通过文本到骨架级联实现可控复杂人体运动视频生成

Ashkan Taghipour，MortezaGhahremani，ZinuoLi，*IEEE*会员，Hamid Laga，Farid Boussaid，*IEEE*高级会员，以及Mohammed Bennamoun，*IEEE*高级会员

arXiv:2603.08028v1 [cs.CV] 9 Mar 2026

摘要——**生成复杂人体运动（如翻转、侧手翻和武术）的视频，对于当前的视频扩散模型而言仍然具有挑战性。纯文本条件控制对于细粒度运动控制存在时间上的歧义，而显式的基于姿态的控制虽然有效，却要求用户提供完整的骨架序列；对于持续时间较长的动态动作而言，这类序列的生成成本很高。我们提出了一种两阶段级联框架，以解决这两个局限。首先，自回归文本到骨架模型根据自然语言描述生成二维姿态序列，并在此前生成的姿态条件下预测每个关节，从而捕捉复杂运动中的长程时间依赖和关节间协调。其次，姿态条件视频扩散模型根据参考图像和生成的骨架序列合成视频，并采用DINO-ALF（自适应层融合）这一多级参考编码器，在大幅度姿态变化和自遮挡情况下保持外观和服装细节。针对缺乏公开可用的复杂人体运动视频生成数据集这一问题，我们引入了一个基于Blender的合成数据集，其中包含2,000个视频，展示了不同角色执行各种杂技和特技动作，并能够全面控制外观、运动和环境。该数据集填补了关键空白，因为现有基准测试严重缺乏对杂技和特技动作的覆盖，同时也避免了网络采集数据带来的版权和隐私问题。在我们提出的合成数据集和Motion-X Fitness基准上的实验表明，我们的文本到骨架模型在FID、R-精确率和运动多样性方面优于此前的方法；同时，我们的姿态到视频模型在VBench的时间一致性、运动平滑度和主体保持指标上，在所有对比方法中取得了最佳结果。更多结果请参见项目页面。**

索引词—**视频生成、文本到姿态、骨架引导扩散。**

I. 引言

DIFFUSION 模型已成为计算机视觉领域占主导地位的生成式框架 [1], [2], [3], [4],[5]。将其扩展到视频后，文本到视频（T2V） [6], [7] 和文本与图像到视频（TI2V）生成 [8], [9] 均取得了显著进展。在 TI2V 中，参考图像定义了主体的外观，而文本描述所需的运动，从而能够合成具有真实运动的照片级真实内容 [10], [11]。尽管取得了这些进展，如何可控地生成复杂人体运动（如翻转、侧手翻、杂技和武术动作），仍是一项尚待解决的挑战 [12],[13]。在本文中，我们将复杂人体运动定义为非重复、高度动态的动作，其特点是姿势变化幅度大且频繁出现自遮挡。

当前的大规模视频扩散模型，无论是开源的 [14], [7], [8] 还是专有的 [15], [16],，都难以处理此类运动，常常生成不合理的肢体轨迹、身体形状或服装的时间不一致性，以及外观漂移 [17]。这些问题限制了其在体育内容创作 [18], 、虚拟教练 [19],、特技预可视化 [20], [21],和虚拟形象动画 [22]中的应用。

复杂运动的 TI2V 生成面临的一个核心挑战是，仅凭文本无法提供充分的控制 [23]。“一个人正在进行后空翻”这样的描述在语义上很清晰，但在时间维度上存在歧义——它没有具体说明逐帧关节轨迹或子运动时序。为了解决这一问题，近期研究采用明确的运动信号作为条件，例如二维姿态或深度图，从而显著提升了可控性 [20], [24],[25]。然而，这些方法要求用户提供完整的姿态序列，这对于复杂动作而言并不实用：生成高质量的二维姿态既耗时，又需要专门的工具 [26], [27]。因此，实践者只能依赖少量模板运动库，从而限制了表现力和可扩展性。

即使有姿态序列可用，在复杂运动下保持参考图像的外观仍然具有挑战性。现有的姿态条件方法使用 CLIP 表征对参考图像进行编码，并通过交叉注意力注入这些表征 [12],[13], [24]。基于 CLIP 的条件控制适用于姿势变化适中、结构化的运动，但在复杂动作中常见的大幅变形、快速过渡和自遮挡情况下表现不佳。CLIP 生成的是缺乏细粒度空间细节的全局语义表征 [28], [29],，因此难以在视角变化或遮挡情况下重建局部外观线索，例如服装纹理和身体部位 [30]。因此，这些模型在动态运动过程中会出现外观不一致性、纹理模糊以及身体部位细节丢失 [31], [32]。

这些观察促使我们提出一种两阶段级联框架，以应对这两项挑战。在第一阶段，自回归文本到骨架模型将自然语言转换为二维姿态序列，从而捕捉复杂、非重复动作中的长程时间依赖和关节间协调，而无需手动生成姿态

Ashkan Taghipour、Zinuo Li 和 Mohammed Bennamoun 任职于澳大利亚西澳大学计算机科学与软件工程系。（电子邮箱：ashkan.taghipour@research.uwa.edu.au；Zinuo.li@research.uwa.edu.au，mohammed.bennamoun@uwa.edu.au）

Morteza Ghahremani 任职于德国慕尼黑机器学习中心（MCML）和慕尼黑工业大学（TUM）。（电子邮箱：morteza.ghahremani@tum.de）

Hamid Laga 就职于澳大利亚默多克大学信息技术学院。（电子邮箱：h.laga@murdoch.edu.au）

Farid Boussaid 任职于澳大利亚西澳大利亚大学电气、电子与计算机工程系。（电子邮箱：farid.boussaid@uwa.edu.au）。

。在第二阶段，姿态条件视频扩散模型根据参考图像和生成的骨架序列合成视频，并利用 DINO-ALF（自适应层融合），通过自适应聚合多个 DINO 层中空间局部化的图像块描述符，在大幅变形和自遮挡情况下保留外观。为了在复杂运动上进行训练和评估，我们还构建了一个基于 Blender 的合成数据集，其中包含 2,000 段以杂技与特技动作为特色的视频，填补了现有基准测试的空白；现有基准主要关注规律、重复的活动。

基于这样一个认识：复杂人体运动的可控生成需要明确地将运动规划与外观合成解耦，我们的主要贡献如下：

- **通过自回归文本到骨架生成进行运动规划**。 我们提出一种自回归文本到骨架模型，将自然语言描述转换为关节级二维姿态序列，并对长程时间依赖和关节间协调进行显式建模。与仅关注姿态真实性的以往文本到姿态方法不同，我们的方法生成结构化运动计划，为复杂的非重复动作提供明确且可编辑的控制信号，无需手动生成姿态。

- **基于 DINO-ALF 的形变感知姿态条件视频扩散**。 我们提出 DINO-ALF（自适应层融合），这是一种用于姿态条件视频扩散的形变感知外观条件机制。
DINO-ALF 利用空间局部化的图像块描述符，并自适应地聚合 DINOv3 多个层中的互补特征，从而在大幅姿态形变和自遮挡情况下保持外观对应关系。该设计能够稳健地保留身份和服装细节。

- **面向复杂人体运动的合成数据集**。我们构建并发布了一个基于 Blender 的合成数据集，包含 2,000 个复杂运动视频，专门针对现有基准中较少涉及的特技和类似特技动作，同时避免了网络采集数据所带来的版权和隐私问题。

II. 相关工作

A. 姿态条件人体视频生成

姿态条件视频生成通过引入二维骨架、深度图或参数化人体模型等显式结构信号来引导人物动画 [12], [13], [24], 从而弥补纯文本控制的局限。这一范式面临两项核心挑战：如何将姿态控制注入视频去噪器，以及如何在整个生成序列中保留参考外观。

姿态控制注入。 不同的架构选择决定了结构控制如何与视频去噪器融合。ControlNet 风格适配器 [12], [33]通过可训练的副本分支注入姿态引导，这些分支向冻结的骨干网络层添加残差。DreamPose [27] 与 DreaMoving [26]分别针对时尚和角色动画微调此类适配器，而 Follow-Your-Pose [34] 则实现姿态引导生成，并使用无姿态训练视频。基于 DiT 的架构支持替代注入模式：VACE [35]将姿态渲染为 RGB控制视频，将其编码为时空对齐的上下文令牌，并通过专用的上下文块注入这些令牌；Human4DiT [36] 借助四维扩散变换器将姿态条件生成扩展至自由视点合成；UniAnimate [37] 将特征编码器条件化适配到 DiT 骨干网络。HumanVid [38] 解耦相机运动与身体运动，以实现相机可控动画，而 SteadyDancer [20] 与 Wan-Animate [21] 则优化时间一致性与首帧保持。

参考外观编码。在遵循驱动姿态的同时保留参考外观同样至关重要。早期方法使用 CLIP 嵌入对参考图像进行编码，并通过交叉注意力将其注入模型 [12], [27]。然而，CLIP 的全局语义表示缺乏细粒度空间细节，因此在大幅变形时会导致外观漂移 [28], [29]。为解决这一问题，MagicAnimate [13] 提出了 ReferenceNet，这是一种通过复制去噪器的空间层来提取稠密外观特征的框架。Champ [24] 利用 SMPL 形状和法线图中的三维参数线索增强姿态引导。DisCo [39] 和 MagicPose [40]通过独立控制通路或多阶段训练，将运动与外观解耦。MimicMotion [41] 强化了对手部和面部等高频失效区域的引导，而 HyperMotion [17]提出了一种空间低频增强旋转位置编码设计，以应对快速姿态变化。StableAnimator [42]和 Animate-X [43]改进了姿态特征与参考特征的针对人体的融合方式；AnimateAnyone 2 [25] 融入环境可供性，SCAIL [44]利用三维一致姿态表示，而 EchoMotion [18] 则通过双模态扩散统一视频生成与动作生成。

尽管这些方法推动了可控人体动画的发展，但大多数方法仅在舞蹈序列等相对规律的运动场景中得到验证。生成复杂且非重复的运动仍然具有挑战性：大多数流程都假设用户提供姿态序列，而对于动态动作 [27], [26], 这类序列很难实际生成。此外，现有的参考编码在面对大幅变形和自遮挡时表现不佳 [31],[32]。

B. 将文本到骨架用作视频运动控制

姿态引导人体视频生成的一项实际局限在于，它假设用户能够提供完整的未来姿态序列[44], [20], [21], , 而对于持续时间较长、较为复杂的动作，这样做成本很高。为减轻这一负担，同时保留明确且可编辑的控制接口，近期研究尝试从语言生成中间运动控制信号，最常见的是二维骨架序列，可直接驱动姿态条件视频生成器 [45],[46], [47], [48]。这种选择与图像空间视频扩散十分契合：二维姿态与视图对齐，且易于编辑，能够提供结构轨迹，同时不会引入额外的相机或深度歧义 [49], [45]。

HumanDreamer [45] 遵循这一范式，先根据文本生成二维姿态序列，再驱动姿态到视频生成器；该方法提出了一种基于 DiT 的姿态生成器，并引入额外的潜空间对齐目标，以更好地匹配语言

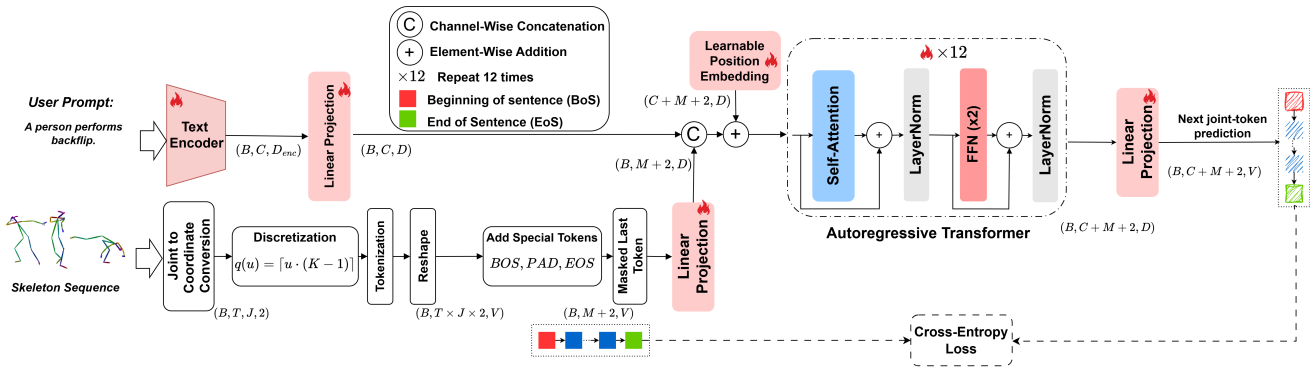

图1. 文本到骨架概述 用于训练的生成架构。文本提示经编码后作为条件前缀前置到姿态词元序列。自回归 Transformer 以所有先前生成的词元和文本描述为条件，预测每个关节词元。

和动作。Holistic-Motion2D [49] 及其扩展版本 Tender 利用全身关键点扩展了文本到二维动作学习，并引入部位感知和置信度感知建模，以更好地处理二维姿态轨迹中的噪声检测结果和遮挡。Motion-2-to-3 [50] 以解耦形式对二维动作进行建模，并将生成扩展到多视角设置，以实现视角一致的动作。从互补的角度来看，Mimic2DM [51]从自然环境中的二维关键点轨迹中学习控制，并采用自回归 Transformer，在分层控制框架中生成二维参考动作。MoSA [52] 在单独的结构阶段根据文本生成人体关键点，并使用投影后的二维骨架作为基于 DiT 的视频生成器的引导。Human-Motion2D Generation [53] 提出了一种扩散模型，可根据文本生成二维骨架序列（也可选择以初始动作帧为条件），并通过强化学习（RL）风格的微调阶段进一步提升真实感和对齐效果，表明生成骨架可以作为骨架引导的视频生成的显式控制信号。

文本到骨架生成还可实现手语视频合成，其中骨架序列作为中间控制，用于渲染为人体外观 [54]。Mixed SIGNals [47] 通过混合动作基元生成手语姿态序列，Sign-IDD [48] 通过骨骼方向和长度约束强制骨架一致性，SignLLM [46] 输出驱动姿态到视频合成的骨架表示。

尽管已有这些进展，现有的文本到二维姿态模型通常是为一般运动或特定下游场景开发的，而它们生成涉及快速过渡和自遮挡的高度动态、非重复姿态轨迹的能力仍较少被探索。特别是，具有快速过渡和自遮挡的高度动态动作对时间一致性和关节间协调提出了比先前基准中通常强调的相对规律运动更高的要求。

III. 所提方法

所提方法采用两阶段级联框架，旨在根据文本生成明确的运动控制，同时在复杂动作中保持外观。若干设计选择是由复杂且非重复动作的特定需求所驱动的。我们采用二维姿态空间而非三维表示，因为二维骨架与视频扩散模型的图像平面直接对齐，可避免三维表示所带来的相机投影歧义。我们对离散化关节坐标采用自回归因子分解，因为复杂动作具有很强的序列依赖关系——每个关节的位置取决于先前的运动轨迹以及其他关节的配置；离散词元则使我们能够通过标准的下一词元目标对这一分布进行建模，同时支持可控采样策略。对于外观条件控制，我们从多个 DINO 层提取特征，而不依赖单个 CLIP 嵌入，因为较早的 DINO 层能够捕捉纹理丰富的局部细节，而较晚的层则编码视角不变语义；对这些特征进行自适应融合，可提供所需的空间局部线索，从而在大幅变形和自遮挡情况下保持身份一致性。第一阶段，文本到骨架生成（第 III-A 节），将自然语言转换为二维姿态序列；第二阶段，姿态条件视频生成（第 III-B 节），则利用 DINO-ALF 外观条件，根据参考图像和生成的骨架合成视频帧。

A. 文本到骨架生成

第一阶段将自然语言运动描述映射为一系列二维骨架关键点（图 1）。我们将运动表示为一系列词元，其中每个词元对应特定时间步的单个关节的二维坐标。随后，基于Transformer的自回归模型 [55] 根据此前生成的所有关节和输入文本，逐个预测关节词元。这种分解方式尤其适用于复杂人体运动，因为它使模型能够捕捉强烈的时间依赖和关节间依赖，而固定的关节顺序和有限的离散词表则能自然地促进生成结构有效的姿态序列。

1) 姿态表示与词元化

将一个运动片段表示为每帧包含 T 个关节、共 J 帧，$\mathbf{P} \in \mathbb{R}^{T \times J \times 2}$，其中每个关节由归一化图像平面坐标 $(x_{t,j}, y_{t,j}) \in [0,1]^2$参数化，对应第 $t \in \{1, \ldots, T\}$ 帧和关节索引 $j \in \{1, \ldots, J\}$。实际应用中，我们将像素坐标按帧宽和帧高进行归一化，使所有关节都位于 $[0,1]^2$ 中，然后再进行离散化。

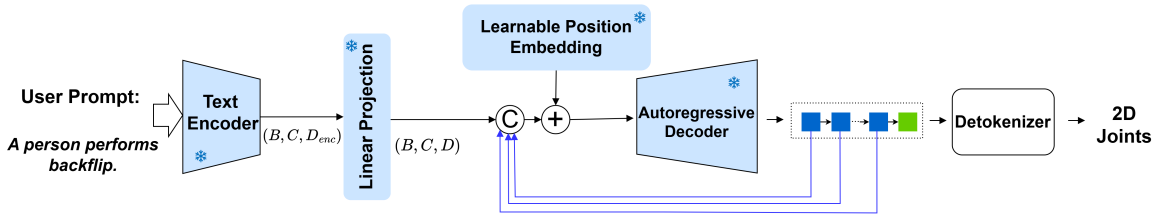

图2. 文本到骨架生成架构的推理概览。

为简化记号，下文将坐标编号 $T \times J \times 2$ 记为 M。

离散坐标词元。 为了使用自回归 Transformer 对二维姿态序列进行建模，我们首先将连续姿态张量转换为离散词元ID。给定归一化坐标 $u \in [0, 1]$，我们按如下方式将其离散化为 K 个分箱

$$q(u) = \lceil u \cdot (K-1) \rceil, \qquad q(u) \in \{0, \dots, K-1\}. \quad (1)$$

我们为特殊符号保留前四个词元ID：PAD =0、BOS = 1、序列结束标记 = 2，以及一个保留符号。所有身体坐标词元均按偏移量 $o = 4$向上平移：

$$s(u) = q(u) + o, \qquad s(u) \in \{4, \dots, K+3\} \quad (2)$$

从而得到总词表大小为 $V = o + K$。这遵循离散轨迹建模中的标准做法，即辅助词元占据一段保留区间，不会与数据词元 [55]发生冲突。在每个坐标都已映射为离散词元之后，我们接下来说明如何将这些词元排列成序列，以进行自回归建模。

一维词元流。 我们将二维姿态张量序列化为一维词元流$\mathbf{z} = [z_1, \dots, z_M]$，其长度为 $M = 2JT$，其中每个关节为其 x 和 y 坐标贡献两个连续词元。我们采用帧主序关节次序，使模型在时间推进之前先完成整个身体姿态，从而确保每一帧内所有关节之间的空间连贯性：对于每一帧，各关节按固定的骨架顺序出现，每个关节贡献连续的（x，y）词元。所得序列为：

$$\mathbf{z} = \big[s(x_{1,1}), s(y_{1,1}), \dots, s(x_{1,J}), s(y_{1,J}), \\ \dots, s(x_{T,J}), s(y_{T,J}) \big]. \quad (3)$$

对于自回归训练，我们在序列前添加一个BOS词元，并在末尾附加一个序列结束标记，以构成完整序列：

$$\mathbf{z}^{\text{full}} = [\text{BOS}, z_1, \dots, z_M, \text{EOS}]. \quad (4)$$

2) 文本条件

给定自然语言动作描述 c （例如，“一个人做出飞膝拳击”），我们使用来自稳定扩散v2 [56]的冻结CLIP文本编码器对其进行编码。对 c 进行词元化并使其通过编码器，可生成由 C 个上下文嵌入组成的序列

$$\mathbf{h}^{\text{text}} = [h_1^{\text{text}}, \dots, h_C^{\text{text}}] \in \mathbb{R}^{C \times D_{\text{enc}}}.$$

这些嵌入会被投影到 Transformer 的隐藏维度，并进行归一化：

$$\mathbf{e}^{\text{cond}} = \text{LN}\left(\mathbf{h}^{\text{text}} W_{\text{cond}}\right), \quad (5)$$

其中， $W_{\text{cond}} \in \mathbb{R}^{D_{\text{enc}} \times D}$ 是一个经过学习的线性投影，LN(·)表示层归一化。所得序列$\mathbf{e}^{\text{cond}} \in \mathbb{R}^{C \times D}$ 会被添加到姿态词元序列的开头，并通过因果自注意力机制对后续每个词元保持可见，从而确保文本条件控制在整个生成过程中持续发挥作用。

3) 自回归解码器

为了在单一自回归模型中同时处理文本词元和姿态词元，我们使用学习得到的嵌入表 $E_{\text{pose}} \in \mathbb{R}^{V \times D}$对$\mathbf{z}^{\text{full}}$ 中的每个词元进行嵌入，其中 V 是式(2)中定义的词表大小， D 是Transformer的隐藏维度。由此得到姿态嵌入序列$\mathbf{E}^{\text{pose}} \in \mathbb{R}^{(M+2) \times D}$，其中 $M+2$ 计入了 M 个身体坐标词元以及BOS词元和序列结束标记。

我们沿序列维度将文本条件序列与姿态嵌入进行拼接：

$$\mathbf{H}^{(0)} = [\mathbf{e}^{\text{cond}}, \mathbf{E}^{\text{pose}}] \in \mathbb{R}^{(C+M+2) \times D}. \quad (6)$$

我们将学习得到的位置嵌入添加到$\mathbf{H}^{(0)}$，并将结果输入由 L 个仅含解码器的 Transformer 模块组成的堆栈。模块采用因果自注意力 [55]，生成隐藏状态 $\mathbf{H}^{(L)} \in \mathbb{R}^{(C+M+2) \times D}$。前 C 个位置对应文本前缀，其余 $M+2$ 个位置对应 BOS 、身体标记和序列结束标记。

线性输出头将最终隐藏状态映射为词汇表上的逻辑值：

$$\mathbf{O} = \mathbf{H}^{(L)} W_{\text{lm}}^{\top} \in \mathbb{R}^{(C+M+2) \times V}, \quad (7)$$

其中 $W_{\text{lm}} \in \mathbb{R}^{V \times D}$。沿词汇表维度应用 softmax，可得到以文本为条件的自回归分布：

$$p_\theta(\mathbf{z}^{\text{full}} \mid c) = \prod_{i=1}^{M+1} p_\theta(z_i \mid \mathbf{e}^{\text{cond}}, z_{<i}), \quad (8)$$

其中 z_1，……，z_M是身体坐标令牌， z_{M+1} ≙是序列结束标记， $z_{<1}$ 仅由 BOS 词元构成。文本嵌入 $\mathbf{e}^{\text{cond}}$ 在整个生成过程中充当固定的条件前缀。

4) 训练目标

训练期间，我们采用教师强制，以真实先前标记而非模型预测作为条件。我们只对身体标记和最终的序列结束标记进行监督；文本前缀和 BOS 词元仅用作上下文，不参与损失计算。

训练目标是标准的下一个词元交叉熵：

$$\mathcal{L}_{\text{text2pose}} = -\sum_{i=1}^{M+1} \log p_\theta\left(z_i \mid \mathbf{e}^{\text{cond}}, z_{<i}\right), \quad (9)$$

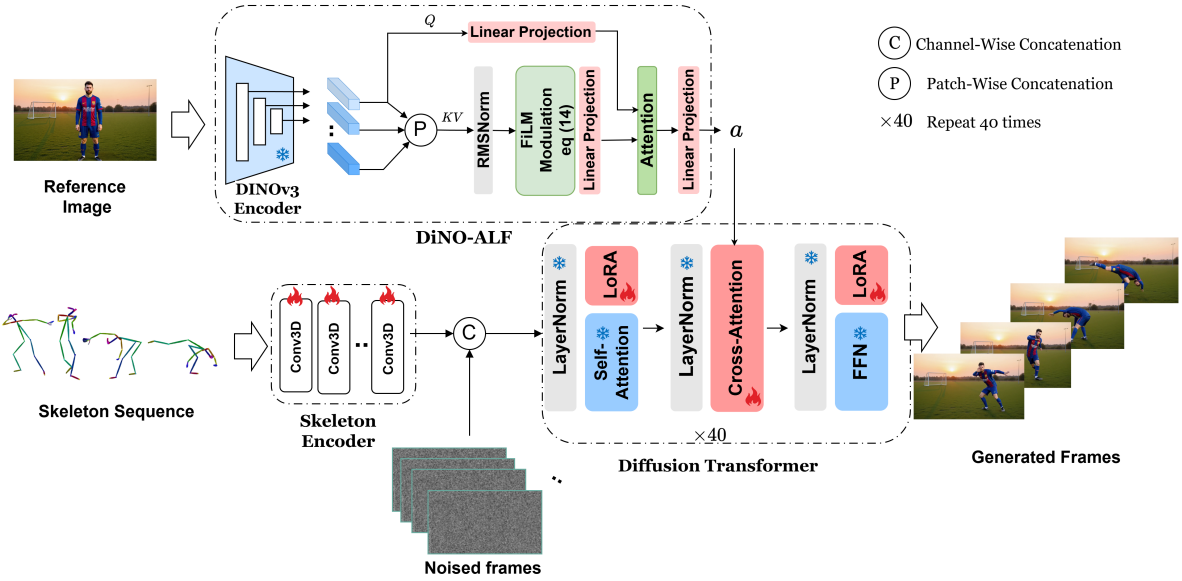

图 3. 姿态条件视频生成架构概览。参考图像通过 DINO-ALF 编码，生成外观令牌；骨架序列经光栅化后，由 3D 卷积神经网络编码为在时空上对齐的运动令牌。两路条件流均注入 DiT 去噪器，以合成输出视频。

其中 z_1，……，z_M 是身体坐标令牌，z_{M+1} =是序列结束标记。

5）推理

推理过程如图 2 所示，模型以自回归方式预测下一个词元，直至生成序列结束标记词元。

B. 姿态条件视频生成

第二阶段根据两个输入合成视频：一张指定应出现谁的参考图像 I_{ref}，以及一个指定身体应如何运动的骨架序列 $\hat{\mathbf{P}} \in \mathbb{R}^{T\times J\times 2}$，其中$\mathbf{P}$ 由第 III-A 节所述的文本到骨架模型生成。输出视频中，参考角色将执行所需的复杂运动（图 3）。

由于$\mathbf{P}$是预测得到的，而非真值，因此骨架生成阶段的误差可能会传播到视频合成中。为提升鲁棒性，我们在训练期间对真值骨架应用随机增强，每种增强都模拟生成阶段的一种典型失效模式：(i) 关节抖动——每个坐标添加高斯噪声（σ=3像素）；(ii) 关节丢弃——每个关节以 p_j=0.05 的概率归零；以及 (iii) 时间偏移——位置偏移 ±1 帧。

遵循先前的姿态引导人体动画研究 [20],[44], [21], ，我们将$\mathbf{P}$ 栅格化为逐帧二维姿态控制图像，绘制关节及其骨架连接，从而生成在空间上与视频帧对齐的图像序列。我们以预训练的 Wan2.1 TI2V 扩散骨干网络 [14]为基础，并将其调整用于姿态驱动生成。由于现在由骨架序列而非文本明确指定运动，我们以基于姿态的条件控制取代文本交叉注意力路径。

我们的设计引入了三个关键组件：（i）DINO-ALF，一种多层次的 DINOv3 [57] 外观编码器，用于提取空间局部化的图块特征，以在大幅度姿态变化下保留服装、纹理和局部身体细节；（ii）一种改进的条件控制接口，用完全可训练的 DINO-ALF 交叉注意力替换基于 CLIP 的参考交叉注意力，并结合选定 DiT 自注意力层与多层感知机层上的 LoRA 适配器；（iii）一种时空对齐运动编码器，将渲染的姿态序列映射为上下文令牌，以提供显式运动引导。

1) 潜空间视频扩散骨干网络

令 $\mathbf{v} \in \mathbb{R}^{3\times T\times H\times W}$ 表示包含 T帧的训练视频片段。预训练时空变分自编码器将$\mathbf{v}$ 编码为潜在标记 $\mathbf{x}_0 \in \mathbb{R}^{C_v\times T'\times H'\times W'}$，其中 C_v为潜在通道维度，(T', H', W') 表示在时间和空间维度下采样后的分辨率。我们遵循标准的扩散训练流程：采样时间步 τ和噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，并构造含噪潜变量

$$\mathbf{x}_\tau = \alpha_\tau\, \mathbf{x}_0 + \sigma_\tau\, \boldsymbol{\epsilon}. \tag{10}$$

DiT 去噪器 $\epsilon_\theta(\cdot)$在外观和运动控制的条件下预测噪声（或根据调度器预测速度）：

$$\hat{\boldsymbol{\epsilon}} = \epsilon_\theta\Big(\mathbf{x}_\tau, \tau \mid I_{\text{ref}}, \hat{\mathbf{P}}\Big), \tag{11}$$

并最小化加权均方误差目标

$$\mathcal{L}_{\text{vid}} = \lambda(\tau)\, \big\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\big\|_2^2, \tag{12}$$

其中，$\lambda(\tau)$是依赖于调度器的损失权重。

2) DINO-ALF：用于外观编码的自适应层融合

预训练骨干网络通过拼接 CLIP 图像/文本标记，并经由交叉注意力将其注入 DiT 模块，从而以参考图像和文本作为生成条件控制。然而，在伴随大幅变形和自遮挡的复杂运动中，这种全局 CLIP 式条件控制往往不足以保留细粒度外观线索（例如，

服装纹理和局部身体部位），从而导致明显的偏移 [28], [29], [30]。为提供更强且具有空间定位性的参考线索，我们从冻结的 DINOv3 编码器中提取多层图像块描述符，并对其进行自适应融合，再将其作为外观令牌注入以进行条件控制。

多层提取。 令 $\{\mathbf{h}^{(\ell)}\}_{\ell=1}^{L_D}$ 表示来自所有 L_D=12 DINO Transformer 层（图块大小为 16×16）的隐藏状态，并令 $\mathbf{p}^{(\ell)} \in \mathbb{R}^{N\times d_D}$ 表示对应的图块令牌（移除 CLS/注册令牌），其中 N 为图块数量，d_D 为编码器的嵌入维度。

自适应层融合。 不同的 DINO 层捕获互补信息：如图 4 所示，较早的层在纹理丰富的区域表现出较高激活，而较晚的层产生更加均匀的响应，因此我们选用早期层作为自适应融合的查询。我们不选择单个层，而是通过交叉注意力模块 A_γ学习聚合每个图块的各层特征。首先，我们将逐层的图块令牌堆叠起来：

$$\mathbf{P} = [\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(L_D)}] \in \mathbb{R}^{N\times L_D\times d_D}. \tag{13}$$

为稳定训练，并使模型能够强调或抑制各个层，我们采用 FiLM 风格调制：

$$\mathbf{P}'_{:,\ell,:} = \mathrm{Norm}(\mathbf{P}_{:,\ell,:}) \odot (1+\boldsymbol{\beta}_\ell) + \boldsymbol{\delta}_\ell, \tag{14}$$

其中，每个层 ℓ均具有可学习的 $(\boldsymbol{\beta}_\ell,\ \boldsymbol{\delta}_\ell)$。随后，交叉注意力聚合调制后的特征：

$$\tilde{\mathbf{p}} = A_\gamma(\mathbf{q}, \mathbf{P}') \in \mathbb{R}^{N\times d_D}, \tag{15}$$

其中，$\mathbf{q} \in \mathbb{R}^{N\times d_D}$ 是由第一层（ℓ=1）得到的查询。最后，我们将聚合后的词元投影到 DiT 的隐藏维度：

$$\mathbf{a} = p_\eta(\tilde{\mathbf{p}}) \in \mathbb{R}^{N\times d}, \tag{16}$$

其中，p_η 是一个带有层归一化的多层感知机。DINO 编码器保持冻结；仅训练 A_γ 和 p_η。如图 5 所示，生成的 DINO-ALF 交叉注意力仍集中于人体主体，而 CLIP 交叉注意力则分散在整个图像上，从而实现运动与外观之间更好的对齐。

3）将原生参考交叉注意力替换为 *DINO* 交叉注意力

Wan2.1 通过每个 DiT 模块中的两条交叉注意力路径注入条件控制：一条关注 CLIP 文本令牌，以引导运动和场景演变；另一条关注 CLIP 图像令牌，以控制外观。在我们的姿态驱动设置中，运动已由骨架序列明确指定，因此我们禁用这两条基于 CLIP 的路径，并将其替换为单一的 DINO 条件交叉注意力，使其关注 DINO-ALF 外观令牌 $\mathbf{a}$，从而提供空间定位更准确的外观线索，在大幅度姿势变化下更好地保留服装和精细细节。

4）姿态控制作为时空对齐的运动令牌

直接将原始 2D 关键点注入高容量视频去噪器并不可行，除非条件控制与潜在时空网格对齐。因此，我们按照既定方式将骨架序列光栅化为 RGB 姿态控制视频$\mathbf{s} \in \mathbb{R}^{3\times T\times H\times W}$（例如骨架图／关键点连线），

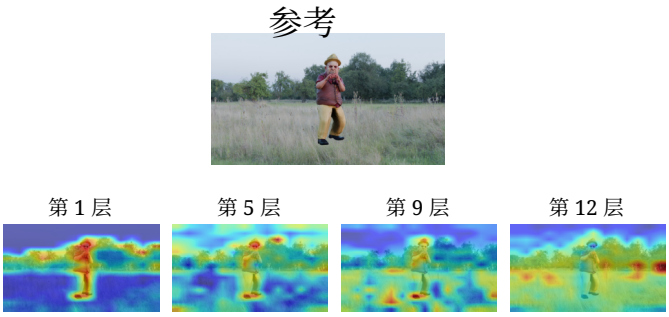

图 4. DINOv3 各层的图块特征幅值图（ℓ_2-范数）。较早的层在主体和纹理丰富的区域呈现较高激活，而较晚的层的幅值则更均匀。这促使我们使用早期层作为自适应层融合的查询。

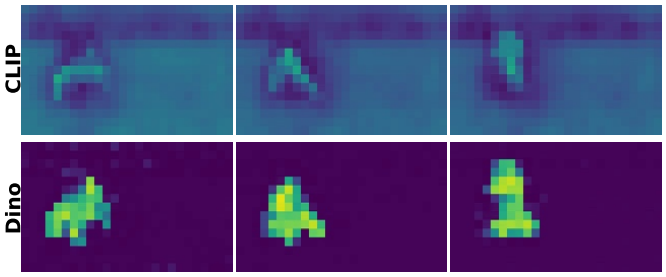

图 5. 后空翻序列中 CLIP（上）与 DINO-ALF（下）的交叉注意力图。DINO-ALF 对运动主体的关注更精准，而 CLIP 的注意力则较为分散。

并遵循 [58], 的渲染设置，使用三维卷积神经网络运动编码器 g_ϕ 对其进行编码：

$$\mathbf{m} = g_\phi(\mathbf{s}), \qquad \mathbf{m} \in \mathbb{R}^{N_m\times d}. \tag{17}$$

在我们的实现中，g_ϕ 使用步幅三维卷积对$\mathbf{s}$进行下采样，使其与变分自编码器潜变量具有相同的（T'，H'，W'）网格，并通过展平时空网格生成词元序列（因此 $N_m = T'H'W'$）。我们将词元通道维度设为与 DiT 隐藏维度相匹配，从而使$\mathbf{m}$ 能够作为一条一等条件流注入。

5) DiT 去噪器中的条件控制接口

我们通过 (i) 运动令牌 $\mathbf{m}$和 (ii) DINO 外观标记 $\mathbf{a}$对 DiT 去噪器进行条件控制。运动令牌提供时空对齐的引导，而 DINO 交叉注意力分支则注入局部外观线索。总体而言，去噪器的运作方式如下：

$$\hat{\boldsymbol{\epsilon}} = \epsilon_\theta\Big(\mathbf{x}_\tau, \tau \mid \underbrace{\mathbf{m}}_{\text{motion}}, \underbrace{\mathbf{a}}_{\text{DINO-ALF}}\Big), \tag{18}$$

其中，I_{ref} 仅通过提取的 ˆ 令牌 $\mathbf{a}$影响生成，而姿态序列 $\mathbf{P}$ 仅通过渲染的姿态控制 $\mathbf{s}$ 及其编码 $\mathbf{m}$发挥作用。

6) 使用 *LoRA* 进行训练与适配

对大型视频 DiT 进行全面微调成本高昂，而且在有限的复杂运动数据上容易过拟合。因此，我们冻结预训练骨干网络，仅训练适配模块：(i) 在选定的自注意力和多层感知机投影中使用 LoRA 模块，(ii) DINO 交叉注意力分支，(iii)

运动编码器 g_ϕ，以及 (iv) DINO 聚合/投影头（A_γ 和 p_η）。对于 DiT 内部的线性投影 $\mathbf{W}$（例如，q、k、v、o或多层感知机层），LoRA 会替换

$$\mathbf{W}\mathbf{x} \;\mapsto\; (\mathbf{W} + \tfrac{\alpha}{r}\mathbf{B}\mathbf{A})\mathbf{x}, \tag{19}$$

其中$\mathbf{A} \in \mathbb{R}^{r\times d_{\text{in}}}$ 和$\mathbf{B} \in \mathbb{R}^{d_{\text{out}}\times r}$ 是可训练的低秩矩阵，r 是秩，α 是缩放因子。

7) 用于提升稳健性的条件丢弃

为减少对任何单一条件控制通道的过度依赖并提升泛化能力，我们在训练期间应用丢弃：以较小的概率将运动令牌$\mathbf{m}$和/或 DINO 外观标记$\mathbf{a}$置零。这促使模型在显式运动控制与外观保持之间分配职责，从而在具有挑战性的运动下产生更稳定的生成结果。

a) 推理。

在测试时，我们（i）根据文本生成骨架序列（第 III-A 节），（ii）将其栅格化为姿态控制视频$\mathbf{s}$ 并编码为运动令牌$\mathbf{m}$，（iii）从 I_{ref}中提取 DINO 令牌$\mathbf{a}$，（iv）使用训练好的条件控制模块在潜空间中运行扩散采样器。遵循 [21]，我们在视频合成之前应用姿态重定向后处理步骤，调整骨骼长度并细化细微的关节位置，使其与参考图像匹配。最终视频通过使用预训练的 VAE 解码器对去噪后的潜变量进行解码而获得。

IV. 实验

我们通过首先评估文本到骨架生成质量，随后评估姿态条件视频合成，最后报告两阶段中关键设计选择的消融研究，来评估我们的两阶段框架。

A. 数据集

1) 文本-姿态数据集

对于文本到骨架评估，我们共使用了 8,000 对文本–姿态序列，数据来自：(i) 我们内部生成的、由 Blender 渲染的复杂运动视频；以及 (ii) HumanDreamer [45] 采用的 *Fitness*类别（来源于 Motion-X [59]）。其中，2,000 个序列来自我们由 Blender 渲染的合成视频，其文本提示直接取自 Mixamo 的运动名称（例如，将“后空翻”运动描述为“一个人完成后空翻”）；其余序列来自 Motion-X Fitness 子集。我们在序列层面随机划分数据，其中 90% 用于训练，10% 作为留出的测试集，并在测试集上报告所有评估结果。在按照第 III-A 节所述进行离散化之前，每个运动均表示为包含 $J{=}62$ 个关节、与视角对齐的二维姿态序列。

2）合成视频数据集

据我们所知，目前没有公开可用的视频数据集专门聚焦于适用于视频生成研究的复杂人体动作。TikTok 数据集 [60] 是一个公开可用、无版权限制的人类视频基准测试，但其关注的是舞蹈和上半身动作，而非杂技动作。UCF 101 [61], HMDB51 [62], KTH [63],和 NTU RGB+D [64] 等人体动作识别数据集并不适用，原因有二：它们存在分辨率低、压缩伪影和视觉噪声等问题，会影响现代视频生成模型的微调；而且其动作类别局限于简单、孤立的动作（例如拍手、跳跃、挥手），而非后空翻、侧手翻和武术踢击等复杂杂技动作序列。以往的数据集通常依赖网络来源视频 [17], [60]。此外，即使是*Sora* [15] 和 *Kling* [16],等先进的闭源生成器，也难以生成复杂动作，因此无法可靠地用于构建这一领域的高质量基准测试。为避免网络来源数据固有的潜在版权和隐私问题，并解决上述局限，我们使用*Blender* 构建了一个合成数据集，从而能够全面控制角色、摄像机设置和环境。

我们从 Mixamo [65] 中整理出一组丰富多样的复杂动作 FBX动作，并将其与 Mixamo 人类角色和皮肤搭配，以增加外观的多样性。为丰富场景变化，我们使用 Poly- Haven [66] 提供的 HDRI环境贴图作为背景和照明。在 Blender 中，我们通过放置动画角色、选择摄像机设置，并使用一致的参数进行渲染，组装每个场景。最终的数据集包含 2,000 段合成视频，涵盖不同角色和环境中的杂技及特技动作。该数据集的规模超过官方 TikTok舞蹈数据集（340 段视频）；且由于完全由合成内容构成，不会引入网络采集视频所涉及的隐私和同意问题 [60]。我们在项目页面（数据集部分）提供了若干具有代表性的视频样例。

B. 文本到骨架评估

基线方法。 我们使用 Human- Dreamer [45] 的公开实现，将我们的方法与其进行比较；同时采用相同的基线方法集合及适配/评估流程，以确保与以往文本到运动领域的研究进行一致的比较。具体而言，我们评估了以下基线方法：T2M-GPT [67] （由两阶段 VQ-VAE标记器和一个以文本为条件的 GPT风格自回归变换器组成）、PriorMDM [68]（利用预训练动作扩散模型作为生成先验的文本条件扩散模型），以及 MLD [69] （将动作VAE与在学习到的动作潜在空间中进行的条件潜在扩散相结合）。

评估指标。遵循标准的文本到运动协议 [70], [71], [72], [73], ，所有指标均在 [45] 所定义的学习型文本–姿态嵌入空间中进行评估。我们报告了：**(i) FID(↓)**：衡量生成运动与真实运动之间的分布相似度，以评估视觉真实感；**(ii) R-精确率** top-k (k=1、2、3；↑)：通过衡量在给定生成运动的情况下，从一组干扰项中正确检索出真实文本的频率，量化语义准确度；**(iii) 多样性(↑)**：衡量生成样本之间的平均几何距离，以确保整个数据集中的动作具有广泛多样性；以及 **(iv)MM-Dist(↓)**：计算文本与动作特征之间的多模态距离，以反映提示词遵循度。

实现细节。 我们使用 K=256个分箱和偏移量 o=4进行词元化，并采用top-k 采样（k=10）进行解码。我们使用AdamW进行训练，学习率为 10^{-5}，权重

表 I 我们的 8,000 对基准测试上文本到骨架生成的定量评估。↓ 表示越低越好，↑ 表示越高越好。

方法	FID↓	Rp-top1↑	Rp-top2↑	Rp-top3↑	多样性↑	MM-Dist↓
T2M-GPT [67]	524.61	0.191	0.287	0.473	40.11	49.85
PriorMDM [68]	585.31	0.216	0.325	0.501	42.58	44.29
MLD [69]	467.22	0.335	0.503	0.653	41.67	47.66
HumanDreamer [45]	322.16	0.411	0.598	0.722	45.33	41.53
Ours	**255.19**	**0.487**	**0.667**	**0.784**	**48.33**	**38.65**

衰减为0。01，$\beta =$（0。9，0。95），批量大小为8，使用单个 NVIDIA A100 图形处理器进行训练。

定量分析。如表 I 所示，我们的方法在所有评估指标上均优于现有方法。我们模型的 FID 达到 255.19，优于此前由 HumanDreamer 保持的最佳成绩 322.16。较低的 FID 分数表明，我们生成的骨架序列与真实人体动作的分布更加接近，这意味着物理合理性和视觉真实感均有所提升。此外，我们的方法取得了最高的多样性分数（48.33），表明它能够生成种类更丰富的不同动作，并避免了模式崩溃这一常见问题，即模型生成重复的运动。

在语义对齐和指令遵循方面，R-精度分数表现出优势；我们的 top-1 准确率达到 0.487，这意味着在包含多个干扰项的候选集中，正确的文本描述近乎有一半的概率会被检索为最佳匹配。多模态距离（MM-Dist）也印证了这一点，我们的模型取得了最低分 38.65。综合来看，这些结果表明，我们的方法在输入提示的语言细微差异与生成的姿态序列之间建立了更精确的映射。Rp-top1、top-2 和 top-3 均持续提升，进一步证实了我们的模型能够准确、稳健地遵循用户提示。

定性分析。图 6 展示了提示“A person performs a cartwheel”（一个人做侧手翻）所生成姿态序列的定性比较。我们的方法生成的轨迹更加连贯，也更符合物理规律，尤其是在具有挑战性的倒立阶段和快速过渡期间。相比之下，HumanDreamer 经常生成不合理的头部和腿部姿态，如红框所示。MLD 在若干帧中也未能生成合理的头部和腿部姿态（红框）。PriorMDM 则出现了不自然的腿部姿态、身体部位过长等伪影（红框），而 T2M-GPT 混淆了姿态顺序，并频繁生成扭曲的四肢（红框）。其他提示的更多定性结果请见我们的项目页面。

C. 姿态到视频评估

基线方法。我们将本文方法与第 II 节中讨论的最先进姿态条件人类视频生成方法进行了比较：Human- Vid [38], MimicMotion [41], Hyper-Motion [17], UniAnimate- DiT [37], 和 VACE [35]。这些方法代表了姿态条件生成的主要架构范式，包括 ControlNet 风格适配器，置信度感知引导、空间 RoPE 设计和基于 DiT 的条件控制。

实现细节。我们基于预训练的 Wan2.1 [14] I2V 14B 视频扩散模型初始化，并在 4 个 NVIDIA A100 图形处理器上进行所有实验。所有视频均包含 81 帧。我们以 2×10^{-5} 的学习率对模型进行 40,000 步微调。训练使用包含 2,000 个视频的 Blender 合成数据集，并采用参数高效的 LoRA 微调策略，秩设为 64（我们微调 LoRA 适配器以及第 III-B 节中引入的额外条件控制模块）。我们按视频对数据集进行随机划分，其中 1,800 个视频用于训练，200 个用于评估，确保训练数据与测试数据之间没有重叠。

评估指标。我们报告了视频生成中使用的帧级图像质量指标，包括 SSIM [74]（结构相似性）、LPIPS [75]（深度特征空间中的感知距离）和 PSNR [76]（逐像素重建保真度），以及视频级 Fr´echet 视频距离（FVD） [77], , 该指标用于衡量真实视频与生成视频在学习得到的视频特征空间中的分布距离。此外，我们还使用了细粒度的 VBench-I2V [78] 指标：主体一致性（通过预训练视觉特征相似度衡量外观保持）；背景一致性（通过 CLIP 特征相似度衡量背景保持）；时间闪烁（通过帧间差异衡量时间闪烁）；运动平滑度（通过帧插值重建误差衡量短期动态）；动态程度（通过 RAFT [79] 光流衡量非静态视频的比例）；美学质量（LAION 美学评分）；以及成像质量（MUSIQ [80] 评分，用于评估低级失真）。

定量结果。表 II 将我们的方法与最先进的基线方法在 VBench-I2V 指标、帧级质量指标和视频级 FVD 上进行了比较。我们的方法在七项 VBench-I2V 指标中的五项上表现最佳，其余两项的得分也具有竞争力。具体而言，与表现最强的基线方法（VACE）相比，我们在主体一致性上高出 +3.0（91.31 对 88.31），在背景一致性上高出 +2.68（93.79 对 91.11），并在运动平滑度上高出 +2.58（97.39 对 94.81）。这些提升表明，我们的 DINO-ALF 外观编码能够更好地保留主体身份和背景连贯性，而结构化姿态条件控制则生成了更平滑的运动轨迹。动态程度提升 +3.47（78.50 对 75.03），进一步证实我们的方法生成的视频具有更丰富、非静态的运动，而非简单的静态解。

在帧级指标方面，我们取得了最高的结构相似性指数 SSIM

表 II 不同方法在 VBench 上的比较。(↑) 越高越好；(↓) 越低越好。

模型	VBench-I2V							帧级			视频级
	主体一致性↑	背景一致性↑	运动闪烁。↑	运动平滑度。↑	动态程度↑	美学质量↑	图像质量↑	结构相似性指数↑	LPIPS↓	峰值信噪比↑	FVD↓
HumanVid [38]	85.33	86.91	91.01	88.21	72.72	54.61	64.19	0.737	0.213	25.54	637.1
MimicMotion [41]	86.01	88.47	92.11	89.95	73.05	55.07	65.14	0.744	0.198	26.66	611.5
Hyper-Motion [17]	86.66	90.23	93.22	90.08	73.65	55.21	65.42	0.767	0.181	27.13	568.8
UniAnimate-DiT [37]	87.21	91.17	95.78	92.33	74.87	56.32	66.87	0.751	0.189	27.48	527.4
VACE [35]	88.31	91.11	95.31	94.81	75.03	**57.11**	**67.91**	0.771	**0.173**	27.91	498.7
Ours	**91.31**	**93.79**	**97.50**	**97.39**	**78.50**	56.37	67.14	**0.795**	0.174	**28.50**	**471.3**

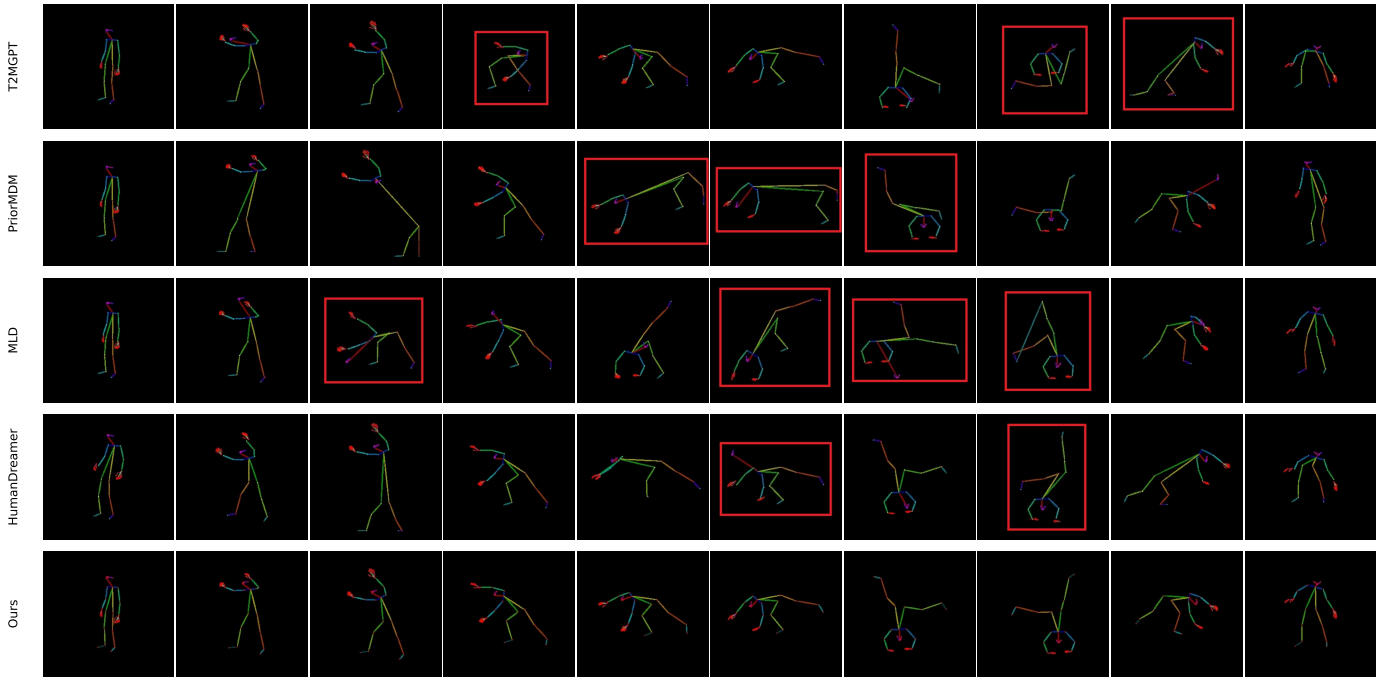

图 6. 提示词“A person performs a cartwheel”对应的生成二维姿态序列定性比较。我们的方法生成的轨迹更加连贯且符合物理规律，尤其是在倒立阶段。竞争方法的失败案例以红框标出。为便于观察，我们裁剪场景，仅展示人体骨架。

（0.795）和峰值信噪比 PSNR（28.50），LPIPS（0.174）也具有竞争力（VACE 为 0.173）。在视频层面，我们的 FVD 为 471.3，是最低的，这表明生成的视频与真实视频分布更为吻合。尽管 VACE 的美学质量 略高（57.11 对 56.37），且成像质量 也略高（67.91 对 67.14），但我们的方法在时间一致性和运动真实感方面表现显著更优。这些结果表明，我们的级联设计——将运动控制与外观合成解耦——能够在不牺牲视觉质量的前提下改善运动动态。

定性结果。 图 7 展示了我们的方法与基线方法之间的视觉对比。我们的方法在各种运动中都能紧密遵循目标二维骨架，同时保留参考外观。相比之下，所有基线方法在复杂姿势下都出现了伪影。手部生成尤其具有挑战性：HumanVid [38] 在早期和晚期帧中生成了错误的手部，Hyper-Motion [17] 在旋转动作中失效，UniAnimate-DiT [37]遗漏了双手的部分区域，而 VACE [35] 错误地渲染了右手（均以红框标出）。此外，MimicMotion [41] 在踢腿动作中还遗漏了一条腿。我们的方法避免了这些问题，因为它利用了 DINO-ALF，能够更好地关注精细的外观细节——例如，红色领带始终得以保留（对比第三列和第八列）。更多结果请参见我们的项目页面。

消融实验。 我们在表 III 中对关键的架构设计和推理选项进行了消融，考察了词元化粒度（K）、解码器深度（L）和解码策略。

（1）词元化粒度。 提高 K 可通过减少离散化误差来提升真实感以及文本与运动之间的对齐效果，但过细的词元化会导致性能略有下降，这是因为词汇表更大且更稀疏，因而更难建模。我们使用K=256。

（2）解码器深度。更深的解码器能够更好地捕捉长程时间结构和全局身体协调能力，但随着深度增大，性能提升趋于饱和。我们使用 L=18。

（3）解码策略。贪心解码（每一步取 argmax）具有过强的确定性，会降低多样性，并且常常传播早期错误。随机采样（核采样 / Top-k）在质量与多样性之间取得了更好的平衡；我们使用带有 k=10的 Top-k 。

我们还对姿态到视频生成阶段进行了消融研究，以分析关键设计选择。我们首先对外观编码设计进行了消融，

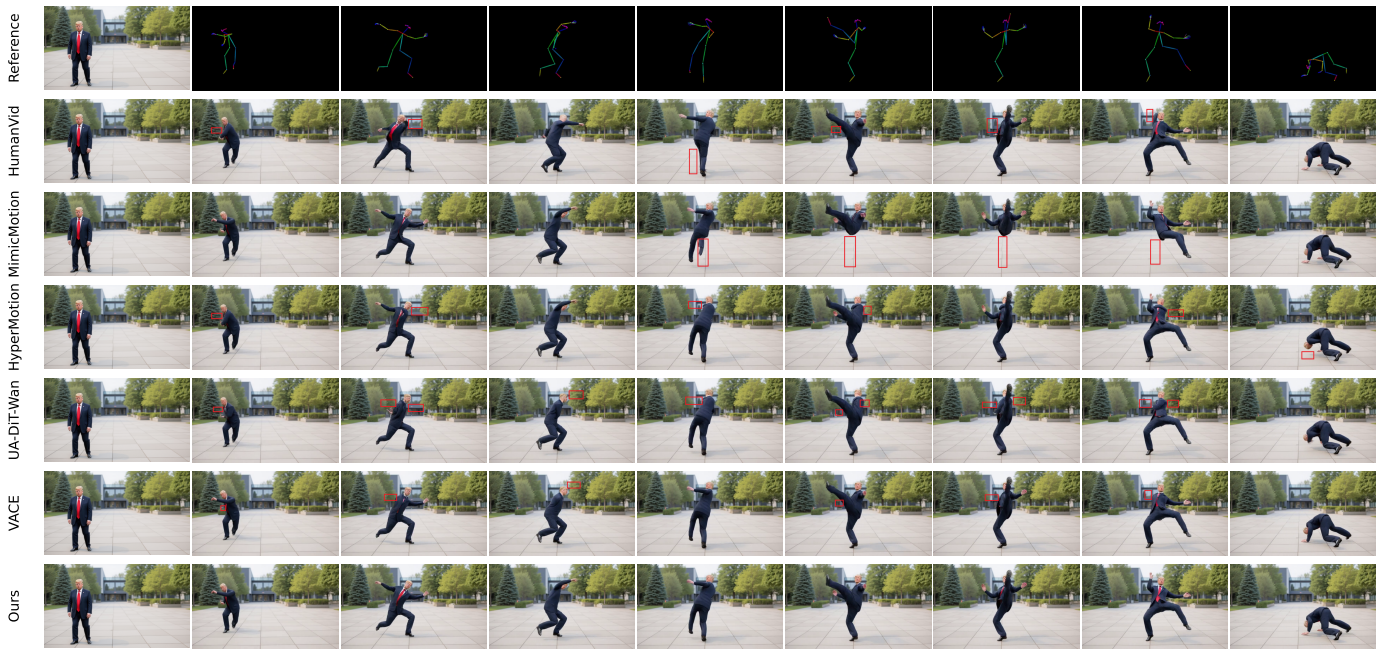

图 7. 与姿态引导基线方法的定性视频比较。我们的方法在保留参考外观的同时，最能遵循目标二维骨架。其他方法在复杂动作下会出现手部/腿部伪影（红框），而我们的 DINO-ALF 条件控制能更好地保留精细细节（例如红色领带）。更多结果请见我们的项目页面。

表 III 文本到骨架架构的消融实验。我们分析了分词粒度（K）、解码器深度（L）和解码策略的影响。↓表示越低越好；↑表示越高越好。

配置	FID↓	Rp-top1↑	多样性↑	MM-Dist↓
分词粒度（K 分箱）				
$K = 64$	312.41	0.401	44.21	43.17
$K = 128$	278.55	0.452	46.88	40.52
$K = 256$ (基线)	**255.19**	**0.487**	**48.33**	**38.65**
$K = 512$	261.33	0.479	47.65	39.21
解码器深度（L层）				
$L = 6$	301.55	0.418	44.77	43.52
$L = 12$	272.38	0.459	46.91	40.33
$L = 18$ (基线)	**255.19**	**0.487**	**48.33**	**38.65**
$L = 24$	258.91	0.481	47.88	39.01
解码策略				
贪心解码	289.66	0.441	42.15	41.77
核采样（p=0.9）	267.81	0.468	47.22	39.88
前-k （k=10，基线）	**255.19**	**0.487**	**48.33**	**38.65**

然后在表 IV 中分析了网络参数。

DINO-ALF 的效果。我们首先研究了使用 DINO-ALF 进行参考外观条件化的影响。我们没有聚合多个 DINO 层的特征，而是评估了一种常见变体，该变体仅使用 DINOv3 最后一层（DLL）的特征。与该变体相比，在大幅度姿态变化下，DINO-ALF 始终能更好地保持主体。如图 8（第三行，DINO LL）所示，仅使用 DLL 时会遗漏精细运动细节，尤其是较小的手部（红框）。这证实了结合

表 IV 姿势到视频网络参数的消融实验。我们分析 LoRA 秩、运动编码器架构和条件丢弃率的影响。（↑）越高越好；（↓）越低越好。

配置	主体↑	背景↑	运动平滑度↑	结构相似性指数↑	LPIPS↓	FVD↓
LoRA 秩（r）						
$r = 16$	84.22	90.45	92.11	0.712	0.221	512.8
$r = 32$	88.77	91.58	95.88	0.758	0.193	491.5
$r = 64$ (基线)	91.31	**93.79**	**97.39**	**0.795**	**0.174**	**471.3**
$r = 128$	**91.33**	93.72	97.28	0.789	0.178	478.9
运动编码器架构						
二维卷积神经网络 + 时间注意力。	87.91	90.55	93.22	0.755	0.195	521.4
3D 卷积神经网络浅层（4 层）	89.22	91.88	95.11	0.771	0.185	498.2
3D 卷积神经网络深层（8 层，基线）	**91.31**	**93.79**	**97.39**	**0.795**	**0.174**	**471.3**
条件丢弃率（p_{drop}）						
$p_{\text{drop}} = 0$ （不丢弃）	89.11	92.05	95.22	0.775	0.182	492.1
$p_{\text{drop}} = 0.05$	90.22	92.88	96.33	0.785	0.178	481.7
$p_{\text{drop}} = 0.1$ (基线)	**91.31**	**93.79**	**97.39**	**0.795**	**0.174**	**471.3**
$p_{\text{drop}} = 0.2$	88.55	91.13	95.81	0.769	0.191	501.5

低级纹理线索与高级语义特征对于实现稳健的外观一致性至关重要。表 V 中的定量结果进一步支持了这一观察：DINO-ALF 提升了主体一致性和运动相关指标（运动平滑度和运动闪烁），表明参考注入更加稳定。

将 DINO 特征替换为 CLIP图像嵌入。接下来，我们将提出的基于 DINO 的外观编码器替换为 Wan2.1 原始架构中使用的默认 CLIP图像嵌入，并通过标准交叉注意力注入。该变体在复杂运动下的细粒度身份细节方面出现明显退化（图 8 第二行 CLIP，红框所示），包括手部部位缺失以及服装

表 V 外观编码设计消融实验。我们比较了 DINO-ALF（本文方法）、DLL（单层）、CLIP 嵌入以及无外观编码器。（↑）越高越好；（↓）越低越好。

模型	VBench-I2V							帧级			视频级
	主体一致性↑	背景一致性↑	运动抖动。↑	运动平滑度。↑	动态程度↑	美学质量↑	图像质量↑	结构相似性指数↑	LPIPS↓	峰值信噪比↑	FVD↓
DLL（单层）	87.65	90.03	94.11	93.77	74.13	55.11	66.23	0.752	0.181	27.02	502.5
CLIP 嵌入	85.44	89.75	93.63	91.21	73.02	54.76	66.08	0.741	0.195	26.98	517.7
无外观编码器	75.13	85.14	82.22	82.11	65.53	47.11	60.09	0.673	0.245	24.33	593.4
Ours	**91.31**	**93.79**	**97.50**	**97.39**	**78.50**	**56.37**	**67.14**	**0.795**	**0.174**	**28.50**	**471.3**

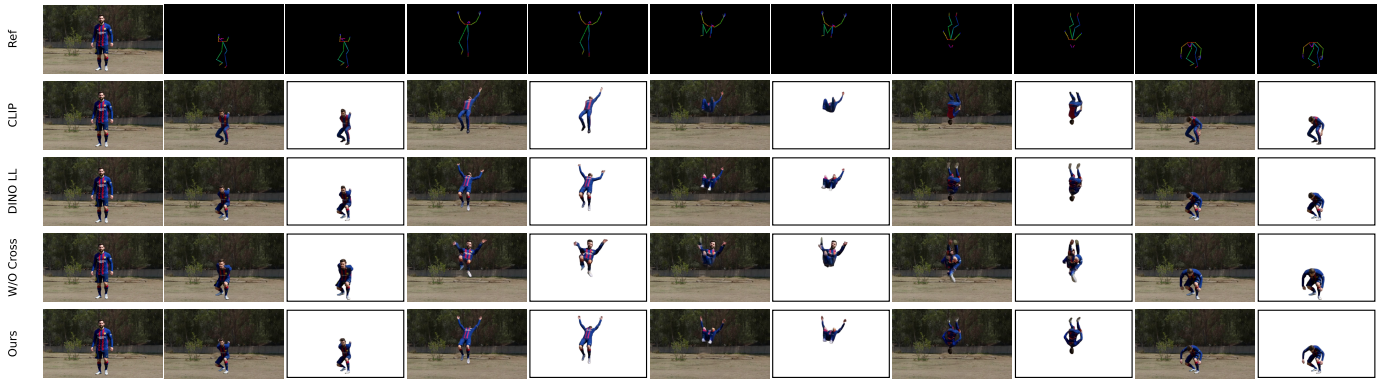

图 8. 姿态到视频阶段的消融实验。从上到下各行为：参考（参考帧和目标姿态）、CLIP（CLIP 嵌入）、DINOLL（DLL 单层）、无交叉（无外观编码器）以及我们的方法。DINO-ALF 在大幅度姿态变化下实现了最佳的外观保持，而仅使用 CLIP 或不使用外观编码器会导致手部/服装伪影和身份漂移（红框）。

逐渐变化（例如，短裤变成长裤），鞋子颜色也发生变化（从白色变为黑色）。这些结果表明，仅凭 CLIP 特征不足以在快速、复杂的运动下保持外观。这一点也体现在表 V 中：在主体一致性和运动指标（运动平滑度、动态程度和运动闪烁）以及帧级指标（结构相似性指数、LPIPS）方面，CLIP 的表现均逊于 DINO-ALF。

移除外观编码器。我们完全移除了外观编码器，方法是禁用额外的交叉注意力分支，仅保留预训练参考路径。这种设置导致严重的身份漂移，以及帧间外观不稳定。如图 8（第四行 W/O cross，红框）所示，在后空翻过程中，生成的视频不再可靠地遵循参考二维姿态。表 V 也显示，美学、运动和主体一致性指标均明显下降，证明 DINO-ALF 对可靠的姿态驱动视频生成至关重要。

网络参数敏感性。表 IV 对 LoRA 秩、运动编码器架构和条件丢弃进行了消融实验。对于 LoRA 秩，性能从 r=16 提升至 r=64 （主体一致性：84.22 → 91.31），但在 r=128 时趋于饱和。对于运动编码，深层三维卷积神经网络优于带时间注意力的二维卷积神经网络（FVD：471.3 对比 521.4），因为它能够捕捉与潜在网格对齐的局部时空动态。对于条件丢弃，p_{drop}=0.1 达到了最佳平衡；不进行丢弃会导致模型过度依赖某一条件流，而 p_{drop}=0.2 则使训练中的条件控制不足。所有实验均采用 r=64、深层三维卷积神经网络和 p_{drop}=0.1 。

通过姿态增强提升误差鲁棒性。表 VI 评估了视频模型由预测骨架（而非真值骨架）驱动时，姿态增强的效果。

表 VI 姿态增强对误差鲁棒性的消融实验。除第一行外，所有行在推理时均使用预测的骨架；真实标注行作为上界。(↑)越高越好；(↓)越低越好。

配置	主体↑	背景↑	M.Smooth↑	结构相似性指数↑	LPIPS↓	FVD↓
真实标注骨架（上界）	91.31	93.79	97.39	0.795	0.174	471.3
不使用数据增强	85.11	86.57	89.07	0.652	0.229	574.7
仅关节抖动	86.45	87.26	91.39	0.651	0.213	568.1
仅关节丢弃	86.57	87.73	91.11	0.683	0.208	541.5
仅时间偏移	87.31	88.05	92.71	0.713	0.193	528.3
完整增强（我们的方法）	89.23	91.17	95.61	0.781	0.183	492.5

没有数据增强时，视频模型对生成骨架中的伪影十分敏感，因为它只在干净的真值姿态上进行训练。每种数据增强都针对骨架生成阶段的一种特定失效模式：关节抖动缓解了空间坐标噪声，关节丢弃处理了漏检或未检测到的关节，而时间偏移则解决了帧级时序误差。三者结合取得了最佳结果，证实训练时数据增强对于实现鲁棒的级联生成至关重要。为单独评估骨架生成阶段引入的误差，表 VI 还列出了 GT骨架上界（取自表 II）；GT 与完整增强之间的差距量化了使用预测骨架所带来的代价。

五、局限性

当前框架旨在生成单人动作，且未对多人交互进行建模，例如

协调群体动作、身体接触或碰撞规避。扩展至多人场景需要联合建模人际空间关系及其交互，这将留待未来工作。此外，在快速旋转和杂技式过渡过程中，手指和面部特征等精细细节可能会丢失或变得模糊，因为姿态条件模型难以保留这些身体末端部位的高频外观线索。

VI. 结论

我们提出了一种用于可控复杂人体运动视频生成的级联框架，将运动规划与外观合成解耦。自回归文本到骨架模型根据自然语言生成二维姿态序列，而采用 DINO-ALF 的姿态条件视频扩散模型在大幅变形和自遮挡条件下保持外观。我们还引入了一个包含 2,000 个复杂运动视频的合成基准，以解决现有数据集中杂技动作代表性不足的问题。实验表明，我们的框架在文本到骨架和视频生成指标上均优于先前方法。

参考文献

[1] Z. Wang、B. Hu、M. Zhang、J. Li、L. Li、M. Gong 和 X. Gao，《基于扩散模型的视觉补偿引导和视觉差异分析用于无参考图像质量评估》，*IEEE*图像处理汇刊，第 34 卷，第 263–278 页，2025 年。[2] G. Fan、Y. Zhou、J. Zhou、Y. Ju、G.-Y. Chen、J. Li 和 A. C. Kot，《Dcd-uie：用于水下图像增强的解耦色度扩散模型》，*IEEE*图像处理汇刊，第 1–1 页，2026 年。[3] T. Bordin 和 T. Maugey，《用于低比特率扩散图像压缩的线性变换颜色引导》，*IEEE*图像处理汇刊，第 34 卷，第 468–482 页，2025 年。[4] A. Taghipour、M. Ghahremani、M. Bennamoun、A. M. Rekavandi、H. Laga 和 F. Boussaid，《Box it to bind it：文本到图像扩散模型中的统一布局控制与属性绑定》，*IEEE*多媒体汇刊，第 27 卷，第 8393–8407 页，2025 年。[5] W. Lei、J. Wang、F. Ma、G. Huang 和 L. Liu，《人物视频生成综合综述：挑战、方法与洞见》，*arXiv* 预印本 *arXiv:2407.08428*，2024 年。[6] Y. HaCohen、N. Chiprut、B. Brazowski、D. Shalem、D. Moshe、E. Richard- son、E. Levin、G. Shiran、N. Zabari、O. Gordon 等，《Ltx-video：实时视频潜扩散》，*arXiv* 预印本 *arXiv:2501.00103*，2024 年。[7] B. Wu、C. Zou、C. Li、D. Huang、F. Yang、H. Tan、J. Peng、J. Wu、J. Xiong、J. Jiang 等，《Hunyuanvideo 1.5 技术报告》，*arXiv* 预印本 *arXiv:2511.18870*，2025 年。[8] Z. Yang、J. Teng、W. Zheng、M. Ding、S. Huang、J. Xu、Y. Yang、W. Hong、X. Zhang、G. Feng 等，《Cogvideox：采用专家变换器的文本到视频扩散模型》，*arXiv* 预印本 *arXiv:2408.06072*，2024 年。[9] A. Blattmann、T. Dockhorn、S. Kulal、D. Mendelevitch、M. Kilian、D. Lorenz 等，《Stable video diffusion：将潜视频扩散模型扩展至大规模数据集》，*arXiv* 预印本 *arXiv:2311.15127*，2023 年。[10] D. Qiu、Z. Fei、R. Wang、J. Bai、C. Yu、M. Fan、G. Chen 和 X. Wen，《Skyreels-a1：视频扩散变换器中的富有表现力的肖像动画》，*arXiv* 预印本 *arXiv:2502.10841*，2025 年。[11] A. Taghipour、M. Ghahremani、M. Bennamoun、A. Miri Rekavandi、Z. Li、H. Laga 和 F. Boussaid，《更快的图像转视频生成：深入探究 CLIP 图像嵌入对时空交叉注意力的影响》，*IEEE Access*，第 13 卷，第 141 313–141 327 页，2025 年。[12] L. Hu、X. Gao、P. Zhang、K. Sun、B. Zhang 和 L. Bo，《Animate anyone：用于角色动画的一致且可控的图像转视频合成》，*arXiv* 预印本 *arXiv:2311.17117*，2023 年。[13] Z. Xu、J. Zhang、J. H. Liew、H. Yan、J.-W. Liu、C. Zhang、J. Feng 和 M. Z. Shou，《MagicAnimate：使用扩散模型实现时间一致的人物图像动画》，*arXiv* 预印本 *arXiv:2311 16498*，2023。[14] T. Wan、A. Wang、B. Ai、B. Wen、C. Mao、C.-W. Xie、D. Chen、F. Yu、H. Zhao、J. Yang 等，《Wan：开放且先进的大规模视频生成模型》，*arXiv* 预印本 *arXiv:2503.20314*，2025 年。

[15] OpenAI，《Sora：大规模文本到视频模型》，https://sora.chatgpt.com，访问日期：2025-12-11。[16] K. Technology，《Kling：高保真文本到视频生成系统》，https://klingai.com/global/，访问日期：2025-12-11。[17] S. Xu、S. Zheng、Z. Wang、H. Yu、J. Chen、H. Zhang、B. Li 和 P.-T. Jiang，《HyperMotion：基于 DiT 的姿态引导复杂动作人物图像动画》，arXiv 预印本 arXiv:2505.22977，2025 年。[18] Y. Yang、H. Sheng、S. Cai、J. Lin、J. Wang、B. Deng、J. Lu、H. Wang 和 J. Ye，《EchoMotion：通过双模态扩散Transformer实现统一的人物视频与动作生成》，arXiv 预印本 arXiv:2512 18814，2025 年。。[19] S. Shi、J. Xu、Z. Li、C. Peng、X. Yang、L. Lu、K. Hu 和 J. Zhang，《一对多动画：无对齐角色动画与图像姿态迁移》，arXiv 预印本 arXiv:2511.22940，2025 年。[20] J. Zhang、S. Cao、R. Li、X. Zhao、Y. Cui、X. Hou、G. Wu、H. Chen、Y. Xu、L. Wang 等，《SteadyDancer：保持首帧的一致协调人物图像动画》，arXiv 预印本 arXiv:2511.19320，2025 年。[21] G. Cheng、X. Gao、L. Hu、S. Hu、M. Huang、C. Ji、J. Li、D. Meng、J. Qi、P. Qiao 等，《Wan-Animate：通过整体复现实现统一的角色动画与替换》，arXiv 预印本 arXiv:2509 14055，2025 年。。[22] M. ControlNeXt、S. UniAnimate 和 S. UniAnimate，《MTVCraft：将 4D 动作标记化以实现任意角色动画》。[23] G. Boduljak、L. Karazija、I. Laina、C. Rupprecht 和 A. Vedaldi，《接下来会发生什么？通过生成点轨迹预测未来动作》，arXiv 预印本 arXiv:2509.21592，2025 年。[24] S. Zhu、J. L. Chen、Z. Dai、Z. Dong、Y. Xu、X. Cao、Y. Yao、H. Zhu 和 S. Zhu，《Champ：通过 3D参数化引导实现可控且一致的人物图像动画》，载于《欧洲计算机视觉会议》。Springer，2024 年，第 145–162 页。[25] L. Hu、G. Wang、Z. Shen、X. Gao、D. Meng、L. Zhuo、P. Zhang、B. Zhang 和 L. Bo，《Animate Anyone 2：具有环境可供性的高保真角色图像动画》，arXiv 预印本 arXiv:2502.06145，2025 年。[26] M. Feng、J. Liu、K. Yu、Y. Yao、Z. Hui、X. Guo、X. Lin、H. Xue、C. Shi、X. Li、A. Li、X. Kang、B. Lei、M. Cui、P. Ren 和 X. Xie，《DreaMoving：基于扩散模型的人物视频生成框架》，arXiv 预印本 arXiv:2312.05107，2023 年。[27] J. Karras、A. Holynski、T.-C. Wang 和 I. Kemelmacher-Shlizerman，《DreamPose：通过稳定扩散实现时尚图像转视频合成》，arXiv 预印本 arXiv:2304.06025，2023 年。[28] M. Lan、C. Zhong、Y. Jiang、Q. Ke、Y. Gu、Q. Zhao 和 W. Zou，《ClearCLIP：分解 CLIP 表征以实现密集视觉语言推理》，载于《欧洲计算机视觉会议》（ECCV），2024 年，第 1–17 页。[29] H.-Y. Chen、Z. Zhou、H. Shang、H. Pouransari、Y. Wu、A. Fang、E. Harber、S.-N. Lim、S. Jayasuriya 和 O. Tuzel，《对比式局部化语言-图像预训练》，载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，2024 年，第 14 212–14 222 页。[30] D. Wu、C. Liang、X. Chen、Z. Gao、S. Han、Y. Li、K. Li、F. Wang、R. Yan 和 X. Xie，《从 CLIP 到 DINO：视觉编码器在多模态大语言模型中发声》，载于《国际学习表征会议》（ICLR），2024 年。[31] S. Yuan、J. Huang、X. He、Y. Ge、Y. Shi、L. Chen、J. Luo 和 L. Yuan，《通过频率分解实现身份保持的文生视频》，载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，2025 年，第 8218–8228 页。[32] Z. Wang、J. Zhang、J. Hu、J. Zhang、L. Zhang 和 Z. Liu，《从大角度到一致面孔：通过人脸专家混合模型实现身份保持的视频生成》，arXiv 预印本 arXiv:2508.09476，2025 年。[33] B. Peng、J. Wang、Y. Zhang、W. Li、M.-C. Yang 和 J. Jia，《ControlNeXt：用于图像与视频生成的强大而高效的控制》，arXiv 预印本 arXiv:2408.06070，2024 年。[34] Y. Ma、Y. He、X. Cun、X. Wang、S. Chen、X. Li 和 Y. Shan，《跟随你的姿态：使用无姿态视频进行姿态引导的文生视频》，载于《AAAI人工智能会议》论文集，第 38 卷，2024 年，第 4117–4125 页。[35] Z. Jiang、Z. Han、C. Mao、J. Zhang、Y. Pan 和 Y. Liu，《VACE：一体化视频创作与编辑》，载于《IEEE/CVF国际计算机视觉会议》论文集，2025 年，第 17 191–17 202 页。[36] R. Shao、Y. Zhang、Z. Liu、H. Sun、Y. Han、W. Wang 和 X. Zhou，《Human4DiT：采用 4D扩散Transformer的自由视角人物视频生成》，arXiv 预印本 arXiv:2405.17405，2024 年。[37] X. Wang、S. Zhang、C. Gao、J. Wang、X. Zhou、Y. Zhang、L. Yan 和 N. Sang

，“UniAnimate：驯化统一视频扩散模型，实现

一致的人物图像动画”，《中国科学：信息科学》，2025。[38] Z. Wang, Y. Li, Y. Zeng, Y. Fang, Y. Guo, W. Liu, J. Tan, K. Chen, T. Xue, B. Dai 等，“HumanVid：揭秘可控摄像机的人物图像动画训练数据”，神经信息处理系统进展，第37卷，第20 111–20 131页，2024。[39] T. Wang, L. Li, K. Lin, C.-C. Lin, Z. Yang, Z. Liu 和 L. Wang，“DisCo：通过解耦控制实现逼真人物舞蹈生成”，载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，2024，第9326–9336页。[40] D. Chang, Y. Shi, Q. Gao, J. Fu, H. Xu, G. Song, Q. Yan, X. Yang 和 M. Soleymani，“MagicPose：借助身份感知扩散实现逼真人体姿势和面部表情重定向”，arXiv预印本 arXiv:2311.12052，2023。[41] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu 和 F. Zou，“MimicMotion：通过置信度感知姿势引导实现高质量人体动作视频生成”，载于国际机器学习会议，2025。[42] S. Tu, Z. Liao, X. Zhao, Z. Huang, Z. Xu, Y. Liu 和 Z. Liu，“StableAnimator：高质量、保留身份的人物图像动画”，arXiv预印本 arXiv:2411.17697，2024。[43] S. Tan, B. Gu, Y. Wu, L. Zhang, J. Wu, F. Gao, J. Yan 和 X. Liu，“Animate-X：通过增强动作表征实现通用角色图像动画”，arXiv预印本 arXiv:2411.10170，2024。[44] W. Yan, S. Ye, Z. Yang, J. Teng, Z. Dong, K. Wen, X. Gu, Y.-J. Liu 和 J. Tang，“SCAIL：通过上下文学习三维一致姿态表示，迈向工作室级角色动画”，arXiv预印本 arXiv:2512.05905，2025。[45] B. Wang, X. Wang, C. Ni, G. Zhao, Z. Yang, Z. Zhu, M. Zhang, Y. Zhou, X. Chen, G. Huang 等，“HumanDreamer：通过解耦生成实现可控人体动作视频生成”，载于计算机视觉与模式识别会议论文集，2025，第12 391–12 401页。[46] S. Fang, C. Chen, L. Wang, C. Zheng, C. Sui 和 Y. Tian，“SignLLM：用于手语生成的大语言模型”，载于IEEE/CVF国际计算机视觉会议论文集，2025，第6622–6634页。[47] B. Saunders, N. C. Camgoz 和 R. Bowden，“Mixed SIGNals：通过动作基元混合实现手语生成”，载于IEEE/CVF国际计算机视觉会议论文集，2021，第1919–1929页。[48] S. Tang, J. He, D. Guo, Y. Wei, F. Li 和 R. Hong，“Sign-IDD：用于手语生成的象似性解耦扩散”，载于AAAI人工智能会议论文集，第39卷，第7期，2025，第7266–7274页。[49] Y. Wang, Z. Wang, J. Gong, D. Huang, T. He, W. Ouyang, J. Jiao, X. Feng, Q. Dou, S. Tang 等，“Holistic-Motion2D：在二维空间中实现可扩展的全身人体动作生成”，arXiv预印本 arXiv:2406.11253，2024。[50] H. Pi, R. Guo, Z. Shen, Q. Shuai, Z. Hu, Z. Wang, Y. Dong, R. Hu, T. Komura, S. Peng 等，“Motion-2-to-3：利用二维动作数据推动三维动作生成”，arXiv预印本 arXiv:2412.13111，2024。[51] J. Li, X. Chen, T. Huang 和 T.-T. Wong，“通过生成和模仿二维动作，学习控制物理模拟的三维角色”，arXiv预印本 arXiv:2512.08500，2025。[52] M. Liu, S. Yan, Y. Wang, Y. Li, G.-B. Bian 和 H. Liu，“MoSA：基于可扩展自回归建模的动作生成”，arXiv预印本 arXiv:2511.01200，2025。[53] R. Xi, X. Wang, Y. Li, S. Li, Z. Wang, Y. Wang, F. Wei 和 C. Zhao，“迈向丰富的视频二维人体动作生成”，arXiv预印本 arXiv:2506.14428，2025。[54] W. Lei, J. Wang, F. Ma, G. Huang 和 L. Liu，“人物视频生成综合综述：挑战、方法与见解”，arXiv预印本 arXiv:2407.08428，2024。[55] M. Zhang, T. Wu, J. Tan, Z. Liu, G. Wetzstein 和 D. Lin，“Gendop：将自回归摄像机轨迹生成作为摄影指导”，arXiv预印本 arXiv:2504.07083，2025。[56] R. Rombach, A. Blattmann, D. Lorenz, P. Esser 和 B. Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022，第10 684–10 695页。[57] O. Sim´eoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa 等，“DINOv3”，arXiv预印本 arXiv:2508.10104，2025。[58] Y. Xu, J. Zhang, Q. Zhang 和 D. Tao，“ViTPose++：用于通用人体姿势估计的视觉变换器基础模型”，IEEE《模式分析与机器智能汇刊》，第46卷，1230，2024 1212页。–。

[59] Y. Zhang, J. Lin, A. Zeng, G. Wu, S. Lu, Y. Fu, Y. Cai, R. Zhang, H. Wang, and L. Zhang，《Motion-x++：大规模多模态三维全身人体运动数据集》，arXiv预印本 arXiv:2501 05098，2025 年。.[60] Y. Jafarian 和 H. S. Park，《通过观看社交媒体舞蹈视频学习高保真着装人体深度》，载于 IEEE/CVF 计算机视觉与模式识别会议（ CVPR ）论文集，2021 年 6 月，第 12 753–12 762 页。[61] K. Soomro、A. R. Zamir 和 M. Shah，《Ucf101：包含 101 个人体动作类别的野外视频数据集》，arXiv预印本 arXiv:1212 0402，2012 年。.[62] H. Kuehne、H. Jhuang、E. Garrote、T. Poggio 和 T. Serre，《Hmdb：用于人体运动识别的大型视频数据库》，载于 2011 年国际计算机视觉会议，2011 年，第 2556–2563 页。[63] C. Schuldt、I. Laptev 和 B. Caputo，《人体动作识别：局部支持向量机方法》，载于第 17 届国际模式识别会议论文集，2004 年。ICPR 2004，第 3 卷，2004 年，第 32–36 页，第 3 卷。[64] A. Shahroudy、J. Liu、T.-T. Ng 和 G. Wang，《Ntu rgb+ d：用于三维人体活动分析的大规模数据集》，载于 IEEE计算机视觉与模式识别会议论文集，2016 年，第 1010–1019 页。[65] Mixamo，《Mixamo：三维角色与动画》，https://www.mixamo. com/，访问日期：2026-01-03。[66] Poly Haven，《Poly haven：公有领域三维资产（ HDRI、纹理和模型）》，https://polyhaven.com/，访问日期：2026-01-03。[67] J. Zhang、Y. Zhang、X. Cun、Y. Zhang、H. Zhao、H. Lu、X. Shen 和 Y. Shan，《使用离散表示从文本描述生成人体运动》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023 年，第 14 730–14 740 页。[68] Y. Shafir、G. Tevet、R. Kapon 和 A. H. Bermano，《作为生成先验的人体运动扩散》，载于第十二届国际学习表征会议，2024 年。[69] X. Chen、B. Jiang、W. Liu、Z. Huang、B. Fu、T. Chen 和 G. Yu，《通过潜空间中的运动扩散执行你的命令》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023 年，第 18 000–18 010 页。[70] ——，《通过潜空间中的运动扩散执行你的命令》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023 年，第 18 000–18 010 页。[71] M. Petrovich、M. J. Black 和 G. Varol，《TEMOS：根据文本描述生成多样化的人体动作》，载于欧洲计算机视觉会议。Springer，2022 年，第 480–497 页。[72] X. Wang、Z. Kang 和 Y. Mu，《文本控制的运动Mamba：文本指令引导的人体运动时间定位》，IEEE图像处理汇刊，第 34 卷，第 7079–7092 页，2025 年。[73] G. Tevet、S. Raab、B. Gordon、Y. Shafir、D. Cohen-Or 和 A. H. Bermano，《人体运动扩散模型》，arXiv 预印本 arXiv:2209.14916，2022 年。[74] Z. Wang、A. Bovik、H. Sheikh 和 E. Simoncelli，《图像质量评估：从误差可见性到结构相似性》，IEEE图像处理汇刊，第 13 卷，第 4 期，第 600–612 页，2004 年。[75] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，《深度特征作为感知度量的惊人有效性》，2018 年 IEEE/CVF计算机视觉与模式识别会议，第 586–595 页，2018 年。 [在线]。可访问：https: // api.semanticscholar.org/CorpusID: 4766599

[76] ´ A. Hore 和 D. Ziou，《图像质量度量：峰值信噪比与结构相似性指数》，载于 2010 年第 20 届国际模式识别会议，2010 年，第 2366–2369 页。[77] T. Unterthiner、S. Van Steenkiste、K. Kurach、R. Marinier、M. Michalski 和 S. Gelly，《迈向准确的视频生成模型：新指标与挑战》，arXiv预印本 arXiv:1812.01717，2018 年。[78] Z. Huang、Y. He、J. Yu、F. Zhang、C. Si、Y. Jiang、Y. Zhang、T. Wu、Q. Jin、N. Chanpaisit 等，《VBench：面向视频生成模型的综合基准测试套件》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2024 年，第 21 807–21 818 页。[79] Z. Teed 和 J. Deng，《RAFT：用于光流的循环全对字段变换》，载于欧洲计算机视觉会议。Springer，2020 年，第 402–419 页。[80] J. Ke、Q. Wang、Y. Wang、P. Milanfar 和 F. Yang，《MUSIQ：多尺度图像质量变换器》，载于 IEEE/CVF国际计算机视觉会议论文集，2021 年，第 5148–5157 页。