arXiv:2611.32549v1 [cs.CV] 24 Nov 2026

MinPair：用于文本到动作生成细粒度控制的测量最小对

Lennart Schiweck Cristóbal Curio 认知系统组，
罗伊特林根大学，德国

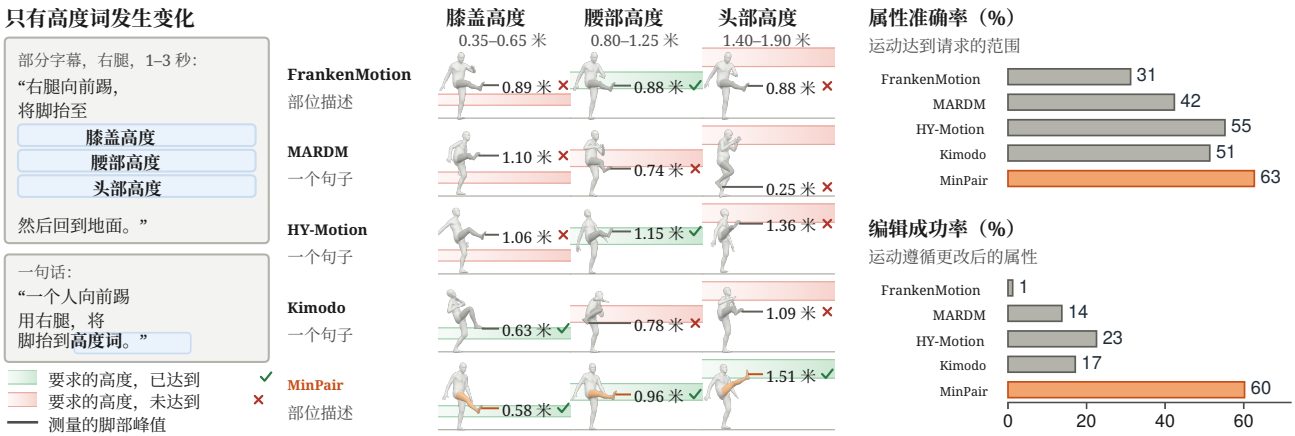


图 1。 **踢腿高度请求**。 仅高度词发生变化。如 FrankenMotion [21], MinPair 读取右腿的身体部位描述文本，而 MARDM [24], HY-Motion [37] 和 Kimodo [31] 读取一个句子。每个模型和高度下展示的动作，是从该提示生成的所有动作中，脚部峰值最接近中位数的动作；如果脚部峰值落在请求范围内，则以绿色显示请求范围，若超出范围则以红色显示。每个高度下方的范围是基准固定的目标范围。MARDM、HY-Motion 和 Kimodo 均是根据其存储的关节拟合得到的 SMPL-H。只有 MinPair 达到了全部三个范围。右侧为涵盖全部 22个目标任务的属性准确率，以及模板编辑得到遵循的比例；每项均使用其所在行的输入（第 4.1节）。

摘要

文本到动作生成在捕捉要执行什么动作方面取得了巨大进展，但在控制动作如何执行方面仍然有限。例如，当踢腿指令中只有高度词发生变化时，所测试的最先进方法中没有一种能分别达到膝盖、腰部和头部的高度。描述文本中的其他词通常会暗示某种属性，例如高度，因此训练损失和标准指标都不要求模型使用明确表达该属性的词。只有当这个词是区分其他方面完全相同的两条描述文本的唯一因素时，模型才必须使用它。我们提出了 *MinPair*，它从真实动作片段中创建此类描述文本对。首先，它使用基于规则的动作描述符测量每个片段的三维身体运动。随后，语言模型根据这些描述符撰写 *MinPair*描述，即带时间信息的身体部位描述，说明每个身体部位如何运动。同一组描述符还能指出同一动作的两个片段在哪些属性上存在差异。一个测量型最小对将一个片段的真实描述文本与其修改版本配对，修改版本中只有一个属性

取值为另一个片段中测得的值。随后，成对损失要求扩散去噪器根据真实描述文本重建片段的准确度高于根据修改版本重建的准确度。训练无需合成动作，推理过程也保持不变。我们在*22*个目标任务以及每次只改变一个属性的描述编辑上测试控制效果。仅使用 *MinPair*描述就能帮助动作达到目标，并改善 *M2M*、*R@3* 和 *FID* 等标准指标；然而，使用这些描述训练的模型仍会忽略大多数单词级编辑

其结果是，成对损失将遵循模板编辑的比例从 *11%* 提高到 *60%*，而标准指标没有进一步改善。借助身体部位描述文本，*MinPair* 达成的目标数量超过了使用多达 *99* 倍动作数据训练的最先进方法。

1. 引言

在文本到动作生成中， [8], 一条诸如踢至头部高度的请求，应当生成一个能够触及头部的踢腿动作。仅指定动作本身，并不能明确踢得有多高、多快以及向前多远

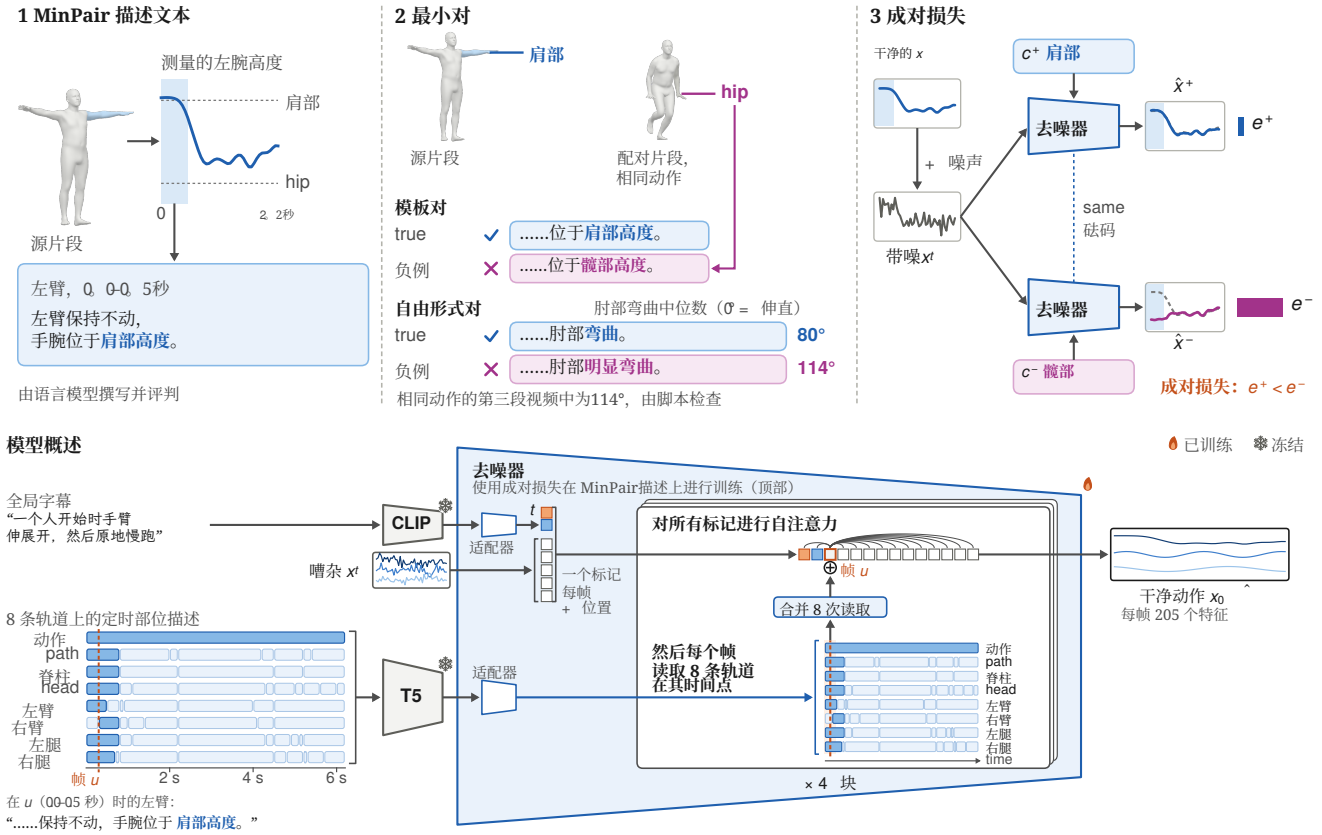


图2。 **方法概览**。 上部，训练。(1) 脚本测量真实片段中的量，此处为左腕高度。语言模型将这些量写入带时间标记的部位描述，随后通过进一步调用将其与测量数据进行比对。(2) 模板对复制真实描述文本，但将数值替换为相同动作的配对片段中的数值，此处为髋部高度。自由形式对会在测量差异（例如肘部弯曲）之后改写一小段文本，并根据测量数据进行检查。(3) 两个描述文本在相同的带噪动作上调节同一个去噪器。成对损失要求真实描述文本在其描述的时间窗口内（阴影部分）产生更小的误差。(3) 中的曲线仅用于示意。下部，第 3.1 节的部位轨迹去噪器，显示第 u 帧。只有文本编码器被冻结。

动作如何执行以及转向幅度多大。我们将这类执行属性称为属性。要求踢至头部高度的动画师，或要求向前走三步的机器人操作员，都期望动作完全按照这一要求执行。扩散模型已成为此任务的标准生成器 [11, 35]。FrankenMotion [21]为每个身体部位提供一条带时间标记的句子轨迹，我们称之为部位描述。HY-Motion [37]和 Kimodo [31]会对数据和参数进行缩放。然而，当踢腿请求中仅改变高度词时，这些最先进方法都无法达到所要求的三种高度（图1）。它们的训练过程并未专门突出属性词。文本标签 [8, 27] 很少说明动作进行的高度、速度或距离，因此由这些标签衍生出的部位描述 [21] 通常也不会说明这些信息。在描述文本明确说明这些属性的情况下，其他词语却会泄露答案。在根据测量数据撰写的描述文本中，线性分类器仅凭留出描述文本中的其余词语，便能在77%的情况下恢复被遮蔽的节奏词，而选择出现频率最高的节奏词时这一比例为41%。去噪损失以及标准指标 M2M、R@3 而 FID 都会对完整描述文本给予奖励，因此模型无需使用属性词，仅凭上下文也能取得成功。只有当某个词单独决定结果时，模型才会学会可靠地遵循该词。

运动语料库已经记录了每个动作的许多次动作执行，例如踢向膝盖和踢至头部高度。对这些动作进行测量，可以显示哪些词语适用于每次动作执行 [4, 22,40]，以及相同动作的哪些片段在某一属性上存在差异。测量最小对利用真实片段构建出这些案例。

我们介绍了 MinPair（图 2）。它使用基于规则的动作描述符测量每个训练片段。语言模型根据这些测量数据生成 *MinPair*描述，说明每个身体部位在各个阶段的运动方式。对于一个描述文本将手腕置于肩部高度的源片段，执行相同动作的配对片段则将手腕保持在髋部高度。我们复制源描述文本，只将这一数值改为配对片段中的数值。对于源片段而言，这份副本仅在一个属性上是错误的，因此充当其 负描述 [41]。新增词语的产生方式

有两种。在 模板对中，属性词沿着描述文本固定的词语尺度移动。在 自由形式对中，语言模型用一个不超过 4 个词的短语表达测量数据本身所体现的差异，我们再将其与测量数据进行核对。由于配对片段是真实片段，因此不会合成运动，负描述所描述的是一个确实发生的动作执行。如同在对比学习 [36], 中，成对损失 要求去噪器从真实描述文本而非负描述中，更准确地重建源动作的同一份带噪副本 [17, 19]。只有发生变化的词语才能造成这种差异。推理过程保持不变。

我们通过两个基准测试（第 4.1节）检验控制能力。 目标任务 要求明确的数值，例如踢到腰部高度，或行走三步。描述编辑会改变留出片段的 MinPair 描述中的一个属性，编辑方式可以是模板编辑，也可以是按照这两类最小对构建的自由形式编辑。使用 MinPair描述但不采用成对损失时，我们的去噪器仅能遵循 11% 的模板编辑。成对损失将这一比例提高到 60%，而 M2M、R@3 和 FID 不再进一步改善。使用身体部位描述文本时，MinPair 达成的目标数量超过了最先进方法 [21, 24, 31, 37], ，尽管 HY-Motion 和 Kimodo 使用多出 23 至 99 倍的动作数据进行训练（第 4.2节）。

我们的贡献如下 .

- 我们提出了测量最小对和对比成对损失，使描述文本中的单个属性能够控制文本到动作扩散模型。每个最小对将片段的真实描述文本与一个负描述相结合，后者从执行相同动作的另一个真实片段中选取一个属性值，体现了一种有针对性的监督信号，其作用可以超过规模效应（第 3.3、3.4和4.2节）。
- 我们提出了 MinPair描述，即根据对每个运动的测量数据撰写的带时间标记的部位描述。它们说明了每个阶段中各身体部位的属性。使用相同的网络，与以往的部位描述相比，它们提高了属性准确率，并提升了全部三个标准指标（第 3.2节和第 4.3节）。
- 我们提出了两个控制基准测试：目标任务和描述编辑，它们分别隔离单个属性词。结果表明，控制单个属性与拟合完整描述文本是两种不同的技能，而 M2M、R@3 和 FID 无法对此进行衡量（第 4.1节和第4.3节）。

2. 相关工作

文本到动作生成与细粒度文本。 HumanML3D [8] 成为文本到动作生成的标准基准，MDM [35]使去噪扩散成为常用生成器。近期工作改进了文本侧。FineMoGen [42]根据细粒度时空描述生成运动，STMC [26] 在采样过程中于多轨时间线上组合定时提示，FrankenMotion [21] 使每个身体部位以其自身的定时句子为条件。FineMotion [38] 与 KinMo [43] 描述单个身体部位或运动学组的运动。FineXtrol [32] 与 MATE [9] 通过分层或词级对比目标学习更精细的文本表示。其他工作则扩展生成器规模。HY-Motion 1.0 [37]在超过 3,000 小时的运动数据上训练了一个十亿参数的流匹配 Transformer，Kimodo [31] 在大规模光学动作捕捉语料库上训练扩散模型，OpenT2M [2] 借助详细的时间标注扩展数据，MotionMaster [14]统一了生成与编辑。PoseScript [4], 、PoseFix [5], 、Kinematic Phrases [22]与 MotionScript [40] 将测得的姿势和运动属性连接到语言，CoMo [13] 通过单个身体部位的姿势编码生成并编辑运动。LaMoGen [16] 利用拉班动作分析 [18]的力效与形状特质引导预训练扩散模型。MinPair 以测得的部位描述为基础，并增加一种训练信号，使生成器能够在同一动作内区分属性值。

困难负样本与后训练。 ARO [41] 使用具备组合意识的负向字幕来揭示视觉语言模型的弱点，而 SugarCrepe [12] 表明，构建不当的负样本包含语言层面的捷径。扩散分类器 [20]根据能够带来最低去噪误差的文本对图像进行分类。HardNeg- DiffusionITM [17]和 SoftREPA [19]以去噪误差为基础，采用对比目标进行训练；TDSM [6] 则通过对骨架动作施加三元组损失来识别这些动作。Fu- jiwara 等 [7]使用事件在时间上被打乱的负向文本训练动作语言模型。MinPair 将基于去噪的原理迁移到动作生成中，并从同一动作的两次真实执行之间测量得到的差异中挖掘负例。MoDiPO [25] 将扩散偏好优化适配到动作领域，而 Motion- RFT [34]则使用学习得到的动作奖励进行强化微调。OmniME [33]从正例和负例中学习动作编辑。MinPair 既不需要奖励模型，也不需要生成候选项。它对同一个真实运动的两个描述文本进行对比。

指令遵循评估。 SATO [3]测试语义相似文本下的稳定性。T2MBench [39] 评估模型在众多提示和模型上的指令遵循、物理合理性和精细控制能力，并为全身及成对身体部位设定数值目标。Motion-Fix [1] 通过源动作、目标动作和编辑文本三元组对动作编辑进行基准测试。我们的基准测试在相同采样噪声下，为单个身体部位添加目标范围，并设置成对的描述编辑，从而隔离单个属性变化所产生的影响。

3. 方法

MinPair 测量每个训练动作，将测量结果写入带时间标记的部位描述，并将这些描述与负描述配对；每条负描述仅改变一个测量属性（图 2）。

3.1. 部分轨迹扩散模型

与 FrankenMotion [21], 一样，一个描述文本由覆盖整个动作片段的全局字幕和八条轨道组成，分别对应动作、路径（根轨迹）、脊柱、头部、左臂、右臂、左腿和右腿。我们将带有时间跨度的一句话称为一个片段。轨道无需共享边界。

一个动作片段 $x \in \mathbb{R}^{T\times D}$ 包含 T 帧，每帧具有 $D = 205$个特征，采用 FrankenMo- tion 的动作表示。其文本条件 c 就是这样的部分字幕。MinPair 是一个扩散模型 [11, 35]。前向过程将 x 与高斯噪声混合，

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad (1)$$

对于采用余弦调度的扩散步 $t \in \{1, \ldots, 100\}$ $\bar{\alpha}_t$。我们将 x 称为干净动作，将 x_t 称为带噪动作。去噪器 $f_\theta(x_t, t, c)$ 根据带噪动作预测干净动作。对于帧索引集合 W（例如一个片段的时间跨度），其在描述文本c 条件下的重建误差为

$$e_W(c) = \frac{1}{|W|} \sum_{u\in W} \left\| f_\theta(x_t, t, c)_u - x_u \right\|_\omega^2, \qquad (2)$$

其中 ω 砝码对动作特征进行加权，与 FrankenMotion 中的方式相同。该误差涵盖 W中各帧的所有特征，而不仅仅是一个身体部位的关节。去噪损失 $\mathcal{L}_{\text{denoise}} = \mathbb{E}_{(x,c),\ t,\ \epsilon}[e_M(c)]$ 在集合 M中对片段的所有有效帧的该误差取平均，其中 t均匀抽取。

部位轨迹去噪器。 该去噪器是一个基于帧的 Transfor mer（图 2，底部）。每一帧 x_t 都变为一个词元。另有两个词元分别携带扩散步 t 和全局字幕，冻结的 CLIP [29]对其进行编码。冻结的 T5-Large [30]对每个片段进行编码。已训练的适配器将两种文本特征映射到去噪器中。四个块中的每一个首先对所有词元执行自注意力，随后经过一个前馈层。然后，每一帧通过交叉注意力读取八条轨迹中在该帧处处于活动状态的片段。一个小型多层感知机将这八次读取的结果与标记活动轨迹的标志进行融合，并将结果加到帧词元上。线性头预测干净动作。

3.2. MinPair描述

MinPair 根据运动本身为几乎每条描述文本编写七条身体部位轨道。全局字幕和动作轨道仍沿用 FrankenMotion [21], 的内容；FrankenMotion 由语言模型根据现有文本标签合并而成，这些标签主要来自 BABEL [27] 和 HumanML3D [8]。

首先，我们使用与每段录制内容相对应的身体模型和体型，计算每个片段的关节位置。基于规则的运动描述符会将这些位置概括成一段简短文本，每隔 0.75 到 1 秒说明哪些四肢在移动、手和脚抬起的高度、关节弯曲的程度，以及骨盆移动和转向的距离。Qwen3.6-27B [28]读取这份摘要，并结合一张记录各身体区域运动幅度的表格，将片段切分为全身运动的各个阶段。在此步骤中，它看不到任何文本标签。随后，对于每个阶段和身体部位，我们将测量值转换为固定词语尺度上的等级词，例如用于表示高度的膝盖高度、肩部高度 或头顶高度 ，以及用于表示节奏的静止、缓慢、中等或快速 。移动的距离转换为步数，而转向则转换为四舍五入到 5° 的角度。

语言模型的第二次调用根据摘要、阶段和这些等级词撰写句子。提示要求每个身体部位和时间跨度各写一句，并为没有间隔的轨道撰写描述。它要求描述一个片段中诸如行走之类的重复动作，并说明其次数；标明每条手臂和腿的侧别；给出相对于身体的高度和方向，例如跨过身体。除转向角和次数外，只允许使用数字。该调用还会将片段的 BABEL 和 HumanML3D 标签作为假设一并参考。只有当这些标签得到该身体部位和时间跨度的测量数据支持时，才能据此命名活动。第三次调用会根据摘要判断每个句子。其判定结果会被保存以供检查，但不会删除任何句子。图 2 中的描述文本 (1) 就是通过这种方式根据手腕高度曲线生成的。

MinPair描述涵盖12,726个训练片段中的450,179个片段，每个片段的中位数为29个。相同的流程还为1,583个测试片段撰写MinPair测试描述文本，这些片段在训练过程中不会被任何模型看到。

3.3. 测量最小对

最小对由同一片段的两条描述文本组成，二者仅在某一属性的用词上有所不同。负向描述文本 复制源片段的真实描述文本，只将这些词替换为配对片段中的取值；配对片段是相同动作的另一次真实动作执行。其他所有词语、片段和轨道均保持不变。配对片段仅提供这一取值，不会被重建。模板对和自由形式对的区别在于新词语的产生方式。

模板对。 在模板对中，属性词沿着描述文本的固定词语尺度移动。我们寻找两个关于相同动作的训练视频片段，其描述文本在同一轨迹上陈述某一属性的不同取值，例如腿部轨迹中最高的高度词、节奏词、转向角或步数。只有当该片段的描述符能够确认某个取值时，该取值才计入。随后，用搭档片段中的词替换源词，例如髋部高度 替换为 肩部高度 （图2（2））。第二次搜索则直接比较两个

关于相同动作的片段，比较 51 个测量量。其中，差异必须超过测量噪声和可见性阈值，例如超过该人物身体半个身高等级或 30° 的转向角。模板对涵盖高度、节奏、转向、次数和侧别，以及四种姿态属性：躯干倾斜、躯干扭转、头部速度和膝部弯曲。

自由形式对。 在自由形式对中，语言模型会用一个不超过 4 个词的片段来描述测量差异本身。这也涵盖了描述文本的词语尺度未提及的差异，例如肘部弯曲或站姿宽度。对于每个训练视频片段，我们选取最多 2 个相同动作的配对对象，并对这两个片段的 61 项测量量进行比较。语言模型会接收带有时间跨度的源描述文本，以及每项可见差异对应的两个片段的描述符值及其文字表述。对于每一对，模型会返回一个真实句子和一个负例句子；负例句子会将其中一个片段替换为配对对象的对应值。真实句子可以将源值补充到未说明该值的片段中，随后该片段也会被替换到去噪损失中。我们会检查句子是否对应差异发生的轨迹和时间跨度，并在能够从文字表述中读取两个值时，确认修改后的词语是否指明了从源值到配对对象值的变化。被拒绝的配对会连同拒绝原因一起返回给语言模型，再处理一次。提议的配对中有 72.4% 通过了检查。

训练使用了 190,141 条负向字幕，涵盖 12,822 个训练片段中的 11,824 个，其中 48,155 条来自自由形式对。

3.4. 成对损失

对于带有受影响时间窗口 W_j 的负描述 c_j^-，我们比较同一段带噪运动在两条描述文本下的重建误差，

$$z_j = \frac{e_{W_j}(c^+) - e_{W_j}(c_j^-)}{\tau \max\{\mathrm{sg}[e_{W_j}(c^+)], 10^{-8}\}}, \tag{3}$$

其中，c^+ 是真实描述文本或其释义，sg 表示停止梯度，$\tau = 0.1$。对于包含 J 条负描述的片段，成对损失为

$$\mathcal{L}_{\mathrm{pair}} = \log\Big(1 + \sum_{j=1}^{J} \exp z_j\Big). \tag{4}$$

除以停止梯度的正例误差后，差值变为相对值，因此误差较大的片段和扩散步不会主导损失。

与对比学习的关系。 将描述文本的得分定义为 $s_W(c) = -e_W(c)/(\tau \mathrm{sg}[e_W(c^+)])$。则 $z_j = s_{W_j}(c_j^-) - s_{W_j}(c^+)$。对于共享一个窗口的负例 W，公式（4）正是 InfoNCE损失 [36]

使用这些得分来处理一个正例和 J 个负例描述文本。因此，训练实质上是针对描述文本的对比学习，其中去噪器充当评分函数，正如图像领域中的 HardNeg-DiffusionITM [17] 和 SoftREPA [19]。MinPair 从测量数据中挖掘负例，并将每次比较限制在发生变化的片段所在的窗口内。

目标函数将成对损失加到去噪损失上，

$$\mathcal{L} = \mathcal{L}_{\mathrm{denoise}} + \lambda\, \mathbb{E}_{\mathrm{sel}}\big[\mathcal{L}_{\mathrm{pair}}\big], \quad \lambda = 0.05, \tag{5}$$

其中，成对损失在一个批次中选定的片段上取平均，而去噪损失涵盖所有片段。

3.5. 训练

我们使用公式（5）从随机初始化开始训练去噪器，共进行 1,500 个周期（批次大小为 64 时共 300,000 步）。我们使用 AdamW，学习率保持恒定，为 10^{-4}。对砝码采用指数移动平均（衰减率为 0.9995），得到用于评估的模型。带有负向字幕的片段在噪声较大的步数 $t \geq 50$中始终有资格参与成对损失；在下方噪声较小的步数中，其参与概率为 0.2。每个批次最多选择其中 64 个片段中的 21 个，优先选择带有负例的片段。对于模板对，真实片段的释义（例如将肩部高度 替换为 肩部高度 ）以 0.5 的概率重新表述，而负例保持原有措辞，因此新的措辞不会成为负例的信号 [15]。没有配对片段的片段最多接收 2 次词汇交换，即用同一尺度上的另一个词替换所述数值。另有两个负例以 0.5 的概率分别加入。一个会在所有文本和轨迹中交换左与右，另一个完全不包含文本。对于脚或手腕位置较高且较为罕见的片段，训练时会以最高 2.95 倍的频率抽取。使用一块 H100 图形处理器大约需要 21 小时。

4. 实验

4.1. 设置

数据。 我们使用 FrankenMotion [21] 的 AMASS [23]子集，该子集划分为 12,845 个训练片段（总时长 30 小时）、1,651 个验证片段和 1,601 个测试片段。运动的帧率为 20 fps。我们保留时长在 1 到 20 秒之间的片段，最终得到 12,822 个训练片段和 1,596 个测试片段。

基线。 我们将 MinPair 与四个已发表的生成器进行比较。FrankenMotion [21] 是与 MinPair 共享输入格式的部位轨迹模型。我们使用其已发布的检查点，该检查点已训练于相同的 30 小时 AMASS 划分。两个部件轨迹模型使用相同的采样器，且不使用无分类器引导 [10]。MARDM [24]，HY-Motion 1.0 [37] 和 Kimodo [31] 读取单一文本。MARDM 是在 HumanML3D [8]上已训练的掩码自回归扩散模型。HY-Motion 和 Kimodo 分别已训练于超过 3,000 小时和 700 小时的运动。我们使用各自的采样设置运行官方砝码

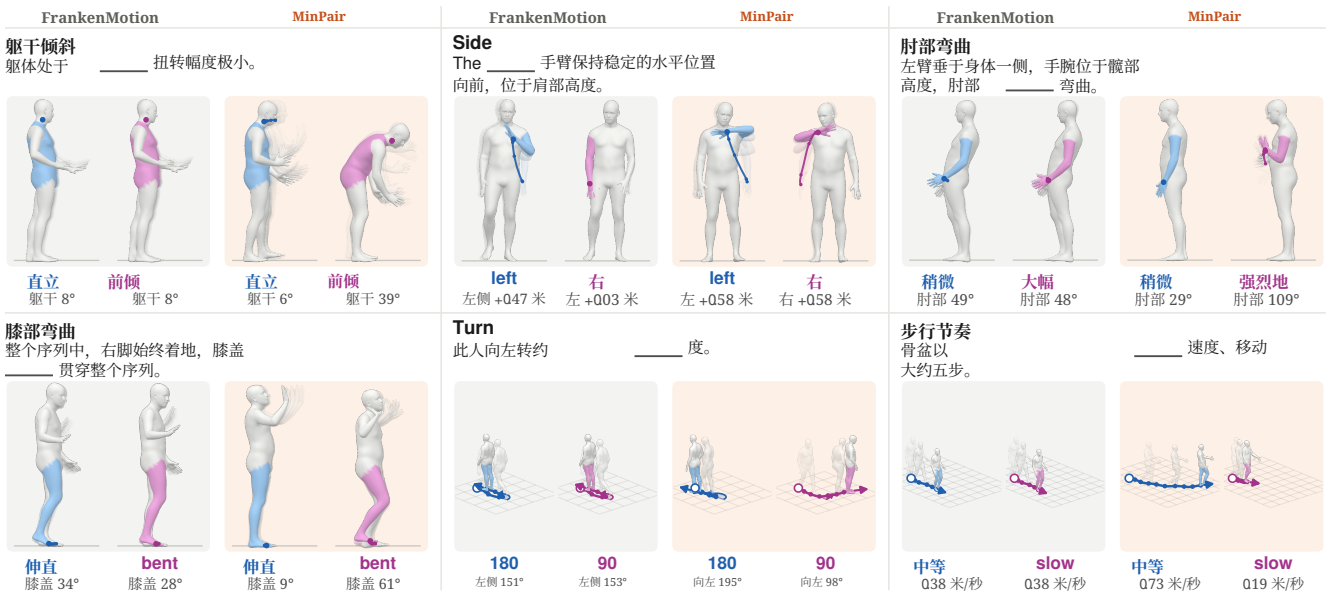


图3. **描述编辑**。 每个区块都在相同的采样噪声下，对一个留出测试片段的部位描述中的单个属性进行更改（从蓝色到洋红色）。肘部和膝盖采用自由形式编辑，其余则采用模板编辑。转向和节奏展示了骨盆路径，数字表示测量值。在这六类的所有测试编辑中，MinPair的成功率分别为83%、49%、72%、53%、48%和43%，而FrankenMotion最高仅为3%。每个区块都展示了一次干净的编辑，其要求的改变量较大，生成的运动也接近要求值。

发布版本，并将关节重采样至 20 fps。MARDM 最多生成 9.8秒，HY-Motion 最多生成 12秒。Kimodo 在一次提示中按其长度生成每个片段。

输入。 每个模型都会读取其专门设计的输入。FrankenMotion和MinPair读取部位描述，因此，要求会按照MinPair描述的句式写入相关身体部位的定时片段中。MARDM、HY-Motion和Kimodo读取单条文本，因此会将整个要求写成一个句子。

目标任务。 第一个基准测试模型能否在22项任务中达到指定值，例如踢腿高度、手臂高度、转向角、行走距离、深蹲次数、躯干倾斜和膝部弯曲。每项任务设有三个指定等级，例如膝盖、腰部或头部高度。每个等级都以两种输入形式编写，并使用各模型共用的20个随机种子生成结果。脚本 m_q 测量 q 运动中该任务的量值，例如脚部峰值高度，所生成的运动为 $G(c_\ell,\ \xi)$，由 c_ℓ 等级为 ℓ 的提示词生成，并加入 ξ采样噪声。若样本的测量值落入 R_ℓ 该等级的目标范围内，则判定为命中，

$$\mathrm{Hit}_A = \mathbf{1}\big[m_q(G(c_\ell,\xi)) \in R_\ell\big]. \tag{6}$$

属性准确率是目标任务中命中次数的占比，并按输入的任务和级别单元格取平均值。

描述编辑。 第二个基准测试单个词语的变化是否会改变运动。描述编辑只修改文本。它会替换留出测试片段的MinPair描述中的一个属性值。随后，模型使用相同的噪声 ξ，分别根据未编辑的描述文本 c 和编辑后的描述文本 c' 生成一个运动，因此两种运动之间的差异完全来自文本。编辑有两个来源。**模板编辑** 由脚本生成。根据模板对的规则，它们在 1,574 个片段中进行了 3,292 次编辑，修改手臂高度、腿部高度、节奏、转向、次数或侧别；此外，还进行了 2,577 次编辑，修改躯干倾斜、躯干扭转、头部速度或膝部弯曲。侧别编辑会镜像左右轨迹。**自由形式编辑** 来自自由形式对的语言模型，该模型在 4,007 次编辑中修改 18 种属性中的一种。脚本会像检查自由形式对那样检查这些编辑。设 $\Delta_g = m_q(G(c',\ \xi)) - m_q(G(c,\ \xi))$ 分别表示编辑后数量的变化 q 以及 Δ_r 真实片段中两个属性值之间观察到的变化。若一项编辑遵循，则

$$\mathrm{Hit}_B = \mathbf{1}\big[\mathrm{sign}(\Delta_r)\,\Delta_g \geq \tfrac{1}{2}|\Delta_r|\big]. \tag{7}$$

运动必须沿请求方向移动，且至少达到实际变化的一半。对于四种姿态属性和自由形式编辑 Δ_r，测量差异是相对于一位训练伙伴的差异；对于姿态编辑，按片段计算；对于自由形式编辑，则按完整视频片段计算。编辑成功率是针对某一源所遵循编辑的占比，其中每次编辑均计数一次，因为某些类型仅包含少量编辑。图3展示了

表1。**与已发布的生成器进行比较**。每个模型读取其自身的输入格式。单句输入不包含附加编辑。成功率和 R@3 以百分比表示，每列的最佳值以粗体显示。

模型	输入	数据（小时）	目标任务 ↑	模板编辑 ↑	自由形式编辑 ↑	All 编辑 ↑	M2M ↑	R@3 ↑	FID ↓
MARDM [24]	一句话	<29	42.3	13.7	7.0	10.8	0.694	81.4	0.117
HY-Motion 1.0 [37]	一句话	>3,000	55.2	22.5	16.7	20.0	0.754	**89.0**	0.059
Kimodo [31]	一句话	700	51.4	17.1	13.4	15.5	0.618	56.7	0.431
FrankenMotion [21]	部位描述	30	31.2	1.2	0.2	0.8	0.745	86.0	0.057
MinPair（我们的方法）	部位描述	30	**62.7**	**60.2**	**55.0**	**58.1**	**0.830**	85.9	**0.038**

表2. **字幕和配对损失的影响**。 每一行都对我们的去噪器进行训练，第一行也不例外；第一行使用 FrankenMotion 的描述文本和 FrankenMotion 的训练过程。每一行都只改变上一行的一项设置。最后一行是 MinPair。成功率和 R@3 均以百分比表示。

我们的去噪器使用以下内容进行训练			目标	模板	自由形式	All			
字幕	训练	成对损失	任务 ↑	编辑 ↑	编辑 ↑	编辑 ↑	M2M ↑	R@3 ↑	FID ↓
FrankenMotion	FrankenMotion	–	27.5	1.4	0.1	0.9	0.729	81.6	0.089
MinPair	FrankenMotion	–	50.0	11.4	6.0	9.2	0.819	86.9	0.061
MinPair	MinPair	–	44.5	10.5	4.7	8.2	**0.837**	**89.0**	**0.033**
MinPair	MinPair	模板对	60.7	**61.4**	27.6	47.7	0.834	88.0	0.034
MinPair	MinPair	自由形式对	57.0	36.7	**59.8**	46.0	0.827	86.6	0.036
MinPair	MinPair	both	**62.7**	60.2	55.0	**58.1**	0.830	85.9	0.038

躯干倾斜、侧别、转向和步行节奏的模板编辑，以及肘部和膝盖弯曲的自由形式编辑。

质量与对齐。 我们还报告 FrankenMotion评估器的三个指标 [21] ，这些指标是在包含全局字幕的 1,562 个测试片段上计算的，并对三次生成运行取平均值。M2M 是生成运动与参考运动嵌入之间经重新缩放的余弦 $(1+\cos)/2$ 相似度。R@3 是前 3 名文本检索准确率。FID 是嵌入分布之间的 Fréchet距离。每个运动均根据真实测试片段的完整描述文本生成，并与该片段进行比较。在 FrankenMotion 描述文本上训练的模型接收 FrankenMotion 测试描述文本；在 MinPair描述上训练的模型则接收同一片段中留出的 MinPair描述，这些描述由第 3.2 节的流水线根据每个测试片段的测量数据生成。评估器使用 FrankenMotion 描述文本进行训练，并针对这些描述文本计算所有文本指标。因此，我们省略其动作到文本相似度 M2T，并将 M2M 用作主要质量指标，因为它比较的是运动与运动。三个单文本模型接收全局字幕和片段长度。

对照实验。 为了将描述文本与网络区分开来，我们还使用 FrankenMotion 的训练过程和单独的去噪损失训练去噪器：一次使用 FrankenMotion 的部位描述，另一次使用 MinPair描述。两次训练共享片段、数据划分、设置和训练轮数，唯一的区别在于描述文本。

4.2. 与已发布的生成器进行比较

表 1为每个模型提供其专属输入。使用部位描述时，MinPair 达到 62.7% 的属性准确率，并遵循 60.2% 的模板编辑和 55.0% 的自由形式编辑。这一表现超过了所有已发布的生成器，尽管 HY-Motion 和 Kimodo 使用的动作数据多出 23 至 99 倍。当两个部位轨迹模型都改为读取 MARDM 的一句话全局字幕，以及 HY-Motion 和 Kimodo 在全局字幕和全部八条轨道中的输入时，FrankenMotion 遵循模板编辑和自由形式编辑的比例分别为 10.0% 和 5.0%。MinPair 仍分别达到 50.7% 和 50.0%，超过所有已发布的生成器。在最长为 9.8 秒、且处于每个模型长度限制以内的测试片段上，MinPair 遵循 58.2% 的编辑，而已发布生成器的比例为 0.9% 至 19.7%。

4.3. 字幕和配对损失的影响

描述文本与成对损失。 表 2每一行训练相同的去噪器，并且每行只改变一项设置。该去噪器在 FrankenMotion 的描述文本上训练，其表现并不优于 FrankenMotion 本身（表1）。其属性准确率为 27.5%，低于 31.2%；对模板编辑的遵循率为 1.4%，低于 1.2%。因此，该网络无法解释后续的性能提升。仅使用 MinPair描述，就能将属性准确率提升至 50.0%，将模板编辑成功率提升至 11.4%。第 3.5 节的训练若不使用成对损失，则不会带来任何控制能力（44.5% 和 10.5%）。随后，结合模板对的成对损失将模板编辑成功率提升至 61.4%，并将属性准确率提升至 60.7%。因此，MinPair描述教会模型理解完整描述文本所描述的动作——

表 3。**成对损失训练的消融实验**。下方每一行的 MinPair 均对其训练的一个部分进行了修改，并从头开始训练。成功率和 R@3 均以百分比表示。

模型	目标任务 ↑	模板编辑 ↑	自由形式编辑 ↑	All 编辑 ↑	M2M ↑	R@3 ↑	FID ↓
MinPair	**62.7**	**60.2**	55.0	**58.1**	0.830	85.9	0.038
仅在高噪声下使用成对损失	57.6	49.4	44.3	47.3	0.829	87.0	**0.036**
在每个步骤使用成对损失	57.7	59.4	**56.0**	58.0	**0.831**	**87.3**	**0.036**
不含词汇交换	58.3	58.5	53.9	56.6	0.827	85.9	0.039
不含稀有位置过采样	61.9	57.9	53.0	55.9	0.830	86.9	0.038

执行。成对损失则使其能够对单个词语的变化作出响应。

成对来源。 表 2 的最后三行仅在其负描述上有所不同。每个来源在其所涵盖的编辑上表现最佳。模板对覆盖了 61.4% 的模板编辑，但仅覆盖了 27.6% 的自由形式编辑。仅使用自由形式对时，覆盖了 59.8% 的自由形式编辑，但仅覆盖了 36.7% 的模板编辑。MinPair 结合了这两个来源，覆盖了 60.2% 的模板编辑和 55.0% 的自由形式编辑。因而，每个来源都在自身的编辑上领先，而在两个来源上训练的 MinPair 则以 62.7% 的属性准确率领先，并且在全部编辑上达到 58.1%。

标准指标。 标准指标遵循描述文本，但不遵循成对损失。使用 MinPair描述时，序列 M2M 从 0.729 提升至 0.819；通过对 Sec. 的训练则提升至 0.837。3.5。使用成对损失时，该指标为 0.830，尽管模板编辑成功率从 10.5% 提升至 60.2%。原因在于，这些指标测试的是完整描述文本。完整描述文本会冗余地描述一次动作执行。当我们遮蔽留出描述文本中的手臂高度、腿部高度或节奏词时，在训练描述文本上拟合的线性分类器可以根据剩余词语，在 86%、73% 和 77% 的情况下恢复这些信息；而最常见值的比例分别为 63%、60% 和 41%。因此，模型无需使用属性词，也能根据上下文生成正确的动作执行。事实上，由未编辑测试描述文本生成的运动，无论是否使用成对损失，与真实片段达成一致的频率相同：在 8 个测量量上，该比例均为 87.8%。

4.4. 成对损失训练的消融实验

表3 展示了每一行对成对损失训练中的一个部分进行更改，并从头开始训练。在仅对高噪声下的步数使用成对损失时，步数为 $t \geq 50$，模型对模板编辑和自由形式编辑的遵循率分别为49.4%和44.3%，而不是60.2%和55.0%。因此，噪声较低的步数也需要使用成对损失。在每个步数都使用成对损失时，模型对模板编辑和自由形式编辑的遵循率分别为59.4%和56.0%。不进行词汇交换时，模板编辑成功率为58.5%；不进行稀有位置过采样、让所有动作片段以相同频率抽取时，模板编辑成功率为57.9%。

5. 结论

我们提出了MinPair，它根据真实动作片段的测量数据，为可控文本到动作生成构建监督信号。同一套基于规则的动作描述符生成MinPair描述以及成对损失所需的测量最小对。在使用相同网络的情况下，与FrankenMotion描述相比，MinPair描述提高了属性准确率、M2M、R@3和FID，但去噪器对模板编辑的遵循率仍仅为11%。成对损失将这一比例提高到60%，但三个标准指标没有进一步改善。因此，控制单个属性与拟合完整描述文本是两种不同的技能。只有当一个词单独区分两个其他方面完全相同的描述文本时，模型才能可靠地习得这种技能。使用身体部位描述文本时，MinPair所达到的目标数量也超过了使用多得多的动作进行训练的最先进方法，这表明针对性监督可以胜过规模。

影响。 MinPair 将描述文本中的词语转化为控制信号。动画师只需修改一个词，就能改变动作的高度、速度或幅度。当机器人模仿生成的运动，且其操作员用语言描述动作执行情况时，同样的控制方式也适用。这种监督信号既不需要新的录制数据，也不需要新的人工标注，只需进行计算，因此可以随语料库规模扩展。成对损失只在训练中增加一项，不会改变推理过程，因此现有的扩散生成器也可以采用。两个基准测试衡量了标准指标无法捕捉的控制能力，并适用于任何文本到运动模型。测量得到的最小对并不局限于运动领域。只要一个词背后的属性可以被测量，就可以构建相应的最小对。

局限性。 MinPair 涵盖了可从关节位置中测量、并可通过词语尺度或短语表达的属性。风格、情感以及与物体的接触不在此范围内。我们在一个语料库上使用一个部位轨迹去噪器训练了成对损失，因此其他生成器系列仍有待测试。此外，将成对损失应用于批次中的所有片段仍未得到探索。

未来工作。 到目前为止，测量词语通过身体部位轨迹发挥作用。下一步是将它们融入全局描述中，使单个句子也能引导动作执行。只要每一对都能依据测量数据进行核验，还可以增加描述动作发生时间，或描述多个身体部位如何协同运动的配对。监督信号的创建也可以变得更加智能化：利用语言模型智能体发现生成器仍然遗漏的内容，并为其添加相应的测量数据和配对。

致谢

本研究由欧盟区域发展基金（ERDF）通过“AIDA”项目资助（拨款编号 FEIH 2476184）。文中所表达的观点和意见仅代表作者，不一定反映欧盟的观点。欧盟和拨款机构均不对此承担责任。

写作和编程助手（Anthropic 的 Claude、OpenAI 的 ChatGPT）协助起草和修订本稿件、编写和运行实验代码，以及准备表格和图。作者核实了每个数字、论断和参考文献，并对内容承担全部责任。

参考文献

[1] Nikos Athanasiou、Alpár Cseke、Markos Diomataris、Michael J. Black 和 Gül Varol。MotionFix：文本驱动的三维人体动作编辑。发表于 *SIGGRAPH Asia 2024* 会议论文集，第 1–11 页，2024 年。 3[2] Bin Cao 、 Sipeng Zheng 、 Hao Luo、Boyuan Li、Jing Liu 和 Zongqing Lu。OpenT2M：采用开源、大规模、高质量数据的简洁动作生成。发表于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 30640–30649 页，2026年。 3[3] Wenshuo Chen、Hongru Xiao、Erhang Zhang、Lijie Hu、Lei Wang、Mengyuan Liu 和 Chen Chen。SATO：稳定的文本到动作框架。发表于第 *32* 届 *ACM* 多媒体国际会议论文集，第 6989–6997 页，2024 年 3– 。[4] Ginger Delmas、Philippe Weinzaepfel、Thomas Lucas、Francesc Moreno-Noguer 和 Grégory Rogez。PoseScript：从自然语言生成三维人体姿势。发表于*ECCV*，2022年。 2，3[5] Ginger Delmas、Philippe Weinzaepfel、Francesc Moreno-Noguer 和 Grégory Rogez。PoseFix：使用自然语言校正三维人体姿势。发表于 *ICCV*，第 15018–15028 页，2023年3– 。 [6] Jeonghyeok Do 和 Munchurl Kim。弥合骨架与文本之间的模态差距：扩散驱动的模态对齐用于零样本骨架动作识别。发表于 *IEEE/CVF* 国际计算机视觉会议（*ICCV*）论文集，第 12757–12768 页，2025年。 3[7] Kent Fujiwara、Mikihiro Tanaka 和 Qing Yu。用于动作语言模型时间定位的时间顺序准确检索。发表于*ECCV*，第 323–339 页，2024年。 3

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 从文本生成多样且自然的 3D 5161、2022 5152 1、2、3、4、5D 人体动作。载于 CVPR，第 – 页。[9] Ke Han, Yueming Lyu, and Nicu Sebe. 用于文本到动作生成的运动对齐词嵌入。载于国际学习表征会议，第 28276–28289 页，2026年。3[10] Jonathan Ho and Tim Salimans. 无分类器扩散引导。arXiv:2207.12598，2022年。5[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 去噪扩散概率模型。载于 NeurIPS，第 6840 6851 页，2020年 2、4 –。[12] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe：修复易被操纵的视觉语言组合性基准测试。载于 NeurIPS，第 31096–31116 页，2023年。3[13] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo：通过语言引导的姿势代码编辑实现可控动作生成。载于 ECCV，第 180–196 页，2024年。3[14] Nan Jiang, Yunhao Li, Lexi Pang, Zimo He, Siyuan Huang, and Yixin Zhu. MotionMaster：可泛化的文本驱动动作生成与编辑。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 30629–30639 页，2026年。3[15] Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 关于视觉语言组合性的困难正样本真相。载于 ECCV，第 37–54 页，2024年。5[16] Heechang Kim, Gwanghyun Kim, and Se Young Chun. LaMoGen：用于文本到动作生成的拉班运动引导扩散。arXiv:2509.24469，2025年。3[17] Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. 扩散模型是视觉与语言推理器吗？载于《神经信息处理系统进展》，第 8385–8405 页，2023年。3、5[18] Rudolf Laban and Lisa Ullmann.《动作掌握》。Macdonald & Evans，伦敦，第 3 版，1971年。3[19] Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. 在扩散模型中，将文本与图像对齐比你想象的更容易。载于 NeurIPS，2025年。3、5[20] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. 你的扩散模型其实是一个零样本分类器。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 2206–2217 页，2023年。3[21] Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion：人体动作的部件级生成与组合。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 16420–16431 页，2026年。1、2、3、4、5、7[22] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. 通过运动学短语弥合人体动作与动作语义之间的差距。载于 ECCV，第 223–240 页，2024年。2、3

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS：作为表面形状的动作捕捉档案。载于 ICCV，第 5442–5451 页，2019年。5[24] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 重新思考文本驱动的人体动作生成中的扩散：冗余表示、评估与掩码自回归。载于计算机视觉与模式识别会议（CVPR）论文集，第 27859–27871 页，2025年。1、3、5、7[25] Massimiliano Pappa, Luca Collorone, Giovanni Ficarra, Indro Spinelli, and Fabio Galasso. MoDiPO：通过 AI 反馈驱动的直接偏好优化实现文本到动作对齐。《Frontiers in Computer Science》，8:1707808，2026年。3[26] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. 用于文本驱动 3D 人体动作生成的多轨时间线控制。载于 CVPR研讨会，第 1911–1921 页，2024年。3[27] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL：带英语标签的身体、动作与行为。载于 CVPR，第 722–731 页，2021年。2、4[28] Qwen团队。Qwen3.6-27B。https://huggingface.co/Qwen/Qwen3.6-27B，2026年。模型卡片，访问于2026年10月22日。4[29]Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 从自然语言监督中学习可迁移视觉模型。载于第 38 届国际机器学习会议（ICML）论文集，第 8748–8763 页，2021年。4[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》，21(140):1–67，2020年。4[31] Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo：扩展可控人体动作生成。arXiv:2603.15546，2026年。1、2、3、5、7[32] Keming Shen, Bizhu Wu, Junliang Chen, Xiaoqin Wang, and Linlin Shen. FineXtrol：通过细粒度文本实现可控动作生成。《AAAI人工智能会议论文集》，40(11):8905–8913，2026年。3[33] Zhenwu Shi, Jingyu Gong, Peiwei Wang, Xingzan Wang, Tianwen Qian, Wenxi Li, Yuan Fang, Jiao Xie, Lizhuang Ma, and Shaohui Lin. 全监督动作编辑：通过正负学习平衡变化与不变性。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 30595 30606 页，2026年 3 –。[34] Xiaofeng Tan、Wanjiang Weng、Hongsong Wang、Fang Zhao、Xin Geng 和 Liang Wang。MotionRFT：用于文本到动作生成的统一强化微调。arXiv:2603.27185，2026年。3[35] Guy Tevet、Sigal Raab、Brian Gordon、Yonatan Shafir、Daniel Cohen-Or 和 Amit H. Bermano。人体动作扩散模型。发表于 ICLR，2023年。2, 3, 4[36] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。arXiv:1807.03748，2018年。3, 5[37] Yuxin Wen、Qing Shuai、Di Kang、Jing Li、Cheng Wen、Yue Qian、Ningxin Jiao、Changhai Chen、Weijie Chen、Yi- ran Wang、Jinkun Guo、Dongyue An、Han Liu、Yanyu Tong、Chao Zhang、Qing Guo、Juan Chen、Qiao Zhang、Youyi Zhang、Zihao Yao、Cheng Zhang、Hong Duan、Xiaoping Wu、Qi Chen、Fei Cheng、Liang Dong、Peng He、Hao Zhang、Ji- axin Lin、Chao Zhang、Zhongyi Fan、Yifan Li、Zhichao Hu、Yuhong Liu、Linus、Jie Jiang、Xiaolong Li 和 Linchao Bao。HY-Motion 1.0：扩展流匹配模型以实现文本到动作生成。arXiv:2512.23464，2025年。1, 2, 3, 5, 7[38] Bizhu Wu、Jinheng Xie、Meidan Ding、Zhe Kong、Jianfeng Ren、Ruibin Bai、Rong Qu 和 Linlin Shen。FineMotion：一个包含空间与时间标注、用于细粒度动作生成和编辑的数据集与基准。发表于 IEEE/CVF国际计算机视觉会议论文集，第 13837–13846 页，2025年。3[39] Bin Yang、Rong Ou、Weisheng Xu、Jiaqi Xiong、Xintao Li、Taowen Wang、Luyu Zhu、Yuetong Fang、Xu Jiang、Ji ng Tan 和 Renjing Xu。T2MBench：在多样化语言指令下评估文本到动作生成。发表于第15届国际自然语言处理联合会议暨第5届亚太地区计算语言学协会会议论文集（第1卷：长篇论文），第 1622–1662 页。计算语言学协会，2026年。3[40] Payam Jome Yazdian、Rachel Lagasse、Hamid Mohammadi、Eric Liu、Li Cheng 和 Angelica Lim。MotionScript：用于表达性三维人体动作的自然语言描述。发表于 IEEE/RSJ智能机器人与系统国际会议（IROS），第 21574–21581 页，2025年。2, 3[41] Mert Yuksekgonul、Federico Bianchi、Pratyusha Kalluri、Dan Jurafsky 和 James Zou。视觉-语言模型何时以及为何表现得像词袋？该如何应对？发表于 ICLR，2023年。2, 3[42] Mingyuan Zhang、Huirong Li、Zhongang Cai、Jiawei Ren、Lei Yang 和 Ziwei Liu。FineMoGen：细粒度时空动作生成与编辑。发表于《神经信息处理系统进展》，第 13981 13992 页，2023 3 –。[43] Pengfei Zhang、Pinxin Liu、Pablo Garrido、Hyeongwoo Kim 和 Bindita Chaudhuri。KinMo：运动学感知的人体动作理解与生成。发表于 IEEE/CVF国际计算机视觉会议论文集（ICCV），第 11187–11197 页，2025年。3