朝向一致的长期姿势生成

Yayuan Li Filippos Bellos Jason J. Corso
University of Michigan
yayuanli@umich.edu

Abstract

当前的姿态生成方法严重依赖中间表示，无论是通过带有量化的两阶段流水线，还是在推理过程中累积误差的自回归模型。这一根本性限制导致性能下降，尤其是在长期姿态生成中，保持时间一致性至关重要。我们提出了一种新颖的单阶段架构，能够直接从最小化的上下文——单个 RGB 图像和文本描述——中在连续坐标空间中生成姿态，同时在训练和推理中保持一致的分布。我们的关键创新是在通过相对移动预测机制直接在姿态坐标上操作，消除了对中间表示或基于标记的生成需求，该机制保持了空间关系，并通过统一的占位符标记方法实现了训练和推理过程中相同行为的单次前向生成。通过对 Penn Action 和第一人称手部动作基准（F-PHAB）数据集的广泛实验，我们证明了我们的方法在长期生成场景中特别显著地优于现有的基于量化和自回归的方法。

1. 介绍

人体姿势生成已成为计算机视觉中的一个基础问题，其应用涉及动画合成、动作理解和运动预测 [8, 16, 27]。最近的研究工作探索了使用不同模式控制这一生成过程的各种方法：从文本描述 [1, 17]，到音频信号 [15, 22]，再到场景上下文 [5, 33]。

创建语义上有意义且在语境上适当的姿势仍然具有挑战性，特别是由于现有方法中的架构限制。这些方法通常分为两种限制性范式。首先，它们依赖于自回归模型，该模型逐帧生成姿势，由于其性质在训练和推断之间引入了分布偏移 [2]。这种分布偏移导致长期性能下降，因为累积性能 [10]，正如我们在本文后面所显示的。其次，它们是两阶段方法，首先将连续姿势坐标转换为离散标记，通过 VAEs [26, 27] 或在生成之前量化为潜在代码 [17]，从而引入信息损失和计算开销。

这些方法在生成较长序列时表现出显著的退化，这是因为量化误差和分布漂移会随着时间的推移而累积（如图 2 所示）。这种退化影响了许多下游应用（例如，在任务指导中，长期语义连贯性至关重要 [9, 28]）。此外，大多数这些方法需要复杂的输入，如 3D 场景信息 [34, 35]，假设这种详细数据的可用性，这限制了它们在广泛的实际应用中的实用性。

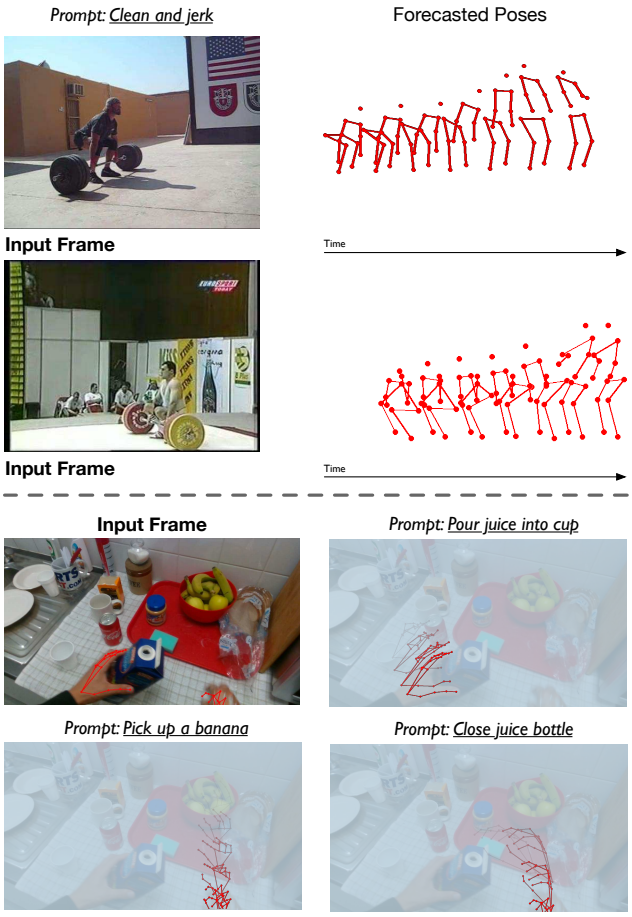

Figure 1. 从单个 RGB 图像和文本描述生成姿态的示例。

为了解决姿态生成中的这些基本限制，我们在我们的方法中引入了两个关键的新颖性：

1. 一种统一的预测机制，确保训练和推理之间分布的一致性，从而实现可靠的长期生成。
2. 一种单阶段姿态生成架构，直接在连续坐标空间中从最小输入——单个 RGB 图像和文本描述——进行操作，保持空间保真度和语义对齐，而不依赖稀缺的 3D 详细场景信息。

我们还探索了语言指导如何能对生成的动作提供语

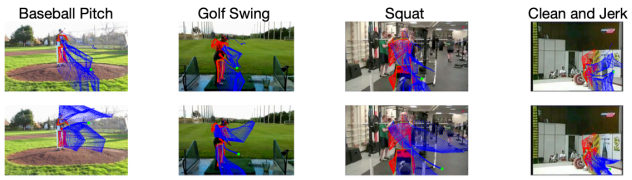

Figure 2. 现有方法中关于长期预测问题的示例。红色姿势是真实数据，蓝色的是预测。由于模型是自动回归训练的，误差会累积。第一行使用了 LSTM，第二行使用了 Transformer 作为骨干网络。

义控制。自然语言提供了一种直观而灵活的方式来指定期望的动作。我们利用简短而精炼的自然语言描述，而不是之前工作中要求的详细运动规范 [11, 21] 。这使得在不需要复杂的运动规范或详细的场景理解的情况下实现有效的控制。这种结合强大的长期生成与语言控制的方法促进了从动画合成到运动规划和任务引导的应用。

我们在不同的姿态目标（人体和手）、视角和领域上，对 Penn Action [37] 和 First-Person Hand Action Benchmark (F-PHAB) [14] 数据集评估我们方法的有效性。通过四个不同粒度的指标来测量性能，我们将我们的模型与五个强力基线进行对比。我们的方法持续优于这些基线，在短期和长期姿态生成中都取得了显著的提升。值得注意的是，我们的方法在涉及大动态和复杂时间动态的挑战场景中表现出色。通过广泛的消融研究和定性结果，我们展示了视觉和文本上下文的整合以及我们的架构设计选择对于实现优越性能的重要性。

2. 相关工作

姿态生成方法 早期的姿态生成方法（有时称为“预测”）专注于仅从姿态历史中预测序列，主要处理关节角度的短期预测 [12] 。后续的研究扩展到了长期预测 [20, 33] ，解决了更复杂的动作比如走、跑和跳。然而，这些方法通常依赖于中间表示或离散的动作类别，从而限制了其泛化能力。

最近的方法已经探索了各种条件信号来引导姿态生成。动作条件方法 [7, 33] 在特定类别中表现出成功，但在细粒度控制方面存在困难。视觉引导的方法 [6, 13] 利用图像上下文，但在长期一致性方面面临挑战。虽然这些方法显示出潜力，但它们通常采用两阶段架构或需要转换为中间表示，从而引入信息损失和计算开销。

语言引导的运动生成 语言指导为生成的动作提供丰富的语义控制。早期的统计模型 [29, 30] 使用简单的二元组表示，而最近的方法则利用深度学习架构 [27, 31] 。像 [26] 这样显著的研究采用了 VQVAE 和 Transformer 架构，但需要对连续的姿势空间进行量化。其他人 [17, 18] 探索文本到动作的生成，但依赖于具有多个阶段的复杂流程。

视觉和语言语境的整合仍然特别具有挑战性。虽然最近的工作 [35] 试图结合这些模态，它们通常需要详细的 3D 场景信息或依赖于中间表示，限制了其实用性。与之相反，我们的方法直接在连续坐标空间中操作，仅以最少的信息作为输入。

生成中的训练-推理一致性 最近的研究已经确定了传统的下一个标记预测方法的基本局限性 [2, 24] ，特别是在连续数据生成方面。虽然一些研究提出量化作为解决方案 [17] ，但这引入了额外的复杂性和潜在的信息损失。我们的工作通过一种统一的预测机制直接解决了这些局限性，该机制在训练和推理过程中保持一致的分布。

3. 方法

3.1. 问题陈述

给定一个自然语言描述和一个视觉场景，我们的目标是生成一系列未来的姿势，这些姿势能够准确地表现所描述的运动，同时保持视觉上下文。

形式上，给定一个二维 RGB 图像 $I \in \mathbb{Z}^{H\times W\times 3}$ ，代表初始场景，以及一个语言描述 $L = [w_1, w_2, ..., w_M]$ ，包含 M 个标记化的单词，我们的目标是生成一个姿势序列 $\boldsymbol{P} = \{P_i\}_{i=1}^{k}$ ，该序列在语义上与语言描述一致，并在视觉上与场景协调，其中每个序列由 k 个未来帧组成。与以往需要复杂 3D 场景数据 [35] 的方法不同，我们的方法仅依赖于单个 RGB 图像。我们使用二维关键点坐标参数化每个姿势 $P_i \in \mathbb{R}^{2N}$ ，其中 N 表示关键点的数量（例如 13 个针对捕捉头部、肩膀和胳膊肘等关节的人体姿势，或 21 个针对代表手腕和手指等关节的手部姿势）。与需要中间表示的以前的方法相比，我们直接在这个连续坐标空间中操作。

3.2. 方法

我们提出了一种新颖的单阶段架构，用于从视觉和文本输入中直接在连续坐标空间中生成姿势。如图 Fig. 3 所示，我们的模型通过三个关键组件融合来自单个 RGB 图像和语言描述的信息。视觉-语言编码器首先通过图像编码器 f_I 处理图像 I 以提取特征 $F_I \in \mathbb{R}^{N_I\times d_I}$ ，同时通过多模态特征融合模块 f_M 处理语言描述 L 以生成融合的图像-文本特征 $F_M \in \mathbb{R}^{N_M\times d_M}$ 。这些特征输入到我们的姿势预测模块，该模块采用占位符令牌机制，以相对于检测到的初始姿势预测未来姿势，从而实现统一的训练和单次前向传递推理。

训练-推理一致性对齐用于一致姿势生成 预测下一个标记在各种任务中表现出色。然而，最近的研究发现，由于训练和推理之间的分布变化存在一些局限性。在分类任务中，这种变化可能限制模型学习长期依赖关系的能力，尽管大型训练数据集可以在一定程度上缓解这个问题 [3] 。然而，我们发现，对于像姿态预测这样的回归任务，这个问题更加显著，导致长期预测能力显著下降。尽管最近的方法 [17] 试图通过在下一个标记预测之前量化姿态来解决这个问题，但它们依赖于

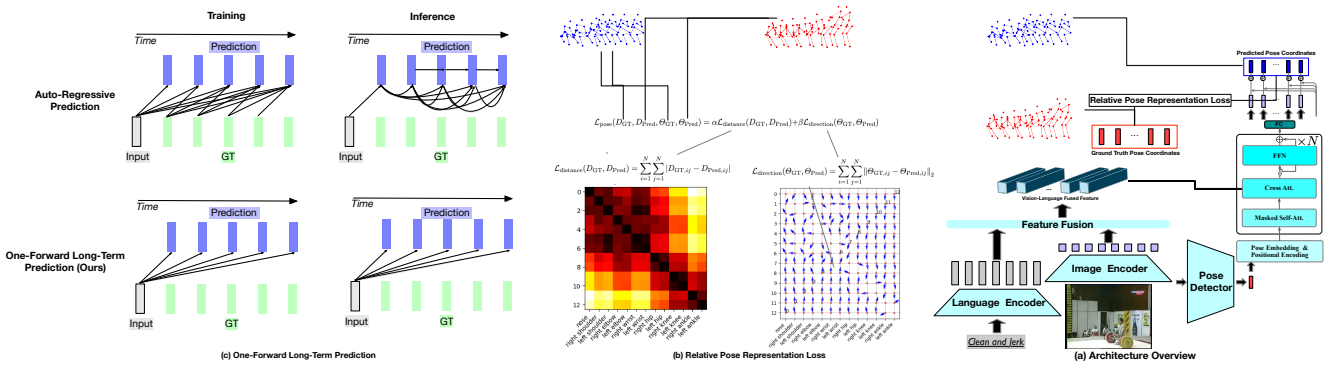

Figure 3. 我们提出的方法概述。给定一个单一的 RGB 图像 I 和自然语言动作描述 L，我们的模型使用多模态编码器提取视觉-语言融合特征。这些特征连同初始姿态 P_0 一起输入到 Transformer 解码器中，该解码器预测一系列未来姿态 $\hat{P}_{1...T}$。我们的方法采用交叉注意力来捕捉视觉和文本输入之间的交互，确保预测的姿态与提供的上下文一致。

两阶段的训练，最终性能在很大程度上依赖于量化和重建步骤的质量。

相比之下，我们提出了一种单阶段方法，直接预测连续的姿态坐标（如 Fig. 3 - (c) 所示），在保持强大表现的同时与之前的两阶段方法相比 Tab. 1。我们的模型在训练和推理期间同时从单个输入中预测多个未来令牌，使用占位符令牌 [PRD] 和非掩蔽自注意力进行高效解码。

我们观察到，训练和推理过程中输入分布的不匹配本质上导致了姿态生成的准确性下降。为了解决这个问题，我们引入了一种对齐输入分布的方法，大大提高了长期预测的准确性。我们的 Transformer 解码器处理初始姿态 P_0，该姿态通过预训练的姿态检测器 [25] 从 I 中提取出来，并结合了融合的视觉-语言特征 F_M，以生成未来的姿态：

$$\hat{P} = \text{Decoder}(P_0, F_M) \in \mathbb{R}^{T \times 2N} \tag{1}$$

然后将预测的姿态 $\hat{P}$ 与真实值 $P_{1...T} \in \mathbb{R}^{T \times 2N}$ 进行比较。

在之前的下一个标记预测（NTP）方法中，训练期间的解码器输入被构建为：

$$X^{NTP} = \begin{pmatrix} x_1^0 & y_1^0 & \cdots & x_N^0 & y_N^0 \\ x_1^1 & y_1^1 & \cdots & x_N^1 & y_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^{T-1} & y_1^{T-1} & \cdots & x_N^{T-1} & y_N^{T-1} \end{pmatrix} \tag{2}$$

其中 $X^{NTP} \in \mathbb{R}^{T \times 2N}$。

掩码自注意力机制确保每个预测 $\hat{P}_i$ ($i = 1, 2, ..., T$) 依赖于前面的姿势 $P_{0...i-1}$。然而，在推理过程中，对预测姿势的依赖会引入误差，随着时间的推移累积，从而降低长期预测的性能。

为了解决这个问题，我们提出了一种新的输入结构，确保了训练和推理的一致性。该结构使用占位符标记来代表未来的时间戳，允许模型在一次前向传递中预测所有未来的姿势：

$$X^{ours} = \begin{pmatrix} x_1^0 & y_1^0 & \cdots & x_N^0 & y_N^0 \\ [\text{PRD}]_1 & [\text{PRD}]_2 & \cdots & [\text{PRD}]_{2N} & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ [\text{PRD}]_1 & [\text{PRD}]_2 & \cdots & [\text{PRD}]_{2N} & \end{pmatrix} \tag{3}$$

其中 $X^{ours} \in \mathbb{R}^{T \times 2N}$。

在这里，占位符标记 $[\text{PRD}] \in \mathbb{R}^{2N}$ 不包含任何信息，但标记了需要预测的位置。这种方法消除了在前向传输过程中（无论是训练还是推理）依赖真实姿态的需要，避免了在推理过程中解码器中的错误累积。位置编码仍然是唯一可区分的信息，引导模型为每个时间戳生成不同的预测。这样的设计允许在坐标空间中进行单阶段的姿态生成，保持推理性能。此外，推理效率得到了提高，因为只需要一次前向传输即可批量生成所有姿态。

生成时间相对的姿态移动 我们的模型通过预测相对运动从初始姿势生成姿态序列。我们利用相对位移预测，而不是预测绝对坐标。考虑一种需要头部关节向下移动的情况：我们的模型不直接预测其在全局坐标中的最终位置（ $x = 0.7, y = 0.9$ ），而是首先检测其在输入图像中的当前位置（ $x = 0.75, y = 0.8$ ），然后预测达到目标所需的位移（ $\Delta x = -0.05$, $\Delta y = 0.1$)。这种方法提供了两个关键优势：它结合了初始姿势的空间上下文（例如，图像的右侧或左侧），并将运动直接建模为位移而非绝对位置。

这个设计的有效性在我们的消融研究中得到了证明。

视觉-语言特征编码 多模态特征通过视觉-语言编码器进行融合。图像编码器 f_I 处理视觉输入 I，以生成图像特征 $F_I \in \mathbb{R}^{N_I \times d_I}$。对于语言表示，我们依赖于简洁的自然语言术语（例如，“挥杆高尔夫”），与之前的工作 [11, 21] 相比，不需要详细且不太实用的文本描述

(例如，“上半身向右侧转动，髋部向左侧扭转，手臂沿着弯曲路径向下移动球杆击球”)。多模态特征融合模块 f_M 将这些简洁的描述与视觉特征集成，以生成融合特征 $F_M \in \mathbb{R}^{N_M \times d_M}$ 。在实践中，我们使用 BLIP 的 [23] 单模态编码器模块用于 f_I ，以及其图像基础的文本编码器用于 f_M ，并保持预训练权重冻结。

为准确捕捉关节之间的内在空间关系并减少全局坐标中的冗余，我们设计了一种结合关节之间相对距离和方向的损失函数 [4] ，其中包括均方误差（MSE）组件。这种公式通过强调关节运动的空间一致性来提高姿态预测精度，确保模型既能学习相对定位又能掌握姿态的整体结构。

我们在一个距离矩阵 $D \in \mathbb{R}^{N \times N}$ 中表示相邻关节之间的成对欧几里得距离：

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4}$$

其中 (i, j) 是相邻关节。

方向表示 方向矩阵 Θ 编码了相邻关节之间的单位方向向量：

$$\Theta_{ij} = \left(\frac{x_j - x_i}{D_{ij}}, \frac{y_j - y_i}{D_{ij}} \right) \tag{5}$$

单个姿态的损失： 总姿态损失是距离损失和方向损失的加权和：

距离损失

$$\mathcal{L}_{\text{distance}}(D_{\text{GT}}, D_{\text{Pred}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} |D_{\text{GT},ij} - D_{\text{Pred},ij}| \tag{6}$$

方向损失

$$\mathcal{L}_{\text{direction}}(\Theta_{\text{GT}}, \Theta_{\text{Pred}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|\Theta_{\text{GT},ij} - \Theta_{\text{Pred},ij}\|_2 \tag{7}$$

总姿势损失

$$\mathcal{L}_{\text{pose}} = \alpha \mathcal{L}_{\text{distance}} + \beta \mathcal{L}_{\text{direction}} \tag{8}$$

，其中 α 和 β 控制每个损失组件的相对贡献。

一系列姿势的损失 给定一个真实位姿序列 $P_{1\ldots k} \in \mathbb{R}^{k \times 2N}$ 和预测位姿 $\hat{P}_{1\ldots k} \in \mathbb{R}^{k \times 2N}$ ，总序列损失计算为所有时间戳的平均值：

$$\mathcal{L}_{\text{seq}} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}_{\text{pose}}(D_{\text{GT}}, D_{\text{Pred}}, \Theta_{\text{GT}}, \Theta_{\text{Pred}}) \tag{9}$$

批处理损失 当对多个样本进行优化时，批量损失是序列损失的平均值：

$$\mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{seq},i} \tag{10}$$

其中 B 是批量大小。

最终损失 用于训练模型的最终损失函数是相对姿态损失和标准 MSE 损失的组合：

$$\mathcal{L} = \mathcal{L}_{\text{rel}}(\alpha, \beta) + \theta \mathcal{L}_{\text{batch},mse} \tag{11}$$

其中 θ 控制 MSE 项的贡献。

4. 实验

在本节中，我们通过定量和定性结果展示我们方法的性能。我们介绍我们的实验设置，包括数据集、评价指标和基线。我们将我们的方法与强基线方法以及以前的 SOTA 姿态生成方法进行比较。我们从预测步数和困难程度两个方面细粒度地研究性能。此外，我们进行了消融研究以展示我们设计的有效性。

4.1. 数据集

我们在各种数据集上评估了我们的方法，以展示其在不同目标（人体和手）和场景中的泛化能力。所有数据集都是基于视频的，每个片段都用自然语言动作描述进行了注释。对于没有姿态注释的数据集，我们使用 Mediapipe 生成伪注释。我们将数据的 90 % 用于训练，10 % 用于测试，使用 45 个时间戳的预测范围进行训练，并评估从 1 到 45 帧的所有预测步骤。

我们使用了两个数据集：Penn Action [37] 和 First-Person Hand Action Benchmark (F-PHAB) [14] 。它们涵盖了两个目标（人体和人手）、多个领域（例如，烹饪、体育）和视角（自我中心和外中心）。

这些数据集使我们能够在不同领域和视点中探索各种用例，从精细的手部动作到更大的人体运动。

在我们的实现中，我们使用冻结的 BLIP 来融合多模态特征。所用的单模态编码器是 ViT-g/14，文本编码器是 BERT。归一化因子 σ 被设置为 0.8。Transformer 模型使用原生 PyTorch 实现进行构建。我们的模型经过 AdamW 优化器的训练以收敛，学习率固定为 10^{-4} 。我们的方法的训练在 1 个 NVIDIA H100 GPU 上进行，每个 GPU 分配的批处理大小为 64。

4.2. 指标

我们使用四个度量标准来评估姿态生成的不同方面的性能：

均方根误差 (RMSE) 度量生成和注释关键点之间的平均距离，提供基于输入图像尺寸的归一化误差。形式

Method	Penn Action				F-PHAB			
	ADE ↓	FDE ↓	PCK ↑	RMSE ↓	ADE ↓	FDE ↓	PCK ↑	RMSE ↓
NN_P	0.0901	0.1050	0.6663	0.0568	0.1684	0.1540	0.3769	0.1091
NN_{VL}	0.2421	0.2461	0.2997	0.1566	0.2583	0.2588	0.2793	0.2143
LSTM [19]	0.1635	0.2622	0.3820	0.1061	0.1938	0.1938	0.3018	0.1358
Naive Transformer [32]	0.1726	0.2303	0.3435	0.1110	0.1922	0.2031	0.3001	0.1459
Quantization + Transformer [17]	0.2549	0.2478	0.1798	0.1664	0.2431	0.2387	0.2077	0.1601
Ours	0.0578	0.0766	0.8179	0.0350	0.0967	0.0855	0.7645	0.0683

Table 1. 我们的方法在 Penn Action 和 F-PHAB 数据集上的 ADE、FDE、PCK 和 RMSE 指标与基准方法进行比较。我们的方法始终优于所有基准，尤其是在长期生成方面。

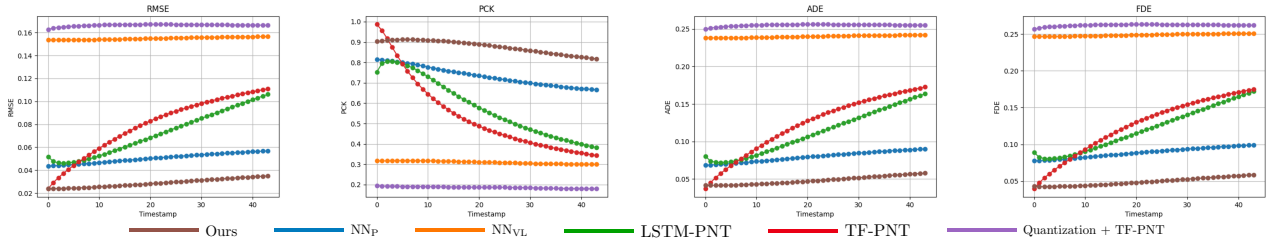

Figure 4. 各个生成时间戳的表现。我们的方法在大多数时间戳上持续优于基线方法。

化为：

$$\text{RMSE}(\hat{Y}, Y) = \sqrt{\frac{1}{T \times N} \sum_{t=1}^{T} \sum_{n=1}^{N} (\hat{Y}_{tn} - Y_{tn})^2} \quad (12)$$

关键点生成位置在注释位置的某个阈值（δ）内的百分比。我们为人体目标设置 $\delta = 0.05$，为手部设置 $\delta = 0.15$：

$$\text{PCK}(\hat{Y}, Y) = \frac{1}{T \times K} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{I}(||\hat{Y'}_{tk} - Y'_{tk}|| < \delta) \quad (13)$$

平均位移误差 (ADE)　衡量在所有时间戳上生成的轨迹与真实轨迹之间的平均 l_2 距离：

$$\text{ADE}(\hat{Z}, Z) = \frac{1}{T} \sum_{t=1}^{T} ||\hat{Z}_t - Z_t|| \quad (14)$$

终位移误差 (FDE)　衡量生成轨迹和真实轨迹在最终时间戳时的 l_2 距离：

$$\text{FDE}(\hat{Z}, Z) = ||\hat{Z}_T - Z_T|| \quad (15)$$

4.3. 基线

我们引入了五个受到先前研究启发的基线，以对我们新提出的从单个 RGB 图像和文本生成指导性姿势的问题进行基准测试。这些基线包括最近邻（具有不同的相似性特征）、基于 LSTM 的姿势生成、原始的 Transformer 解码器，以及带有姿势量化的两阶段生成方法。

最近邻 (NN)　基线根据输入的图像和文本描述，从训练数据中检索最相似的样本，并使用其对应的姿态作为预测。它衡量现有的姿态序列作为测试样本的预测能力。如果这种方法表现良好，则表示测试输出在训练集中得到了良好的表示。我们探索了两种变化：

- NN_P：相似性计算为输入图像和训练图像中关键点坐标之间的欧几里得距离。
- NN_{VL}：相似性基于融合视觉和语言特征之间的欧几里得距离。

我们使用基于 LSTM 的模型进行姿态生成，利用视觉-语言特征作为输入。

朴素的 Transformer　该基线使用原始 Transformer 解码器进行下一个词元预测任务，以将其性能与我们提出的解决方案进行比较。

量化 + 变压器　受单模态姿态生成的最新技术启发，这一两阶段基线首先使用 VQ-VAE 对姿态坐标进行量化，然后应用 Transformer 解码器。这使我们能够评估两阶段方法在我们场景中的效率。

4.4. 结果

我们针对从单个 RGB 图像生成视觉-语言引导姿态的新问题呈现了首次结果。我们使用之前介绍的指标将

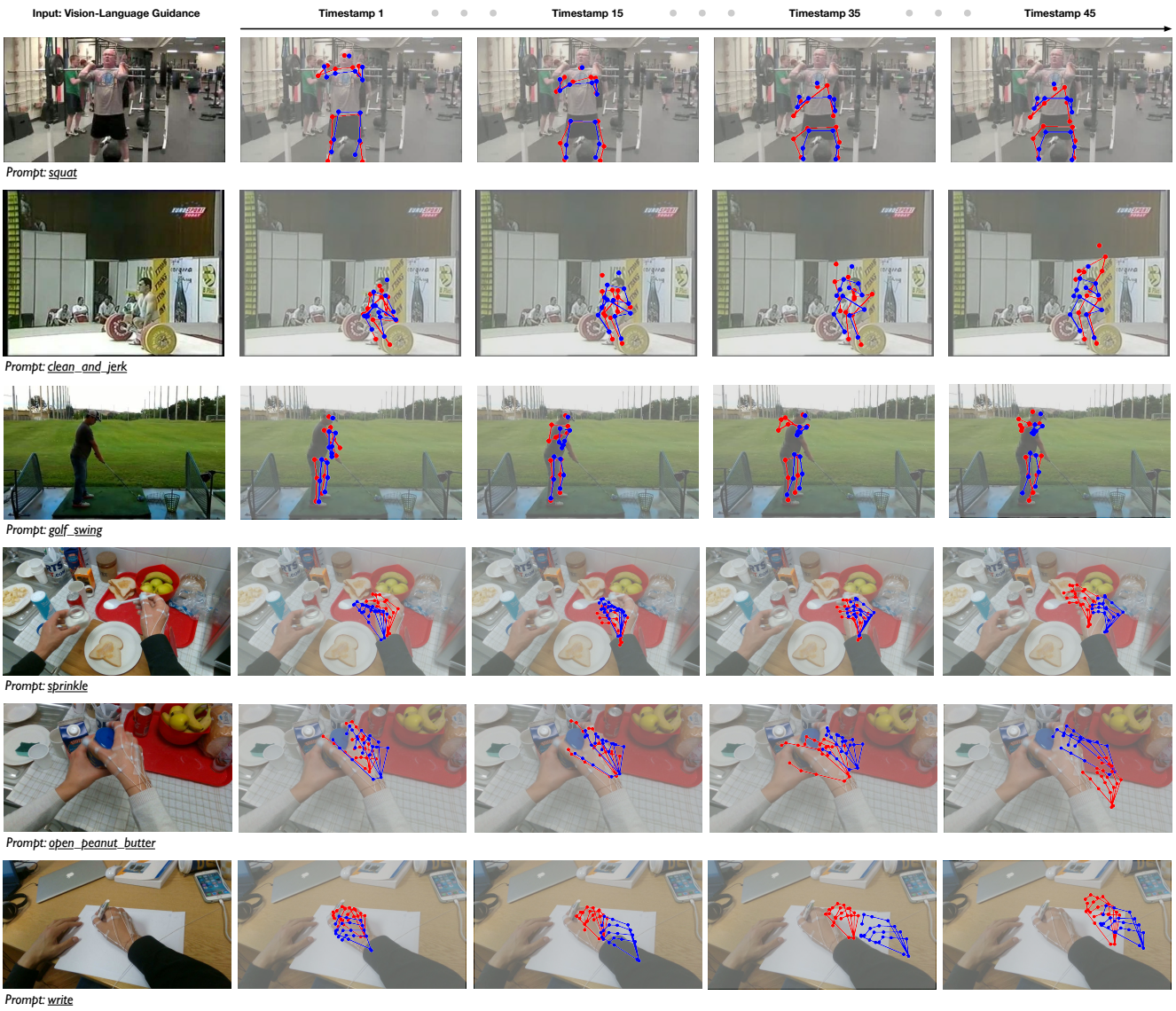

Figure 5. 我们提出的方法的定性结果。红色姿势是注释，蓝色姿势是我们方法的预测。

	Penn Action				F-PHAB			
Method	ADE ↓	FDE ↓	PCK ↑	RMSE ↓	ADE ↓	FDE ↓	PCK ↑	RMSE ↓
TF	0.1726	0.2303	0.3435	0.1110	0.1922	0.2031	0.3001	0.1459
+ Pose Det.	0.1246	0.1548	0.4412	0.0978	0.1638	0.1702	0.3323	0.1287
+ Train in Batch (Full Attn)	0.0693	0.0693	0.7741	0.0431	0.1425	0.1433	0.3847	0.0908
+ Causal Mask	0.0595	0.0742	0.8204	0.0370	0.1231	0.1335	0.4210	0.0823
+ Ours (relative loss)	0.0578	0.0766	0.8179	0.0350	0.0967	0.0855	0.7645	0.0683

Table 2. 消融研究结果比较了在 Penn Action 和 F-PHAB 上使用 ADE、FDE、PCK 和 RMSE 的不同模块选择。

我们的方法与五个基线进行了比较。此外，我们研究了不同时间戳下的性能，分析了难以处理的样本，进行消融研究，并与最先进的单模式方法进行比较，以展示整合视觉和文本信息的必要性。

表 1 总结了我们的方法与五个具有挑战性的基线相比较的性能。我们的方法优于所有基线。当传统的

Method	Penn Action			
	ADE ↓	FDE ↓	PCK ↑	RMSE ↓
NN_P	0.1118	0.1653	0.5491	0.0703
NN_{VL}	0.2247	0.2577	0.1813	0.1446
LSTM	0.1735	0.2886	0.3159	0.1121
Naive Transformer	0.1680	0.2546	0.3265	0.1076
Quantization + Transformer	0.2467	0.2556	0.1235	0.1586
Ours	0.0922	0.1573	0.6824	0.0573

Table 3. 在最难的 10 个% 测试样本上的结果。难度由注释姿势序列的方差定义。

Method	Penn Action			
	ADE ↓	FDE ↓	PCK ↑	RMSE ↓
TM2T [17]	0.2684	0.2924	0.1709	0.2708
PHD [36]	-	-	0.7720	-
Ours	0.0169	0.0170	0.860	0.012

Table 4. 在 Penn Action 数据集上与 SOTA 单模态生成方法的比较。

最近邻居（NN）结合来自先进深度网络（例如姿态检测器或视觉-语言特征提取器）中提取的信息时，它表现出令人惊讶的良好性能（NN_P 和 NN_{VL} ）。此外，LSTM-NTP 和 TF-NTP 展示了在两个流行的深度网络结构——LSTM 和 Transformer 上使用下一个标记预测（NTP）进行训练的局限性。NTP 无法帮助模型在训练期间学习超过一个未来时间戳的运动，并导致姿态生成在推理过程中出现“漂移”行为（图 2 ）。受到先前工作 [17] 的启发，我们在使用 Transformer 进行 NTP 之前进行了量化（量化 + TF-NTP）。然而，正如我们在 Sec. 3.2 中所论述的那样，这种方法严重依赖于量化性能，这需要大量数据以自监督的方式构建离散空间。我们展示了虽然 NTP 在量化空间中表现不错，但将量化的姿态重构为实际的姿态坐标成为性能瓶颈，并导致整体性能不佳。

另一方面，我们的方法旨在通过一次前向预测直接在坐标空间中生成未来姿态（在一次前向中预测多个标记） Sec. 3.2 ，成功克服了 LSTM-NPT 和 TF-NPT 中的“漂移”行为，并避免了量化过程中的信息损失（量化 + TF-NTP）。这种改进在定量 Tab. 1 和定性结果 Fig. 5 中都有所体现。

我们进一步分析了每种方法在不同时间戳上的性能以及在最难的 10 个% 测试样本上的表现。

如图 4 所示，我们的方法在大多数预测时间范围内，尤其是较长的时间范围内，一直优于基线方法。此外，与 LSTM-NTP 和 TF-NTP 性能大幅下降相比，我们的方法在预测时间范围增加时保持了稳健的性能。这突出了我们设计的一次前向预测多个未来时间戳的效率。

此外，Tab. 3 表明我们的方法在测试样本中最难的 10 % 中保持了最佳性能。难度是由关键点运动的方差定义的。也就是说，地面真实未来姿态序列中更多的运动表明捕捉动态更加困难。

消融研究 表 2 显示了我们消融研究的结果。我们比较了不同设计选择的效果，例如编码器选择和使用批内训练。我们的完整模型，通过在一次前向传播中从检测到的初始姿势生成所有未来姿势并结合相对损失函数，始终优于简化版本。这些结果强调了我们的设计选择对于实现最佳性能的重要性。

与最新单模态姿势生成方法的比较 我们将我们的方法与最先进的单模态姿态生成方法进行比较，以突出在生成指导性姿态时同时使用视觉和文本信息的必要性。需要注意的是，这种比较并不完全公平，因为之前的方法利用了更强的输入，如多个帧，而我们仅使用单个 RGB 图像和简单文本。

表 4 显示，我们的方法显著优于仅文本的 SOTA 方法 TM2T [17] 和仅视觉的 SOTA 方法 [36] 。

5. 结论

在这项工作中，我们引入了一种新的用于视觉语言引导的姿态预测的单阶段架构，它直接在连续的坐标空间中运作。通过在训练和推理之间保持一致性并利用相对运动预测，我们的方法在保持空间保真度的同时，实现了优越的姿态生成。通过在标准姿态数据集上的广泛实验，包括 Penn Action 和 F-PHAB，我们的方法在多个指标上展示了最先进的性能，大大超越了现有的基线。此外，我们的设计在处理具有挑战性的场景方面表现出色，如大幅度运动和动态时间序列，凸显了其在长期预测中的鲁棒性。

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. 2019 International Conference on 3D Vision (3DV) , pages 719–728, 2019. 1

[2] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction, 2024. 1, 2

[3] Tom B Brown. Language models are few-shot learners. arXiv preprint arXiv:2005.14165 , 2020. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition , pages 7291–7299, 2017. 4

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. ArXiv , abs/2007.03672, 2020. 1

[6] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In Proceedings of the IEEE conference on com-

puter vision and pattern recognition , pages 548–556, 2017. 2

[7] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 9544–9555, 2023. 2

[8] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In 2019 IEEE winter conference on applications of computer vision (WACV) , pages 1423–1432. IEEE, 2019. 1

[9] Dima Damen, Michael Wray, Ivan Laptev, Josef Sivic, et al. Genhowto: Learning to generate actions and state transformations from instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 6561–6571, 2024. 1

[10] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. ArXiv , abs/2305.18654, 2023. 1

[11] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3d human pose. In CVPR , 2024. 2, 3

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In Proceedings of the IEEE international conference on computer vision , pages 4346–4354, 2015. 2

[13] Tomohiro Fujita and Yasutomo Kawanishi. Future pose prediction from 3d human skeleton sequence with surrounding situation. Sensors , 23(2):876, 2023. 2

[14] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of Computer Vision and Pattern Recognition (CVPR) , 2018. 2, 4

[15] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , pages 9942–9952, 2023. 1

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia , page 2021–2029, New York, NY, USA, 2020. Association for Computing Machinery. 1

[17] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In European Conference on Computer Vision , pages 580–597. Springer, 2022. 1, 2, 5, 7

[18] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 1900–1910, 2024. 2

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Comput. , 9(8):1735–1780, 1997. 5

[20] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 16750–16761, 2023. 2

[21] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: Learning egocentric action frame generation via visual instruction tuning. arXiv preprint arXiv:2312.03849 , 2023. 2, 3

[22] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. ArXiv , abs/1911.02001, 2019. 1

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning , pages 12888–12900. PMLR, 2022. 4

[24] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838 , 2024. 2

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 , 2019. 3

[26] Sihan Ma, Qiong Cao, Jing Zhang, and Dacheng Tao. Contact-aware human motion generation from textual descriptions. arXiv preprint arXiv:2403.15709 , 2024. 1, 2

[27] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 10985–10995, 2021. 1, 2

[28] Jackson Spencer, Sanjiban Choudhury, Matthew Barnes, Christopher Dellin, et al. What matters in learning from offline human demonstrations for robot manipulation. In Conference on Robot Learning , 2022. 1

[29] Wataru Takano and Yoshihiko Nakamura. Bigram-based natural language model and statistical motion symbol model for scalable language of humanoid robots. 2012 IEEE International Conference on Robotics and Automation , pages 1232–1237, 2012. 2

[30] Wataru Takano and Yoshihiko Nakamura. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. The International Journal of Robotics Research , 34:1314 – 1328, 2015. 2

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In The Eleventh International Conference on Learning Representations , 2023. 2

[32] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems , 2017. 5

[33] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 9401–9411, 2021. 1, 2

[34] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In Advances in Neural Information Processing Systems (NeurIPS) , 2022. 1

[35] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2024. 1, 2

[36] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 7114–7123, 2019. 7

[37] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In Proceedings of the IEEE international conference on computer vision , pages 2248–2255, 2013. 2, 4