VideoJAM：增强运动的联合外观-运动表征
视频模型中的生成

Hila 经理 [*1 2] Uriel Singer Yuval Kirstain Yaniv Taigman Shelly Sheynin [1]
亚当·波兰 [1] https://hila-chefer.github.io/videojam-paper.github.io/ 利奥尔·沃尔夫 [1]

图 1. VideoJAM 生成的文本转视频样本。我们介绍了 VideoJAM，这是一个明确灌输强烈运动的框架
在任何视频生成模型之前。我们的框架显著增强了各种运动类型的运动连贯性。

抽象的

尽管最近取得了巨大进展，但生成
视频模型仍然难以捕捉现实世界
运动、动力学和物理。我们证明
这种限制源于传统的像素
重建目标，使模型偏向于外观保真度，而牺牲了运动
连贯性。为了解决这个问题，我们引入了Video- JAM，这
是一个新颖的框架，可以灌输一种有效的
通过鼓励模型学习联合外观-运动，在视频生成器之
前学习运动
表示。VideoJAM 由两个
互补单元。在训练过程中，我们扩展
目标是从单个
学习表征。在推理过程中，我们引入了内部引导，这是一
种引导
通过扩大运动产生相干运动
将模型自身不断演变的运动预测作为动态引导信
号。值得注意的是，我们的
框架可以应用于任何视频模型
只需进行最少的调整，无需修改训练数据或缩放模
型。
VideoJAM 实现了最先进的性能
在运动连贯性方面，超越了极具竞争力的专有模型，
同时增强了
感知到的视觉质量。这些
研究结果强调了外观和运动
可以互补，如果有效整合，可以提高视觉质量和
视频生成的连贯性。

1. 简介

视频生成领域的最新进展展示了在制作高质量剪辑方面取得的显著进步（Brooks et al.，
2024； KlingAI， 2024； Polyak 等人， 2024）。然而，尽管生成的
视频中，这些模型往往无法准确地描绘运动、物理和动态交互（Kang et al.， 2024；
Brooks 等人， 2024 年） （图2）。当被要求生成体操元素等具有挑战性的动作时（例如，
图2(b) 中的车轮），世代经常表现出严重的

*第一作者在 GenAI、Meta 实习期间完成的工作。
1GenAI, Meta 2特拉维夫大学。联系人：Hila
厨师 <hilach70@gmail.com>。

第42届国际机械会议论文集
学习，加拿大温哥华。PMLR 267，2025。

图 2.视频生成中的运动不连贯性。DiT -30B 生成的不连贯示例（Peebles & Xie, 2023）。该模型在以下方面存在困难:(a) 基本运动,例如慢跑（反复用同一条腿踩踏）;(b) 复杂运动,例如体操;(c) 物理运动,例如物体动力学（铁环穿过女性身体）;以及 (d) 旋转运动,无法复制简单的重复模式。

变形,例如出现额外的肢体。在其他情况下,生成过程会表现出与基本物理相矛盾的行为,例如物体穿过其他固体（例如,图2(c) 中的呼啦圈穿过女性）。另一个例子是旋转运动,模型很难复制简单重复的运动模式（例如,图2(d) 中的旋转器）。有趣的是,即使对于模型训练数据中很好地表示的基本运动类型（例如,图2(a) 中的慢跑）,这些问题也很突出,这表明数据和尺度可能不是导致视频模型中时间问题的唯一因素。

在本研究中,我们旨在深入探究视频模型为何难以实现时间相干性,并提出一种能够实现最佳运动生成结果的通用解决方案。首先,我们发现像素质量与运动建模之间的差距很大程度上源于相同的训练目标。通过定性和定量实验（参见第三节）,我们证明了基于像素的目标几乎不受生成步骤中时间扰动的影响,而时间扰动对于确定运动至关重要。

受这些洞见的启发,我们提出了VideoJAM,这是一个新颖的框架,通过教授视频模型联合外观-运动表示,为其提供显式运动先验。这通过两项互补的改进实现:在训练过程中,我们修改目标函数,使其除了外观之外,还能预测运动;在推理过程中,我们提出了一种引导机制,利用学习到的运动先验进行时间相干的生成。

具体来说,在 VideoJAM 训练期间,我们将视频与其对应的运动表征配对,并修改网络以预测两种信号（外观和运动）。为了适应这种双重格式,我们仅在架构中添加了两个线性层（见图4）。第一层位于模型的输入端,将两个信号合并为一个表征。第二层位于模型的输出端,从学习到的联合表征中提取运动预测。然后,修改目标函数以预测联合外观-运动分布,从而鼓励模型依赖新增的运动信号。

在推理阶段,我们的主要目标是视频生成,预测的运动作为辅助信号。为了引导生成过程有效地结合学习到的运动

在此之前,我们引入了Inner-Guidance,一种新颖的推理时间引导机制。与依赖于固定外部信号的现有方法（Ho & Salimans， 2022； Brooks 等人， 2023）不同， Inner-Guidance 利用模型自身不断演变的运动预测作为动态引导信号。这种设置需要解决独特的挑战:运动信号本质上依赖于其他条件和模型权重,这使得先前工作的假设无效,需要新的公式（附录A 第 2 节）。我们的机制直接修改模型的采样分布,使生成过程趋向于外观-运动联合分布,而不是仅基于外观的预测,从而使模型能够在整个生成过程中改进自身的输出。

通过大量实验,我们证明将VideoJAM 应用于预训练视频模型可显著增强不同模型大小和不同运动类型的运动一致性。此外,VideoJAM 确立了运动建模领域的全新领先水平,甚至超越了竞争激烈的专有模型。这些进步无需修改数据或缩放模型即可实现。VideoJAM 设计直观,仅需添加两个线性层,因此既通用又可轻松适配任何视频模型。有趣的是,尽管我们没有明确以像素质量为目标,但 VideoJAM 还提高了生成的感知质量。这些发现强调,外观和运动并非相互排斥,而是内在互补的。

2.相关工作

扩散模型（Ho 等人， 2020)彻底改变了视觉内容生成。从图像生成（Dhariwal & Nichol， 2021； Rombach 等人， 2022； Ho 等人， 2022a； Black Forest Labs， 2024； Dai 等人， 2023； OpenAI， 2024)到编辑和个性化（Gal 等人， 2022； Ruiz 等人， 2023； Chefer 等人， 2024b;a），以及最近的视频生成。最初将扩散模型应用于视频的尝试依赖于模型级联（Ho 等人， 2022b； Singer等人， 2023)或使用时间层直接“膨胀”图像模型（Guo 等人， 2023； BarTal 等人， 2024； Wu 等人， 2023）。其他工作则侧重于添加自动编码器以提高效率（Blattmann 等人， 2023b； An 等人， 2023；

Wang 等人， 2023）,或根据图像调节生成（Blattmann 等人， 2023a； Girdhar 等人， 2024； Hong 等人， 2022）。最近,UNet 主干被Transformer （Polyak 等人， 2024； Brooks 等人， 2024； Genmo， 2024； HaCohen 等人， 2024)取代,主要遵循扩散变换器（DiT） （Peebles & Xie， 2023）。

为了控制生成内容， Dhariwal & Nichol (2021)引入了分类器引导,其中分类器梯度引导生成内容朝向特定类别。Ho & Salimans (2022)提出了无分类器引导 (CFG),用文本替换分类器。与内部引导类似,CFG修改了采样分布。然而,CFG 无法处理噪声条件或多重条件。与我们的工作最接近的是Liu 等人(2022) 的论文,它使用组合分数估计c1, ···来处理多重条件cn 。

$$p_\theta(x|c1,\ldots,cn) = \frac{p_\theta(x,c1,\ldots,cn)}{p_\theta(c1,\ldots,cn)} \propto p_\theta(x,c1,\ldots,cn) = p_\theta(x)\prod_{i=1}^{n} p_\theta(\text{在}|x\text{中})。$$

其中θ表示模型权重， p表示采样分布。上式假设c1, ..., cn彼此独立且与θ无关,但这在我们的例子中并不成立,因为运动是由模型直接预测的,因此本质上取决于θ和条件。同样， Brooks 等人(2023)假设条件和模型权重θ 之间相互独立,这在我们的设定下同样是不正确的。更多讨论请参见附录A。

像素质量与时间相干性之间的差距是一个突出的问题(Ruan et al., 2024; Brooks et al., 2024; Liu et al., 2024b; Kang et al., 2024)。先前的研究探索了基于运动或物理的信号来改进视频生成。一些方法将它们用作指导或编辑的输入(Geng et al., 2024; Ma et al., 2023; Liu et al., 2024a; Cong et al., 2023; Montanaro et al., 2024)。需要注意的是,他们的目标与我们的目标不同,因为我们的目标是教会模型一个时间先验,而不是将其作为输入。其他方法建议将内容生成和运动生成分离（Tulyakov等人， 2018； Ruan 等人， 2024； Qing 等人， 2023； Shen 等人， 2024； Jin 等人， 2024b）。在图像生成方面，Liu 等人（2023)通过结合空间先验（例如表面法线和深度图)以及给定的条件骨架作为输入,提升了人体图像的真实感。最后,与我们的方法最相似的是,近期的一些研究使用运动表征来提升图像到视频生成中的连贯性（Shi 等人， 2024； Wang 等人， 2024）,但这些研究仅限于以图像为条件的模型。

3. 动机

在训练过程中,生成视频模型会采用有噪声的训练视频,并通过比较模型的

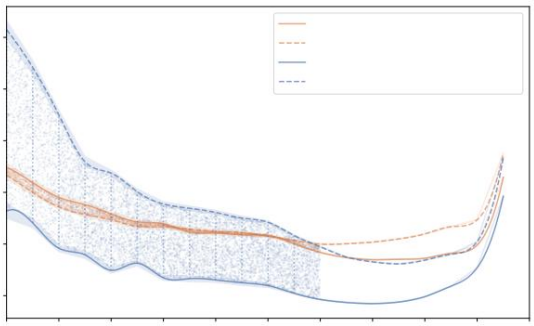

图 3.动机实验。我们使用“原始” DiT（橙色)和我们微调后的模型（蓝色)比较了随机排列视频帧前后模型的损失。原始模型在 t ⩽ 60 时几乎不受时间扰动的影响。

使用原始视频、噪声或两者的组合进行预测（Ho 等人， 2020 年； Lipman 等人， 2023 年）（第4.1节）。我们假设,这种公式会使模型偏向基于外观的特征,例如颜色和纹理,因为这些特征主导了像素级差异。因此,该模型不太倾向于关注时间信息,例如动态或物理信息,这些信息对目标函数的贡献较小。为了证明这一论断,我们进行了实验来评估模型对时间不相干性的敏感性。为了提高效率,以下实验在 DiT-4B （Peebles & Xie, 2023)上进行。

我们进行了一项实验,将两种不同的视频加噪后输入模型:第一组是未经任何干扰的普通视频,第二组是对帧进行随机排列后的视频。假设模型能够捕捉时间信息,我们预计时间非相干（受扰动)的输入将导致比时间相干的输入更高的测量损失。

给定一组随机的35,000 个训练视频,我们将每个视频的噪声设置为随机去噪步长t ∈ [0, 99]。然后我们
检查前后测量的损失差异
排列并按时间步长汇总结果。我们考虑了两种模型 采用基于像素目标的“原始” DiT 模型,以及经过微调的 VideoJAM 模型,该模型添加了显式运动目标（第4 节）。

实验结果如图3 所示。可以看出,原始模型在生成步骤60之前几乎不受帧重排的影响。这意味着该模型无法区分有效视频和时间不连贯的视频。与此形成鲜明对比的是,我们的模型对这些扰动极其敏感,计算损失的显著差异就证明了这一点。

在应用B中,我们包含一项定性实验,证明步长t ⩽ 60决定了视频中的粗略运动。这两个结果都表明,训练目标对时间不连贯性不太敏感,导致模型更倾向于外观而非运动。

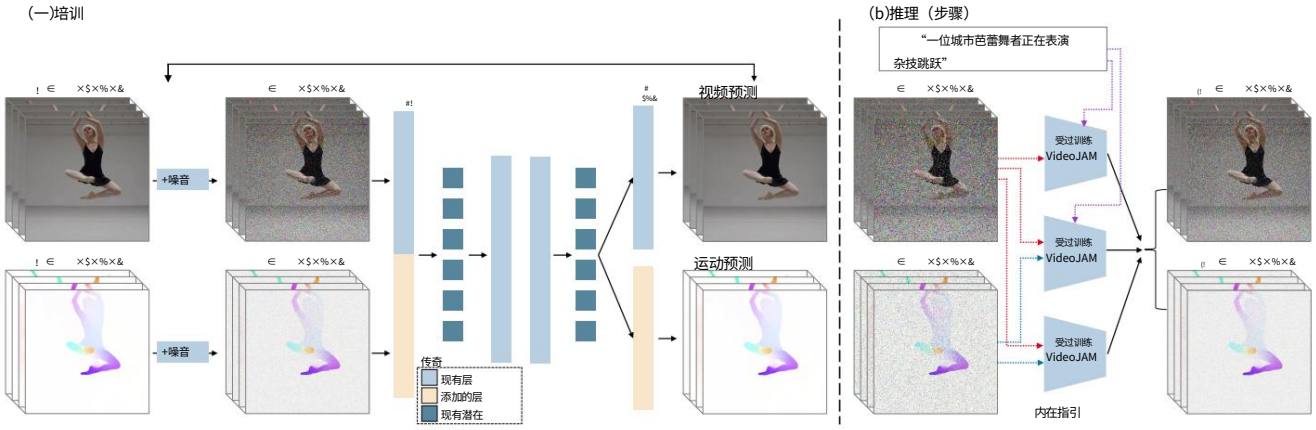

图 4. VideoJAM 框架。VideoJAM由两个单元构成：(a)训练。给定输入视频x1及其运动表征d1,两个信号均被加噪,并使用线性层W+嵌入到单个联合潜在表征中。扩散模型处理输入,两个线性投影层根据联合表征预测外观和运动。(b)推理。 在。

我们提出了内部指导,其中模型自身的噪声运动预测用于指导每一步的视频预测。

4. VideoJAM

受上一节见解的启发,我们提出教会模型一种同时包含外观和运动的联合表征。我们的方法包含两个互补的阶段（见图4）:(i) 在训练过程中,我们修改目标函数以预测联合外观-运动分布;这通过修改架构以支持双输入输出格式来实现,其中模型同时预测视频的外观和运动。(ii)

在推理中,我们添加了 Inner-Guidance,这是一种新颖的公式,它利用预测的运动来引导生成的视频实现连贯的运动。

4.1. 准备工作

我们在扩散变换器 (DiT) 架构上进行实验,该架构已成为视频生成的标准主干(Brooks 等人， 2024 年； Genmo， 2024 年)。

该模型在时间自编码器 (TAE)的潜在空间中运行,该编码器在空间和时间上对视频进行下采样以提高效率。我们使用流匹配(Lipman et al., 2023)来定义目标函数。在训练过程中,给定视频x1、随机噪声x0 N (0, I)和时间步长t ∈ [0, 1],使用x0对x1进行噪声处理,以获得如下中间潜在函数：

$$xt = tx1 + (1 - t)\, x0。 \quad (1)$$

然后优化模型来预测速度,即,

$$速度= \frac{正确的}{日期} = x1 - x0。 \quad (2)$$

因此,用于训练的目标函数变为,

$$L = Ex1,x0 \quad N(0,1),y,t\in[0,1]\ ||u(xt, y, t; \theta) - vt||2\ 2\ , \quad (3)$$

其中y是（可选)输入条件， θ表示权重,u(xt, y, t; θ) 是模型的预测。

预测u 是使用 DiT 获得的。首先,模型将xt “拼接”成p × p 个视频块序列。该序列通过线性投影投影到 DiT 的嵌入空间中， Win ∈ R p CDiT分别是 TAE 和 DiT 的嵌入维度。然后,DiT 应用堆叠注意力层来生成视频的潜在表示,该表示被投影回 TAE 的空间,使用Wout ∈ R CDiT×CTAE·p得出最 ^{2} ·CTAE×CDiT, CTAE和终预测。

^{2} , IE。,

$$u(xt, y, t; \theta) = M(xt \cdot Win, y, t; \theta) \cdot Wout, \quad (4)$$

其中M表示注意力模块。为了提高效率,我们采用如上所述的预训练模型,并使用 VideoJAM 对其进行微调,如下所述。

4.2 联合外观-运动表征

我们首先描述 VideoJAM 采用的运动表示。我们选择使用光流,因为它灵活、通用,并且易于表示为 RGB 视频;因此,它不需要训练额外的 TAE。光流计算光学保持对之间的密集位移场,其中d(u, v)是像素(u, v)从I1到I2 的位移。为了将d转换为 RGB图像,我们计算每个像素的角度和范数,

框架。给定两个框架I1， I2 ∈ R 高×宽×3 ,

流量， d∈R 高×宽×2 ,

$$um = 最小值\ 1, \sigma \frac{\sqrt{\quad^{2} + 在^{2}}}{\sqrt{\ H2 + W2}} \quad ,\alpha=\arctan2(v,u), \quad (5)$$

其中m为归一化运动幅度， σ = 0.15， α为运动方向（角度）。每个角度都被赋予一种颜色,像素不透明度由 m 决定。我们的归一化使模型能够捕捉运动幅度,较大的运动幅度对应于较高的m值。

并降低不透明度。通过使用系数σ = 0.15代替全分辨率(√ H2 + W2),我们可以防止细微运动变得过于不透明,从而确保它们仍然可区分。RGB 光流经 TAE 处理,产生一个带噪声的表示dt （参见公式1）。

接下来,我们修改模型以预测外观和运动的联合分布。我们通过将架构更改为双输入输出格式来实现这一点,其中模型同时接收带噪视频xt和带噪光流dt,并预测这两个信号。这需要修改两个线性投影矩阵Win和Wout （见图4(a)）。

首先,我们扩展输入投影Win ,使其包含两个输入 视频和运动潜在向量xt、 dt。我们通过添加零行来获得双投影矩阵W+ CTAE·p R 2·CTAE·p 2×CDiT ,使得初始化时网络等同输出矩阵 2 于预训练的 DiT,并忽略添加的运动信号。其次,我们扩展Wout ,添加一个 ∈ W+ ∈ R CDiT×2·CTAE·p。添加的层从联合潜在表示中提取运动预测。总之， W+将模型更改为

出去 2 。

在 和W+ 出去，

一种处理和预测外观和运动的双输入输出格式。

如图4(a) 所示,我们的修改保留了DiT 的原始潜在维度。本质上,这要求模型学习一个统一的潜在表示,并利用该表示通过线性投影来预测两个信号。将上述公式代入公式4中,我们得到:

′
u + ([xt, dt], y, t; θ) = M([xt, dt] · W+ in, y, t; θ) · W+其中[·]表示通道维度中的连接，u去，

表示如上所述的扩展模型权重， + ud]表示双重输出,其中第一个和u通道表示外观（视频)预测,而最后一个通道表示运动（光流)预测。

= [u x ，

最后,我们扩展训练目标,使其包含一个明确的运动项,因此,等式3中的目标变为:

′ +
L = E[x1,d1],[x0,d0],y,t ||u +([xt, dt], y, t; θ) − 在 吨 ||2 2 , (6)

+ vd]

学习新的目标分 使 用公式[√ 计算,注意,其 吨 中v虽然我们只修改了两个线性层,但我们联合微调了网络中的所有权重,以允许模型布。

在推理过程中,该模型会根据噪声生成视频及其运动表征。需要注意的是,我们主要关注的是视频预测,而运动预测则引导模型获得时间上合理的输出。

4.3. 内在指引

正如之前所观察到的（Ho & Salimans， 2022）,以辅助信号为条件的扩散模型并不能保证模型会忠实地考虑该条件。因此,我们建议修改扩散得分函数,以使预测朝着合理的运动方向发展。

在我们的设定中,有两个条件信号:提示y 和嘈杂的中间运动预测dt。值得注意的是， dt本质上取决于提示和模型权重,因为它是由模型本身生成的。因此,假设条件和模型权重之间相互独立（例如Brooks 等人(2023))的现有方法不适用于此设定（附录A 第 2 节）。为了解决这个问题,我们建议直接修改采样分布，

$$p \quad \theta'([xt, dt]|y) \propto p\theta'([xt, dt]|y)p\theta'(y|[xt, dt])w1\, p\theta'(dt|xt, y)\, w2 \quad , \tag{7}$$

其中pθ ′ ([xt, dt]|y)是原始采样分布， pθ ′ (y|[xt, dt])估计在给定联合预测的情况下提示的似然值， pθ ′ (dt|xt, y)估计含噪运动预测的似然值。后者旨在提高模型的运动一致性,因为它最大化了生成视频运动表征的似然值。利用贝叶斯定理,等式7等价于:

$$p\theta'([xt, dt]|y) \quad \frac{p\theta'([xt, dt], y)}{p\theta'([xt, dt])}^{w1} \quad \frac{p\theta'([xt, dt], y)}{p\theta'(xt, y)}^{w2}$$

$$\propto p\theta'([xt, dt]|y) \quad \frac{p\theta'([xt, dt]|y)}{p\theta'([xt, dt])}^{w1} \quad \frac{p\theta'([xt, dt]|y)}{p\theta'(xt|y)}^{w2},$$

其中,我们省略所有pθ ′ (y)的出现,因为y是一个外部常数输入。接下来,我们可以通过取对数导数将其转换为相应的得分函数，

$$(1 + w1 + w2)\nabla\theta' \log p\theta'([xt, dt]|y) \quad -w1\nabla\theta' \log p\theta'([xt, dt]) - w2\nabla\theta' \log p\theta'(xt|y)。 \tag{8}$$

按照Ho & Salimans (2022) 的研究,我们通过在30%的训练步长中随机丢弃文本,并在20%的训练步长中随机丢弃光流（设置d = 0），联合训练模型,使其对两个辅助信号 y 和 d进行条件和非条件训练。具体而言,当丢弃光流信号时,损失仅针对外观计算,从而得到第4.1节中描述的训练方案。因此,推理过程中的整体指导公式变为:

$$u + ([xt, dt], y, t; \theta') = (1 + w1 + w2) \cdot u + ([xt, dt]), y, t; \theta') \quad -w1 \cdot u + ([xt, dt], \varnothing, t; \theta') - w2 \cdot u + ([xt, \varnothing], y, t; \theta')。$$

除非另有说明,所有实验都使用w1 = 5、 w2 = 3,其中 w = 5 是基础模型的文本指导尺度。

5.实验

我们进行了定性和定量实验,以证明 VideoJAM 的有效性。我们将我们的模型与其基础版本（预训练版本)以及领先的专有和开源视频模型进行了基准测试,以突出 VideoJAM 增强的运动连贯性。

图 5. VideoJAM-30B 的文本转视频结果。VideoJAM可以生成各种类型的运动,从基本的运动（例如,跑步)到复杂运动（例如,杂技)以及改进的物理（例如,跳过障碍）。

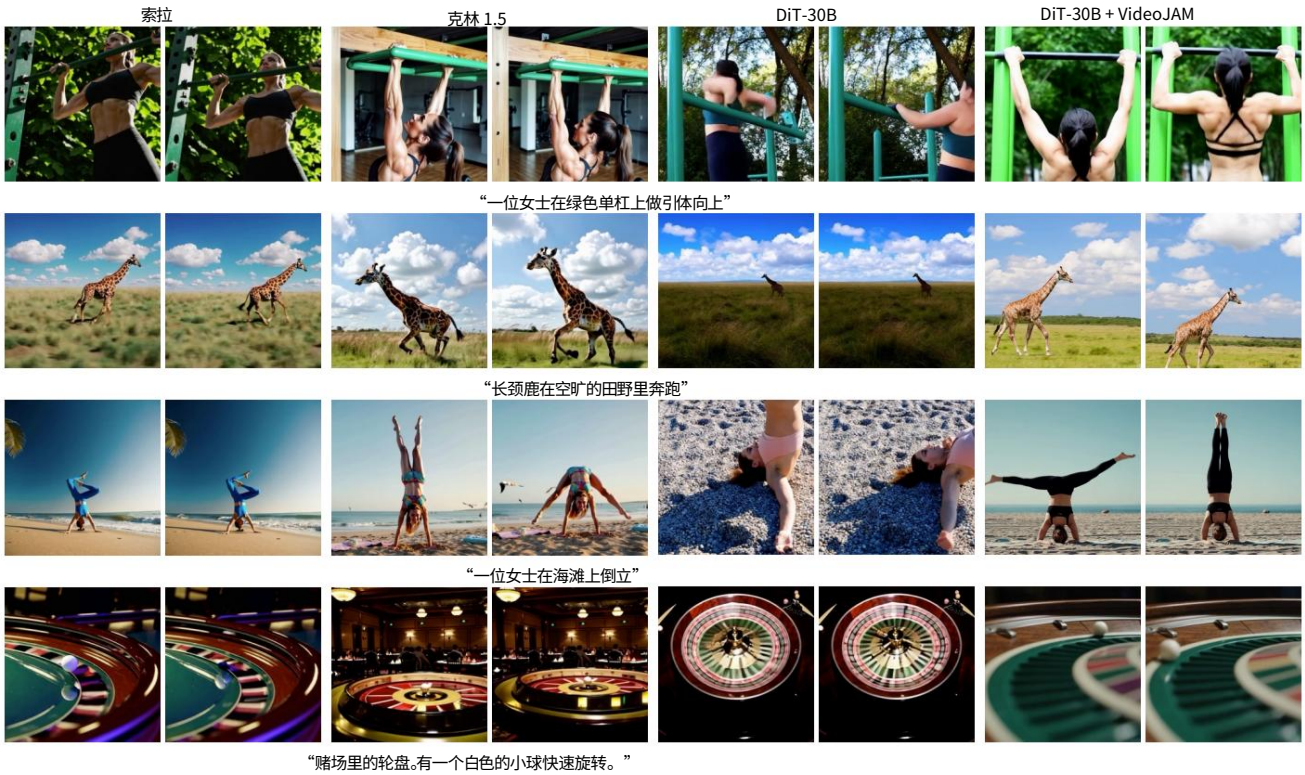

图 6. VideoJAM-30B 与领先基线 Sora、Kling 和 DiT-30B 在代表性数据集上的定性比较 VideoJAM-bench 的提示。基线难以进行基本运动,显示“向后运动”（Sora,第二行)或不自然运动（Kling,第二行）。生成的内容违背了物理的基本定律,例如,人们穿过物体（DiT,第一行）,或者出现或消失的物体（Sora、DiT,第 4 行）。对于复杂运动,基线显示静态运动或变形（Sora、Kling,第一排,第三排）。相反,在所有情况下,VideoJAM 都能制作出时间连贯的视频,更好地遵循物理定律。

实施细节我们考虑两种变体
DiT 文本到视频模型 DiT-4B 和 DiT-30B,以证明运动连贯性是两者的常见问题
小型和大型模型。我们所有的模型都使用
为了提高效率,空间分辨率为256 × 256。模型
训练生成128帧视频,每帧 24 帧
秒,从而生成 5 秒的视频。DiT
使用第4.1节中的框架对模型进行了预训练
一个包含O(100 M)个视频的内部数据集。然后我们进行微调
使用300万个随机样本的 VideoJAM 模型
来自模型的原始训练集,其构成较少
比 3% 的训练视频更好。这使得我们的微调
阶段要轻便高效。在这个微调过程中,我们
采用 RAFT (Teed & Deng, 2020)来获取光流。
有关更多实施细节,请参阅附录C。

基准我们使用两个基准进行评估。
首先,我们介绍专为测试运动连贯性而构建的 VideoJAM-bench。其次,我们考虑
Movie Gen (MGen) 基准（ Polyak 等人， 2024）
显示了我们结果的稳健性。

VideoJAM-bench 解决了包括 MGen 在内的现有基准测试的局限性,这些基准测试未能全面评估现实世界中具有挑战性的运动场景。例如,

MGen 的第二大类别是“异常活动” （23.4%
MGen）,与我们评估现实世界（“通常”）动态的目标形成对比。第三大类别,
“场景” （占 MGen 的 19.9%）,重点关注几乎静态的场景
因此,自然界天生就优先考虑外观,而不是有意义的动作。即使是与我们重叠的类别,例如
作为“动物”,MGen 给出的代表性例子是
“一只好奇的猫从舒适的藏身处向外张望”。

为了构建 VideoJAM-bench,我们考虑了以下提示
对视频生成器构成挑战的四类自然运动（见图2）:基本运动、复杂运动、旋转
运动和物理。我们使用训练数据中的一个保留集（没有模型在该集上进行训练）,并使用
法学硕士选择最适合的128 个提示,至少
四个类别之一,描述一个单一的、具体的、
和清晰的运动。为了避免评估偏向
针对特定的提示风格,我们要求法学硕士修改
提示的长度和细节程度各不相同。完整的
我们的提示列表可以在附录D中找到。

基线我们考虑了各种最先进的
模型,包括专有和开源模型。在较小的类别中,我们包括 CogVideo2B、CogVideo5B （Hong 等人，
2022）、 PyramidFlow （Jin et al.， 2024a)和基础模型
在更大的类别中,我们评估了领先的开源模型（Mochi （Genmo，
2024）、 CogVideo5B)和
具有外部 API 的专有模型（Sora （Brooks 等人，
2024)、 Kling 1.5 （KlingAI， 2024）、 RunWay Gen3
（ RunwayML， 2024）)以及基础型号 DiT-30B。
使用视频排行榜选择领先基线。

表 1. VideoJAM-4B 与之前研究的比较
VideoJAM-bench。人工评估显示投票百分比
支持 VideoJAM;自动指标使用 VBench。

	人类评估			自动指标	
方法	文字信仰	品质	运动	外观	运动
CogVideo2B	84.3	94.5	96.1	68.3	90.0
CogVideo5B	62.5	74.7	68.8	71.9	90.1
金字塔流	76.6	83.6	82.8	73.1	89.6
DiT-4B	71.1	77.3	82.0	75.2	78.3
+VideoJAM	-	-	-	75.1	93.7

表 2. VideoJAM-30B 与之前研究的比较
VideoJAM-bench。人工评估显示投票百分比
支持 VideoJAM;自动指标使用 VBench。

	人类评估			自动指标	
方法	文字信仰	品质	运动	外观	运动
CogVideo5B	73.4	71.9	85.9	71.9	90.1
RunWay Gen3	72.2	76.6	77.3	73.2	92.0
Mochi	56.3	65.6	74.2	69.9	89.7
索拉		68.5	51.7	75.4	91.7
克林 1.5	51.8	45.9	63.8	76.8	87.1
DiT-30B	71.9	74.2	72.7	72.4	88.1
+VideoJAM	-	-	-	73.4	92.4

5.1. 定性实验

图1、5、9展示了使用 VideoJAM- 30B 获得的结果。结果显示了各种各样的运动类型
挑战现有的体操模式（例如空中
分裂）,需要物理理解的提示（例如,
手指压进黏液里、篮球落入网中)等等。

图6将 VideoJAM 与领先的基线进行了比较,
Sora 和 Kling,以及基础型号 DiT-30B,在提示下
来自 VideoJAM-bench。比较结果凸显了最先进模型中的运动问题。即使是简单的运动,
例如奔跑的长颈鹿（第二行）,显示类似
“向后运动”（Sora)或不自然的运动（Kling,
DiT-30B）。复杂的动作,比如引体向上或倒立,
静态视频的结果（Sora,第一行和第三行;Kling,第一行）
行)或身体变形（Kling,第三行）。基线
也表现出物理违规行为,例如物体消失或出现（Sora,DiT-30B,第四行）。相比之下,
VideoJAM 始终如一地产生连贯的动作。

5.2. 定量实验

我们评估外观和运动质量,以及
使用自动指标和人工
评估。在我们所有的比较中,每个模型运行一次
对所有基准提示使用相同的随机种子。
对于自动指标,我们使用 VBench （Huang et al.，
2024）,评估解开的

表 3.消融研究。消融的主要成分
使用 VideoJAM-bench 在 VideoJAM-4B 上构建我们的框架。人类评估显示支持 VideoJAM 的投票百分比。

	人类评估			自动指标	
消融类型	文字信仰。品质运动外观运动				
无文字指导 无内部	68.0	62.5 63.3		74.5	93.3
指导68.9		64.4 66.2		75.3	93.1
无光流	79.0	70.4 80.2		74.7	90.1
IP2P 指导	73.7	85.2 78.1		72.0	90.4
+VideoJAM-4B	-	-	-	75.1	93.7

轴。我们按照论文的方法,将分数汇总为两类 外观和运动。指标评估
每帧的质量、美学、主题一致性、
生成运动的数量以及运动的连贯性。更多
有关指标及其汇总的详细信息,请参阅
附录C.1。

对于人工评估,我们遵循两种选择
强制选择（2AFC)协议,类似于Rombach 等人的协议。
（2022）； Blattmann 等人（2023a）,其中评分者比较两个
视频（一个来自 VideoJAM,一个来自基线)并选择
基于质量、运动和文本对齐的最佳方案。
每个比较由5 位独立用户评分,提供
每个基准的基线至少有 640 个响应。

VideoJAM-bench 上的比较结果
4B、30B模型分别如表1、表2所示。
此外,自动指标的完整细分如下
附录D 中给出了比较结果
Movie Gen 基准测试在 App. E中提供。在所有情况下,
VideoJAM 在所有模型尺寸中的表现均优于所有基线
运动连贯性方面,包括自动和
与人类评估有相当大的差距（表1、2、7）。

值得注意的是,VideoJAM-4B 的表现优于 CogVideo5B
基线,尽管后者大了25% 。对于
30B 版本,VideoJAM 甚至超越了专有
最先进的模型,例如 Kling、Sora 和 Gen3
（运动偏好分别为63.8%、68.5%、77.3%）。
考虑到 Video- JAM 是在明显较低的分辨率（256)下训练的,这些结果尤其
令人印象深刻
与基线（768及更高)相比,并进行了微调
仅基于300万个样本。虽然这种分辨率差异
解释了为什么 Kling 和 Sora 等专有模型在视觉质量上超越了我们的模型（表2）,VideoJAM 始终
表现出更好的运动连贯性。

最重要的是,VideoJAM 显著改善了运动
其基础模型 DiT-4B 和 DiT-30B 的一致性
直接的苹果与苹果的比较。在 DiT-4B 的测试中,人类评分者在82.0%的情况下更倾向于 VideoJAM 的动作
DiT-30B 的评分为 72.7%。VideoJAM 的评分也高于
质量（4B 为 77.3%， 30B 为 74.2%)和文本忠实度
（4B 为 71.1%， 30B 为 71.9%）,表明我们的方法

图 7.局限性。我们的方法在以下情况下效果较差:(a) 运动在“缩小”中观察到（移动物体覆盖了一小部分框架）。(b)物体相互作用的复杂物理。

也增强了这一代人的其他方面。

5.3. 消融研究

我们消除了框架的主要设计选择。
首先,我们在内部指导公式中消除了文本指导和运动指导的使用（通过设置w2 = 0,
w1 = 0（在公式8 中）。接下来,我们省略
推理过程中的运动预测,通过放弃
每个推理步骤的光流（d = 0）。最后,我们
用以下公式代替我们的指导公式
InstructPix2Pix (IP2P) 指导(Brooks 等人， 2023)（参见
第 2节,附录A）。请注意,DiT 模型的结果
表1、2也起到消融的作用,因为它们消融了使用
VideoJAM 在训练和推理过程中的作用。

结果报告于表3 中。所有消融都会导致运动相干性显著下降,其中去除

运动指导的危害比移除
文本引导,表明运动引导部分确实引导模型走向时间连贯

代。此外,在推理时放弃光流预测是最有害的,证实了

联合输出结构的好处在于执行合理的
运动。InstructPix2Pix 引导比较进一步表明,我们的内部引导公式是最

适合我们的框架,因为它给出了第二低的结果
就运动而言。

最后,请注意,人类评估者始终更喜欢
VideoJAM 在视觉质量和文本对齐方面
在所有消融中,进一步证实 VideoJAM
有益于视频生成的各个方面。

5.4. 限制

虽然 VideoJAM 显著提高了时间相干性,但挑战依然存在（见图7）。首先,由于计算限制,我们依赖于有限的训练分辨率

和 RGB 运动表示,这阻碍了模型
能够在“缩小”场景中捕捉动作
移动的物体只占据画面的一小部分。在
在这些情况下,相对运动幅度减小,使得表示的信息量降低（等式5）。例如,
在图7(a) 中,没有打开降落伞,运动显得不连贯。其次,虽然运动和物理

尽管运动相互交织,从而提升了物理性能,但我们的运动表征缺乏明确的物理编码。这限制了模型处理物体交互的复杂物理特性的能力。例如,在图7(b) 中,球员的脚在改变轨迹之前并未触球。

6.结论

视频生成提出了一个独特的挑战,需要对空间交互和时间动态进行建模。

尽管取得了令人瞩目的进步,视频模型仍然难以实现时间连贯性,即使是训练数据集中表现良好的基本动作(图2)。在本研究中,我们将训练目标确定为一个关键因素,优先考虑外观保真度而非运动连贯性。

为了解决这个问题,我们提出了 VideoJAM,这是一个为视频模型配备显式运动先验的框架。其核心理念直观自然:单一潜在表征即可同时捕捉外观和运动。VideoJAM 仅使用两个额外的线性层,无需额外的训练数据,即可显著提升运动连贯性,即使与强大的专有模型相比也能取得最佳效果。我们的方法具有通用性,为未来利用复杂物理等现实世界先验知识增强视频模型提供了众多机会,为现实世界交互的整体建模铺平了道路。

影响声明

这项工作的主要目标是推进视频生成中运动建模技术的发展,使模型能够更忠实地理解和呈现世界。与内容生成领域的任何技术一样,视频生成也存在被滥用的可能性,这一担忧在研究界已引起广泛讨论。然而,我们的工作并未引入任何先前进展中尚未出现的特定风险。我们坚信开发和应用工具来检测偏见并减少恶意使用案例的重要性,从而确保包括我们在内的生成工具的安全和公平使用。

参考

An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.-B., Luo, J.和 Yin, X. Latent-Shift:利用时间移位实现潜在扩散,实现高效的文本转视频生成。arXiv 预印本arXiv:2304.08477,2023 年。

BarTal, O.、Chefer, H.、Tov, O.、Herrmann, C.、Paiss, R.、 Zada, S.、Ephrat, A.、Hur, J.、Li, Y.、Michaeli, T.、Wang, O.、Sun, D.、Dekel, T. 和 Mosseri, I. Lumiere:用于视频生成的时空扩散模型。 arXiv 预印本arXiv:2401.12945, 2024。

黑森林实验室。FLUX,2024。网址:https://blackforestlabs.ai/。

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A. 等人。稳定视频扩散:将潜在视频扩散模型扩展至大型数据集。arXiv 预印本arXiv:2311.15127, 2023a。

Blattmann, A.、Rombach, R.、Ling, H.、Dockhorn, T.、Kim, SW、Fidler, S. 和 Kreis, K. Align your latents:具有潜在扩散模型的高分辨率视频合成。

在 CVPR,2023b 中。

Brooks, T.、Holynski, A. 和 Efros, AA 合著的《InstructPix2Pix:学习遵循图像编辑指令》。CVPR, 2023 年。

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., 以及 Ramesh, A. 视频生成模型作为世界模拟器。2024 年。

Chefer, H.、Lang, O.、Geva, M.、Polosukhin, V.、Shocher, A.、michal Irani、Mosseri, I. 和 Wolf, L. 扩散模型的隐藏语言。发表于第十二届国际学习表征大会, 2024a。网址: https://openreview.net/forum?id=awWpHnEJDw。

Chefer, H., Zada, S., Paiss, R., Ephrat, A., Tov, O., Rubin- stein, M., Wolf, L., Dekel, T., Michaeli, T., 和 Mosseri, I. Still-moving:无需定制视频数据的定制视频生成。arXiv 预印本 arXiv:2407.08674, 2024b。

Cong, Y.、Xu, M.、Simon, J.、Chen, S.、Ren, J.、Xie, Y.、Perez-Rua, J.-M.、Rosenhahn, B.、Xiang, T. 和 He, S. Flatten:光流引导注意力,实现一致的文本到视频编辑。arXiv 预印本 arXiv:2310.05922,2023 年。

Dai, X., Hou, J., Ma, C.-Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A. 等人。Emu:使用大海捞针增强图像生成模型。arXiv 预印本 arXiv:2309.15807,2023 年。

Dhariwal, P. 和 Nichol, A. 扩散模型在图像合成方面胜过 GAN。arXiv 预印本 arXiv:2105.05233,2021年。

Gal, R.、Alaluf, Y.、Atzmon, Y.、Patashnik, O.、Bermano, AH、Chechik, G. 和 Cohen-Or, D. 一图胜千言:使用文本反转实现个性化文本到图像生成。arXiv 预印本 arXiv:2208.01618,2022 年。

Geng, D.、Herrmann, C.、Hur, J.、Cole, F.、Zhang, S.、Pfaff, T.、Lopez-Guevara, T.、Doersch, C.、Aytar, Y.、Rubinstein,