) 站立肩部伸展与臂推伸展，(4) 站立肩部外展与臂抬高，(5) 侧弓步与侧蹲。在以下方面有合理的表现：(6) 站立肩部平面举与臂抬高飞，(7) 站立肩内外旋与臂抬高，(8) 直线弓步与单腿蹲。在一些方面表现较差：(9) 跨栏步与手臂，(10) 站立主动腿举与手臂。运动 9 和 10 是在举起手臂的情况下进行的。

我们通过向模型提供训练样本进一步研究在 UI-PRMD 上的微调效果。当模型提供了 40 个训练样本并在 320 个样本上进行评估时，运动检测提升到 49.5%。当模型提供了 200 个训练样本并在 160 个样本上进行评估时，运动检测提升到 74.5%。

3.6. 消融和替代研究

我们在算法的多个部分中使用消融或替代来测试每个步骤的有效性。我们在表格 3 中显示了模型的性能比较。

为了展示模型在不使用多类别预测情况下的能力，我们将整个练习标签视为一个单一类别。每个完整的标签 L_i 被视为一个独特的类别： $C' = \{C'_1, C'_2, \dots, C'_{|D|}\}$ ，其中 $|D|$ 是数据集中唯一标签的数量。这样就产生了一个由 1,749 个独特标签组成的词汇表 $\mathcal{V}$ ，然后用于初始化模型的分头。在这个实验中，我们没有使用带有类别重叠的“部分积分”。如表 3 所示，获得的 61.8% 准确率是基于全或无的字符串匹配。因此，我们将此得分称为我们的“基线”。如预期的那样，练习检测的准确率低于“部分积分”方法的准确率 76.5%。此外，重复计数的准确率几乎没有受到影响。

Table 4: LiFT 的消融和替代结果。

Exercise Identification	Accuracy	Deviation
Baseline Single-Class	0.618	—
MotionBERT Multi-class	0.765	—
Mediapipe Multi-class	0.654	—
Repetition Counting	OBO	MAE
Baseline Single-Class	0.829	1.11
MotionBERT Multi-class	0.853	0.645
Mediapipe Multi-class	0.717	1.43

82.9% 与 85.3%。这是预料之中的，因为不同的分类方法旨在影响练习检测得分；然而，我们推测略低的重复计数准确率可能是由于潜在预测类别的大词汇表增加所致。

3.6.1. 3D 人体姿态估计对比

我们测试了模型在较低质量数据上的表现能力，即使用 Mediapipe 从 RGB 视频中进行 3D 人体姿态估计。我们使用 Mediapipe 版本 0.10.20 [23]，因为它有在移动应用中运行的潜力。在这里，我们从奥林匹亚使用的视频中提取 3D HPE。我们使用‘full’模型，其已被证明可以在移动设备上以超过每秒 30 帧的速度处理视频。Mediapipe 无法检测到从 MotionBERT 获得的 Human3.6M 格式中存在的关键点根部、脊椎和胸廓。因此，我们通过插值计算右髋和左髋之间的中点作为根关节，右肩和左肩之间的中点作为胸廓，根部与胸廓肩膀之间的中点作为脊椎关键点。后续处理步骤与第 3.2 节中描述的 MotionBERT 关键点使用的相同。结果的准确性显示在表 3 中。

众所周知，Mediapipe 的 3D HPE 预测误差高于 MotionBERT。基于此，我们预计 LiFT 模型在训练和评估时，使用获得的质量较低的运动数据，其预测准确性可能会较低。在使用 Mediapipe 运动时，我们观察到运动检测的准确性降低了 14.5%，而 OBO 重复计数的准确性降低了 15.9%。这与我们的预期一致；然而，它强化了这样的概念：一个端到端的移动系统是可能的，使用我们的运动-语言模型。我们将这一移动实现留作未来的工作，并推测需要更准确、移动友好的 3D HPE。

虽然我们的方法在多任务运动分析中表现出色，但仍然存在一些局限性。首先，模型需要完整的一组重复来分析一个锻炼，这可能会限制实时反馈。其次，我们的 3D 人体姿态估计方法没有检测到锻炼设备（例如哑铃、阻力带），这对于跟踪依从性和监测长期力量进展可能是有价值的。第三，我们的模型尚未评估重复质量，尽管我们相信当前的架构可以通过额

外的训练数据支持这一点。我们推测，从更精确的 3D HPE 模型获得的高保真度 3D 姿态估计可以实现对部分动作的更细致分析。

最后，模型需要进一步改进以能够在移动设备上运行，利用较低质量的 3D 人体姿态估计（HPE）如 Mediapipe。LiFT 模型很适合移动设备，需要 433 MB 的 RAM 来存储使用单精度浮点值的模型权重。相比之下，上述的大型语言模型权重需要 25.6 GB 的 RAM 与单精度浮点值。需要进行更多研究以优化在移动设备上的高效部署。未来的工作将专注于移动设备上的实时 3D HPE、重复质量评估，以及改进对未见过的锻炼的泛化能力。解决这些限制将进一步提高模型在 AI 驱动健身追踪中的适用性，使智能教练更具可访问性、可扩展性和效果。

4. 结论

在这项工作中，我们引入了一种强大的多任务运动分析模型，能够在大规模上执行运动检测和重复计数，远远超过了以前的方法。虽然现有方法通常限制于一小部分运动集，但我们的模型通过利用多模态视觉-语言表示和大规模在线健身数据，在数百种可能的动作中实现了泛化。通过在多个公开的数据集上的严格评估，我们展示了模型在已知和未知的运动类别中都具有强大的性能，突显了模型的适应性和可扩展性。

我们的研究结果表明，视觉-语言模型可以有效地弥合多任务运动分析中的差距，为家用运动追踪和智能健身指导的实际应用铺平道路。由于只需要 RGB 视频，我们的方法实现了一种仅软件、注重隐私的解决方案，可以在消费者的智能手机上部署，使普通用户也能使用 AI 驱动的健身分析工具。

References

- [1] J. L. Dieleman, J. Cao, A. Chapin, C. Chen, Z. Li, A. Liu, C. Horst, A. Kaldjian, T. Matyas, K. W. Scott, A. L. Bui, M. Campbell, H. C. Duber, A. C. Dunn, A. D. Flaxman, C. Fitzmaurice, M. Naghavi, N. Sadat, P. Shieh, E. Squires, K. Yeung, and C. J. L. Murray, “US health care spending by payer and health condition, 1996-2016,” vol. 323, no. 9, pp. 863–884.
- [2] J. R. Zadrow, S. Décar, M. O’Keeffe, Z. A. Michaleff, and A. C. Traeger, “Overcoming overuse: Improving musculoskeletal health care,” vol. 50, no. 3, pp. 113–115. Publisher: Journal of Orthopaedic & Sports Physical Therapy.

- [3] J. Zadro, M. O' Keeffe, and C. Maher, "Do physical therapists follow evidence-based guidelines when managing musculoskeletal conditions? systematic review," vol. 9, no. 10, p. e032329. Publisher: British Medical Journal Publishing Group Section: Rehabilitation medicine.
- [4] C. Schwalm, C. Jamart, N. Benoit, D. Naslain, C. Prémont, J. Prévét, R. Van Thienen, L. Deldicque, and M. Francaux, "Activation of autophagy in human skeletal muscle is dependent on exercise intensity and AMPK activation," vol. 29, no. 8, pp. 3515–3526.
- [5] A. McTiernan, C. M. Friedenreich, P. T. Katzmarzyk, K. E. Powell, R. Macko, D. Buchner, L. S. Pescatello, B. Bloodgood, B. Tennant, A. Vaux-Bjerke, S. M. George, R. P. Troiano, and K. L. Piercy, "Physical activity in cancer prevention and survival: A systematic review," vol. 51, no. 6, pp. 1252–1261.
- [6] M. K. Abramowitz, C. B. Hall, A. Amodu, D. Sharma, L. Androga, and M. Hawkins, "Muscle mass, BMI, and mortality among adults in the united states: A population-based cohort study," vol. 13, no. 4, p. e0194697.
- [7] T. E. Lambert, L. A. Harvey, C. Avdalis, L. W. Chen, S. Jeyalingam, C. A. Pratt, H. J. Tatum, J. L. Bowden, and B. R. Lucas, "An app with remote support achieves better adherence to home exercise programs than paper handouts in people with musculoskeletal conditions: a randomised trial," vol. 63, no. 3, pp. 161–167.
- [8] U. H. Mitchell, H. Lee, H. E. Dennis, and M. K. Seeley, "Quality of knee strengthening exercises performed at home deteriorates after one week," vol. 23, no. 1, p. 164.
- [9] H. Uzawa and S. Davis, "Outcome measures for adherence to home exercises among patients with chronic low back pain: a systematic review," vol. 30, no. 4, pp. 649–653.
- [10] C. Hammer, L. Degerfeldt, and E. Denison, "Mechanical diagnosis and therapy in back pain: Compliance and social cognitive theory," vol. 9, no. 4, pp. 190–197. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/14038190701433241>.

- [11] M. Friedrich, G. Gittler, Y. Halberstadt, T. Cermak, and I. Heiller, “Combined exercise and motivation program: Effect on the compliance and level of disability of patients with chronic low back pain: A randomized controlled trial,” vol. 79, no. 5, pp. 475–487.
- [12] S. Milanko and S. Jain, “LiftRight: Quantifying strength training performance using a wearable sensor,” vol. 16, p. 100115.
- [13] Y. Liao, A. Vakanski, and M. Xian, “A deep learning framework for assessing physical rehabilitation exercises,” vol. 28, no. 2, pp. 468–477. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.
- [14] B. X. B. Yu, Y. Liu, K. C. C. Chan, Q. Yang, and X. Wang, “Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer’ s progression,” vol. 119, p. 108095.
- [15] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba, “Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance,” vol. 24, no. 1, pp. 280–291. Conference Name: IEEE Journal of Biomedical and Health Informatics.
- [16] H. Qu, Y. Cai, and J. Liu, “LLMs are good action recognizers,”
- [17] O. Levy and L. Wolf, “Live repetition counting,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3020–3028. ISSN: 2380-7504.
- [18] K. Chung, A. Kerrigan, and I. Dave, “Long action repetition counting with staged layer distillation,”
- [19] Y. Zhang, L. Shao, and C. G. M. Snoek, “Repetitive activity counting by sight and sound.”
- [20] H. Zhang, X. Xu, G. Han, and S. He, “Context-aware and scale-insensitive temporal repetition counting,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 667–675. ISSN: 2575-7075.
- [21] Y. Tang, W. Luo, J. Zhang, W. Huang, R. Jing, and D. Zhang, “MultiCounter: Multiple action agnostic repetition counting in untrimmed videos.”

- [22] H. Hu, S. Dong, Y. Zhao, D. Lian, Z. Li, and S. Gao, “TransRAC: Encoding multi-scale temporal correlation with transformers for repetitive action counting,” pp. 19013–19022.
- [23] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device real-time body pose tracking.”
- [24] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, “MotionBERT: A unified perspective on learning human motion representations.”
- [25] S. García-de Villa, A. Jiménez-Martín, and J. J. García-Domínguez, “A database of physical therapy exercises with variability of execution collected by wearable sensors,” vol. 9, p. 266.
- [26] S.-H. Cheng, M. A. Sarwar, Y.-A. Daraghmi, T.-U. İk, and Y.-L. Li, “Periodic physical activity information segmentation, counting and recognition from video,” vol. 11, pp. 23019–23031. Conference Name: IEEE Access.
- [27] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, and F. Verdini, “The KIMORE dataset: KInematic assessment of MOvement and clinical scores for remote monitoring of physical REhabilitation,” vol. 27, no. 7, pp. 1436–1448.
- [28] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, “A data set of human body movements for physical rehabilitation exercises,” vol. 3, no. 1, p. 2. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- [29] M. Fieraru, M. Zanfir, S. C. Pirlea, V. Olaru, and C. Sminchisescu, “AIFit: Automatic 3d human-interpretable feedback models for fitness training,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9914–9923, IEEE.
- [30] “Unlabeled action quality assessment based on multi-dimensional adaptive constrained dynamic time warping.”
- [31] R. Gajbhiye, S. Jarag, P. Gaikwad, and S. Koparde, “AI human pose estimation: Yoga pose detection and correction,” vol. 7, no. 5.
- [32] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, “Counting out time: Class agnostic video repetition counting in the wild.”

- [33] A. Postlmayr, B. Garg, P. Cosman, and S. Dey, “PersonalPT: One-shot approach for skeletal-based repetitive action counting for physical therapy,” vol. 34, p. 100516.
- [34] B. Garg, A. Postlmayr, P. Cosman, and S. Dey, “Short: Deep learning approach to skeletal performance evaluation of physical therapy exercises,” in *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, CHASE ’23, pp. 168–172, Association for Computing Machinery.
- [35] J. A. Francisco and P. S. Rodrigues, “Computer vision based on a modular neural network for automatic assessment of physical therapy rehabilitation activities,” vol. 31, pp. 2174–2183. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.
- [36] A. Jaiswal, G. Chauhan, and N. Srivastava, “Using learnable physics for real-time exercise form recommendations,” in *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 688–695.
- [37] V. S. Nair, S. Sree, J. Joseph, and M. Sivaprakasam, “Online action representation using change detection and symbolic programming.”
- [38] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, “MotionGPT: Human motion as a foreign language.”
- [39] L.-H. Chen, S. Lu, A. Zeng, H. Zhang, B. Wang, R. Zhang, and L. Zhang, “MotionLLM: Understanding human behaviors from human motions and videos.”
- [40] Z. Zhou, Y. Wan, and B. Wang, “AvatarGPT: All-in-one framework for motion understanding planning generation and beyond,” pp. 1357–1366.
- [41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer.”
- [42] “Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality | LMSYS org.”
- [43] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez,

- A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and efficient foundation language models.”
- [44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space.”
 - [45] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units.”
 - [46] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, “Fast Word-Piece tokenization.”
 - [47] C. Guo, “EricGuo5513/HumanML3d.” original-date: 2022-02-26T21:02:58Z.
 - [48] Q. Yu, M. Tanaka, and K. Fujiwara, “Exploring vision transformers for 3d human motion-language models with motion patches,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 937–946, IEEE.
 - [49] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” vol. 36, no. 7, pp. 1325–1339. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
 - [50] W. Kim, B. Son, and I. Kim, “ViLT: Vision-and-language transformer without convolution or region supervision.”
 - [51] “Papers with code - InfiniteRep dataset.”
 - [52] “MM-fit - the dataset of full-body physical exercises captured by multiple time-synchronized wearable sensing devices.”
 - [53] R. Aguilar-Ortega, R. Berral-Soler, I. Jiménez-Velasco, F. J. Romero-Ramírez, M. García-Marín, J. Zafra-Palma, R. Muñoz-Salinas, R. Medina-Carnicer, and M. J. Marín-Jiménez, “UCO physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises,” vol. 23, no. 21, p. 8862. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

- [54] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), pp. 311–318, Association for Computational Linguistics.

5. 补充信息

“核心动作”: “weight”: 1.0, “words”: [“深蹲”, “弓步”, “推举”, “推”, “拉”, “划船”, “硬举”, “屈髋”, “抬起”, “跳跃”, “跳动”, “飞行”, “推举器”, “迈步”, “坐下”, “跪下”, “踢”, “摆动”, “肩胛骨上举”, “屈膝礼”, “平板支撑”, “卷腹”, “轻拍”, “开合跳”, “旋转”, “步行”, “倾斜”, “伸展”, “屈伸”, “卷曲”, “反向踢”, “拳击”, “猛击”, “下沉”, “扭转”, “行进”, “桥式”, “洁净”, “折叠”, “桌面”, “拥抱”, “跑步”, “吸气”, “呼气”, “外展”, “内收”, “滚动”, “上勾拳”, “收腹”, “振翅”, “呼吸”, “抓举”, “跳跃”, “挤压”, “弯曲”, “猛推”, “举起”, “休息”, “穿梭”, “射击”, “慢跑”, “钩拳”, “爬行”, “拍打”, “提升”, “握住”, “泵”, “骑行”, “猛攻”, “猛打”, “滑动”, “耸肩”, “推拉”, “起身”, “摇动”, “旋转”, “升起”, “行走”, “鞠躬”, “猛砸”, “平衡”, “拍手”, “仰卧起坐”, “引体向上”, “俯卧撑”], “空间形容词变化”: “weight”: 1.0, “words”: [“左”, “右”, “窄”, “关闭”, “宽”, “错开”, “分裂”, “内”, “外”, “迷你”, “上”, “下”, “圆”, “圆形”, “摆动”, “对角线”, “单侧”, “双侧”, “v”, “x”, “直”, “弯曲”, “前”, “侧”, “其他”, “后”, “深”, “反向”, “向后”, “额前”, “前部”, “提升”, “伸手”, “移动”, “头顶”, “超过”, “交替”, “直立”, “站立”, “单腿”, “释放”, “姿势”, “t”, “完整”, “一半”, “外部”, “内部”, “倒置”, “交织”, “交叉”, “侧”, “点”, “驱动”, “w”, “中性”, “负”, “空心”, “打开”, “小”, “小”, “大”, “大”, “向下”, “通过”, “下降”, “三角形”, “下”, “隔离”, “蹲”, “下降”, “水平”, “垂直”, “坐”, “面向”, “l”, “r”, “低”, “高”, “扭曲”, “等”, “内侧”, “外侧”, “远端”, “近端”, “背部”, “悬挂”, “升高”, “单”, “双”, “顶部”, “底部”, “仰卧”, “俯卧”, “滑”, “后面”, “躺”, “倾斜”, “平行”, “超”, “固定”, “平”, “坐”, “越过”, “向下”, “一起”, “延伸”, “弯曲”, “一个”, “分开”, “跳”, “扭转”, “脉冲”], “名词变化”: “weight”: 1.0, “words”: [“高脚杯”, “相扑”, “死虫”, “保加利亚”, “罗马尼亚”, “黑客”, “鸟”, “超人”, “早安”, “大猩猩”, “阿诺德”, “波比”, “攀爬者”, “俯冲”, “孩子”, “熊”, “蜘蛛”, “螃蟹”, “猫”, “牛”, “骆驼”, “农夫”, “钻石”, “空气”, “自行车”, “rdl”, “驴”, “安静”, “海豚”, “星星”, “蛤蜊壳”, “消防栓”, “锯子”, “骡子”, “剪刀”, “锤子”, “造人者”, “小狗”, “彩虹”, “碎颅”]

锤”，“狗”，“眼镜蛇”，“海豹”，“突击队员”，“针”，“线”，“滑冰者”，“蛤蜊”，“车轮”，“瑜伽”，“热身”，“串联”，“游泳者”，“山”，“斗牛犬”，“捕食者”，“滑雪者”，“俄罗斯式”，“罐头”，“冲浪者”，“海星”，“折刀”，“光环”，“青蛙”，“驾驶”，“蝴蝶”，“拳击手”，“跑步者”，“作弊”，“牧师”，“专注”，“土耳其式”，“娘娘腔”，“公主”，“女孩”，“罗马式”，“三重奏”，“囚犯”，“尺蠖”，“船”，“鸽子”，“人”，“制造者”] “身体部分名词”：“weight”：1.0，“words”：[“肱三头肌”，“手臂”，“腿”，“核心”，“二头肌”，“肩膀”，“膝盖”，“脚趾”，“小腿”，“绳肌”，“四头肌”，“胸部”，“斜方肌”，“背部”，“三角肌”，“手”，“脚后跟”，“颈部”，“臀肌”，“腹肌”，“肩袖”，“臀部”，“脚踝”，“手掌”，“上部”，“下部”，“躯干”，“身体”，“胫骨前肌”，“腹股沟”，“髋部”，“斜肌”，“脊椎”，“臀”，“前锯肌”，“下巴”，“背阔肌”，“脚”，“手腕”，“肘部”，“大腿”]，

“热身”：“权重”：1.0，“单词”：[“热身 1”，“热身 2”，“热身 3”，“热身 4”，“热身 5”，“热身 6”，“热身 7”，“热身 8”，“热身 9”，“热身 10”，“热身 11”，“热身 12”，“热身 13”，“热身 14”，“热身 15”，“热身 16”，“热身 17”，“热身 18”，“热身 19”] “时间副词变体”：“weight”：0.4，“words”：[“慢”，“快”，“保持”，“节奏”，“离心”，“等距”，“拉伸”，“秒”，“节奏”，“弹出”，“迅速”，“静态”，“动态”，“力量”，“等长”，“部分”，“快速”，]，“设备名词”：“权重”：0.1，“词语”：[“smith”，“cable”，“sled”，“trx”，“resistance”，“bodyweight”，“barbell”，“battlerope”，“chair”，“weight”，“rope”，“wall”，“without”，“with”，“support”，“supported”，“bed”，“couch”，“bench”，“dumbbell”，“band”，“medicine”，“kettlebell”，“ball”，“bosu”，“box”，“floor”，“suitcase”，“sky”，“world”，“machine”，“bar”，“heavy”，“light”，“board”，“table”]，“组合连接词”：“权重”：0.5，“单词”：[“和”，“然后”，“+”，“&”，“与”，“组合”，“与”，“到”，“所有”，“整个”，“同时”，“-”]，“未分类”：“weight”：0.1，“words”：[“时尚”，“古怪”，“变化”，“松露”，“远”，“离开”，“道路”，“环绕”，“来自”，“关闭”，“”]

”提示”：“prompt”：“Q: 做了多少次单臂划船? A:”，“prompt”：“Q: 数一数这个视频中的单臂划船次数。A:”，“prompt”：“Q: 单臂划船的总次数是多少? A:”，“prompt”：“Q: 你观察到多少次单臂划船? A:”，“prompt”：“Q: 完成了多少次单臂划船: A:”，“prompt”：“Q: 告诉我单臂划船的次数。A:”，“prompt”：“Q: 正在进行什么练习? A:”，“prompt”：“Q: 命名这个练习: A:”，“prompt”：“Q: 视频中展示的是哪种动作? A:”，“prompt”：“Q: 识别正在演示的练习: A:”，“prompt”：“Q: 这是什么类型的练习? A:”，“prompt”：“Q: 告诉我正在做的练习: A:”，