生成图像动力学

李正奇　　理查德-塔克　　诺亚-斯纳维利　　亚历山大-霍林斯基

谷歌研究

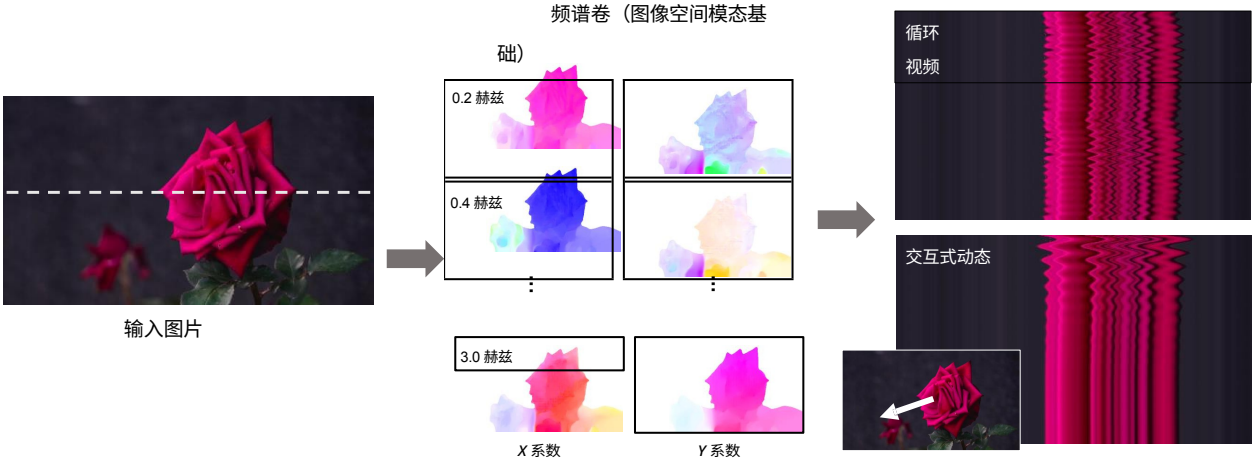

图 1.我们建立了一个关于场景运动的生成式图像空间先验模型：从单张 RGB 图像中，我们的方法生成了一个*频谱体*[23]，这是一种运动表示法，可在傅立叶域中模拟密集、长期的像素轨迹。我们学习到的运动先验可用于将单张图片转化为无缝循环视频，或转化为可响应用户输入（如拖动和释放点）的交互式动态模拟。在右图中，我们将输出视频可视化为时空 X-t 切片（沿左图所示的输入扫描线）。

摘要

我们提出了一种对场景运动进行图像空间先验建模的方法。我们的先验是从真实视频序列中提取的运动轨迹集合中学习的，这些运动轨迹描绘了树木、花朵、蜡烛和随风摇摆的衣服等物体的自然摆动动态。我们将傅立叶域中的密集、长期运动建模为频谱卷，发现它非常适合使用扩散模型进行预测。对于单幅图像，我们训练有素的模型会使用频率协调的扩散采样过程来预测频谱体积，并将其转换为横跨整个视频的运动纹理。与基于图像的渲染模块一起，预测的运动表示法可用于许多下游应用，例如将静态图像转化为无缝循环视频，或允许用户与真实图像中的物体进行交互，产生逼真的模拟动态（通过将光谱体积解释为图像空间模态基）。更多结果请参见我们的项目页面：generative-dynamics.github.io。

1. 导言

自然界总是处于运动之中，即使是看似静止的场景，也会因风、水流、呼吸或其他自然节奏而产生微妙的摆动。模拟这种运动在视觉内容合成中至关重要--人类对运动的敏感性会导致没有运动的图像（或有轻微不真实运动的图像）显得不可思议或不真实。

虽然人类很容易解释或想象场景中的运动，但训练一个模型来学习或产生逼真的场景运动却绝非易事。我们在世界上观察到的运动是场景的基本物理动力学的结果，即施加在物体上的力，这些力会根据物体的独特物理特性--质量、弹性等--做出反应，这些量很难按比例测量和捕捉。幸运的是，在某些应用中并不需要对其进行测量：*例如*，只需分析一些观察到的 2D 运动，就能模拟出场景中可信的动态[23]。

同样的观察到的运动也可以作为学习*跨*场景动态的监督信号--因为尽管观察到的运动是多模态的，并且基于复杂的物理效应，但它往往是可预测的：

蜡烛会以某种方式闪烁，树木会摇摆，树叶会沙沙作响。对于人类来说，这种可预测性在我们的感知系统中根深蒂固：通过观看一幅静止图像，我们可以想象出合理的运动--或者说，由于可能会有很多这样的运动，所以我们可以想象出以该图像为条件的自然运动*分布*。鉴于人类能够对这些分布进行建模，一个自然的研究问题就是对它们进行计算建模。

生成模型，特别是传统扩散模型 [44, 85, 87]的最新进展，使我们能够对丰富的分布进行建模，包括以文本为条件的真实图像分布 [73-75]。这种能力带来了一些新的应用，例如以文本为条件生成多样化的真实图像内容。在这些图像模型取得成功之后，最近的工作又将这些模型扩展到了其他领域，如视频 [7, 43] 和三维几何 [77, 100, 101, 103]。

在本文中，我们为*图像空间场景运动*（即单张图像中所有像素的运动）建立了一个生成先验模型。该模型是根据从大量真实视频序列中自动提取的运动轨迹进行训练的。具体来说，我们从每个训练视频中计算出*频谱体积*形式的运动[22, 23]，频谱*体积*是密集、长距离像素轨迹的频域表示。频谱体积非常适合表现出振荡动态的场景，例如在风中移动的树木和花朵。我们发现，作为场景运动建模扩散模型的输出，这种表示方法也非常有效。我们训练了一个生成模型，该模型以单幅图像为条件，可从学习到的分布中采样光谱体积。然后，预测的光谱体积可直接转化为运动纹理--一组长距离、每个像素的运动轨迹--用于为图像制作动画。光谱体积也可解释为*图像空间模态基础*，用于模拟交互动态[22]。

我们使用扩散模型从输入图像中预测光谱体积，该模型一次生成一个频率的系数，但通过共享注意力模块协调各频段的预测。如图 1 所示，预测的运动可用于合成未来帧（通过基于图像的渲染模型）--将静态图像转化为逼真的动画。

与原始 RGB 像素先验相比，运动先验能捕捉到更基本、更低维度的结构，从而有效解释像素值的长距离变化。因此，生成中间运动会带来更连贯的长期生成和更精细的动画控制。我们在多个下游应用中演示了如何使用我们训练有素的模型，例如创建无缝循环视频、编辑生成的运动以及通过图像空间模态基础实现交互式动态图像，即模拟物体动态对用户施加力的响应[22]。

2. 相关工作

生成合成。生成模式的最新进展使得根据文本提示合成逼真图像成为可能[16, 17, 24, 73-75]。通过将生成的图像张量沿时间维度进行扩展，这些文本到图像的模型可以增强合成视频序列的能力[7, 9, 43, 62, 84, 106, 106, 111]。虽然这些方法能生成捕捉真实片段时空统计的视频序列，但这些视频往往存在运动不连贯、纹理的时间变化不真实、违反物理约束（如质量保证）等问题。

图像动画。其他技术并不完全根据文本生成视频，而是将静态图片作为输入并制作成动画。最近，许多深度学习方法采用 3D Unet 架构直接生成视频卷[27, 36, 40, 47, 53, 93]。这些模型实际上是相同的视频生成模型（但以图像信息而非文本信息为条件），并表现出与上述模型类似的人工痕迹。克服这些限制的一种方法是不直接生成视频内容本身，而是通过基于图像的渲染对输入源图像进行动画处理，即根据外部来源（如驾驶视频 [51, 80-82, 99]、运动或 3D 几何图形优先项 [8, 29, 46, 63-65, 67, 90, 97, 101, 102, 104, 109] 或其他来源）的运动来移动图像内容。用户注释 [6, 18, 20, 33, 38, 98, 105, 108]。根据运动场对图像进行动画处理，可以获得更强的时间连贯性和真实感，但这些先前的方法要么需要额外的引导信号或用户输入，要么利用有限的运动表示。

运动模型和运动先验。在计算机图形学中，自然的振荡三维运动（如水波纹或随风摆动的树木）可以用噪声建模，噪声在傅立叶域中形成，然后转换为时域运动场[79, 88]。其中一些方法依赖于对被模拟系统的基本动态进行模态分析[22、25、89]。Chuang *等人*[20]根据用户注释，将这些频谱技术应用于单张二维图片中的植物、水和云的动画制作。我们的工作尤其受到 Davis [23] 的启发，他将场景的模态分析与该场景视频中观察到的运动联系起来，并利用这种分析从视频中模拟活动间的动态。我们采用了 Davis *等人*的频率空间*频谱体积*运动表示法，从大量训练视频中提取了这一表示法，并证明频谱体积适合用扩散模型预测单个图像的运动。

其他方法在*预测*任务中使用了各种运动表示法，即使用图像或视频来为确定性的未来运动估计提供信息[34, 71]，或为更丰富的可能运动*分布*提供信息[94, 96, 104]。然而

这些方法中的很多都是预测光流运动估计值（即每个像素的瞬时运动），而不是完整的运动轨迹。此外，之前的工作大多集中在活动识别等任务上，而非合成任务。最近的研究表明，使用生成模型对运动进行建模和预测具有以下优势
在一些封闭域环境中，如人类和动物[2, 19, 28, 72, 91, 107]。

作为纹理的视频某些运动场景可被视为一种纹理--称为*动态纹理* [26]，它将视频建模为随机过程的时空样本。动态纹理可以表现波浪、火焰或移动树木等平滑、自然的纹理，已被广泛用于视频分类、分割或编码 [12-15, 76]。与之相关的一种纹理称为*视频纹理*，它将移动场景表示为一组输入视频帧以及任意一对帧之间的过渡概率 [66, 78]。许多方法通过分析场景运动和像素统计来估计动态或视频纹理，目的是生成无缝循环或无限变化的输出视频 [1, 21, 32, 58, 59, 78]。与大部分此类工作不同的是，我们的方法可以提前学习先验，然后将其应用于单个图像。

3. 概述

给定一张图片 I_0，我们的目标是生成一个视频 $\{\hat{I}_1, \hat{I}_{(2)}, \ldots, \hat{I}_T\}$ 具有振荡运动的特征，例如
我们的系统由两个模块组成：运动预测模块和基于图像的渲染模块。我们的系统由两个模块组成：运动预测模块和基于图像的渲染模块。我们的管道首先使用潜在扩散模型（LDM）预测输入 $I_{(0)}$的光谱体积 $S = S_{(f)(0)}, S_{(f)(1)}, \ldots, S_{(f)(K)-1}$。 然后，通过反离散傅立叶变换，将预测的频谱量转换为运动纹理 $F = (F_1, F_2, \ldots, F_T)$。该运动纹理决定了每个输入像素在未来每个时间步长的位置。

给定预测的运动纹理后，我们将使用基于神经图像的渲染技术制作输入 RGB 图像的动画（第 5 节）。我们将在第 6 节中探讨这种方法的应用，包括制作无缝循环动画和模拟交互动态。

4. 运动预测

4.1. 运动表示

从形式上看，运动纹理是一系列随时间变化的二维位移图 $F = \{F_t \mid t = 1, \ldots, T\}$，其中每个像素坐标 $\mathbf{p}$ 处的二维位移向量 $F_{(t)}(\mathbf{p})$来自
输入图像 I_0定义了该像素在未来某一时刻的位置。

时间 t 时该像素的位置 [20]。要在时间 t 生成未来帧，可以使用相应的位移对 $I_{(0)}$中的像素进行拼接

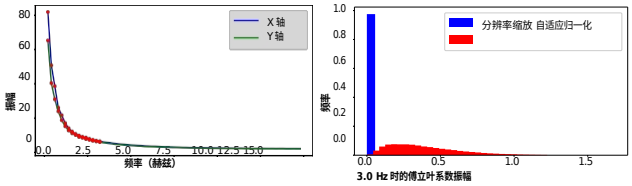

图 2.**左图：**我们将从真实视频中提取的 X 和 Y 运动分量的平均功率谱可视化，显示为蓝色和绿色曲线。自然振荡运动主要由低频成分组成，因此我们使用前 $K = 16$ 项，用红点标记。**右图：**我们展示了 3.0 Hz 频率下傅立叶项振幅的直方图，该直方图是在(1)根据图像宽度和高度缩放振幅（蓝色）或(2)频率自适应归一化（红色）之后绘制的。我们的自适应归一化可防止系数集中在极端值上。

图 D_t，得到前向扭曲图像 I'_t：

$$I'_t(\mathbf{p} + F_t(\mathbf{p})) = I_{(0)}(\mathbf{p}). \quad (1)$$

如果我们的目标是通过运动纹理制作视频，那么一种选择是直接从输入图像预测时域运动纹理。然而，运动纹理的大小需要与视频的长度成比例：生成 T 个输出帧意味着预测 T 个位移场。为了避免在长视频中预测如此大的输出表示，之前的许多动画制作方法要么自动生成视频帧 [7, 29, 57, 60, 93]，要么通过额外的时间嵌入独立预测每个未来输出帧 [4]。然而，这两种方法都无法确保生成视频的长期时间一致性。

幸运的是，许多自然运动可以描述为少量谐波振荡的 叠加，这些谐波振荡以不同的频率、振幅和相位表示 [20, 23, 25, 50, 69]。由于这些基本运动都是准周期性的，因此在频域对其进行建模是很自然的。因此，我们采用了 Davis *等人*[23]提出的视频运动的高效频率空间表示法，即*频谱体*，如图 3 所示。频谱卷是从视频中提取的每像素运动轨迹的时间傅里叶变换。

鉴于这种运动表示，我们将运动预测问题表述为多模态图像到图像的转换任务：从输入图像到输出运动频谱卷。我们采用潜在扩散模型（LDM）来生成由 4K 通道二维运动频谱图组成的频谱卷，其中 $K << T$ 是建模的频率数，在每个频率上，我们需要四个标量来表示 x 维和 *y 维的*复傅里叶系数。请注意，像素在未来时间步长的运动轨迹 $F(\mathbf{p}) = \{F_t(\mathbf{p}) \mid t = 1, 2, \ldots T\}$ 和
其表示为光谱卷 $S(\mathbf{p}) = \{S_f(\mathbf{p}) \mid k =$

$\frac{T}{0, 1, \ldots}$ $_2 - 1\}$ 通过快速傅立叶变换 (FFT) 进行关联：

$$S(\mathbf{p}) = \mathrm{FFT}(F(\mathbf{p})). \quad (2)$$

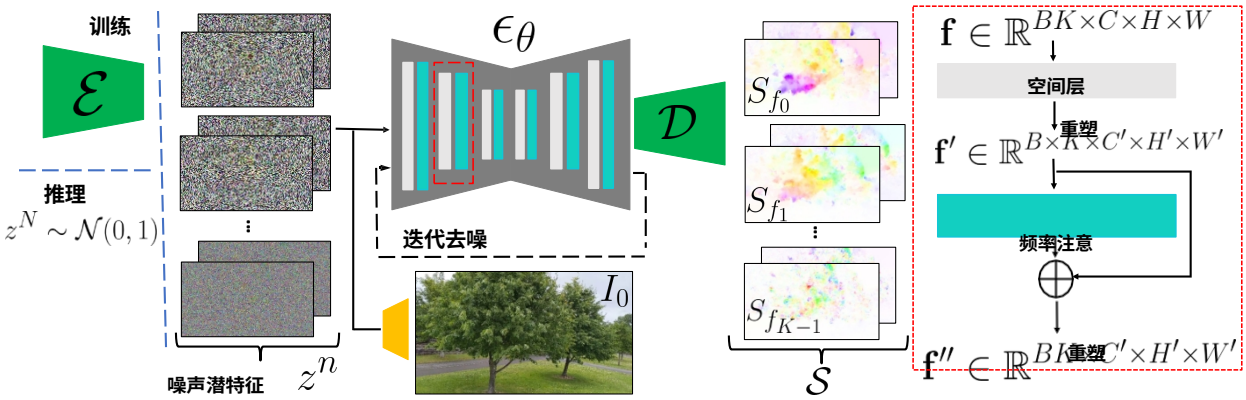

图 3 运动预测模块**运动预测模块。**我们通过频率协调去噪模型预测频谱体积 S。扩散网络$\epsilon_{(\theta)}$的每个区块将二维空间层与注意力层交错在一起（**右侧红框**），并对潜在特征 z^n进行迭代去噪。在训练过程中，我们通过编码器 E 将下采样输入 I_0 与真实运动纹理编码的噪声潜特征连接起来，并在推理过程中用高斯噪声 z^N替换噪声特征（**左图**）。

我们应该如何选择 K 输出频率？先前的实时动画研究发现，大多数自然振荡运动主要由低频成分组成 [25, 69]。为了验证这一观察结果，我们计算了从 1000 个随机取样的 5 秒真实视频片段中提取的运动的平均功率谱。如图 2 左图所示，运动的功率谱随着频率的增加呈指数下降。这表明，大多数自然振荡运动确实可以用低频项来很好地表示。在实践中，我们发现前 K= 16 个傅立叶系数足以真实再现一系列真实视频和场景中的原始自然运动。

4.2. 用扩散模型预测运动

我们选择潜空间扩散模型（LDM）[74] 作为运动预测模块的骨干，因为 LDM 比像素空间扩散模型的计算效率更高，同时还能保持合成质量。标准的 LDM 由两个主要模块组成：(1) 变异自动编码器 (VAE)，通过编码器 z= E(I)将输入图像压缩到潜空间，然后通过解码器 I= D(z)从潜特征重建输入；(2) 基于 U-Net 的扩散模型，从高斯噪声开始学习迭代去噪特征。我们的训练不是将这一过程应用于 RGB 图像，而是应用于来自真实视频序列的光谱卷，对其进行编码，然后按照预先确定的方差表进行 n 步扩散，以产生噪声潜变量 z^n。通过迭代估计噪声$\epsilon_{(\vartheta)}(z^{(n)}; n, c)$，训练二维 U-Nets 对噪声潜点进行去噪处理，用 于 在每一步 n 更新潜点特征$\in$ (1, 2, ..., N)。LDM 的训练损失可写成

$$L_{\text{LDM}} = \mathsf{E}_{(n)\,(\in U)\,[1(N)\,(]),(,\epsilon)^{n}(\in N)\,(0,1)} \left|\left| \epsilon^{(n)} - \epsilon_{(\vartheta)}(z^{(n)}; n, c) \right|\right|^2 \quad (3)$$

其中，c 是任何条件信号的嵌入，例如文本，或者在我们的例子中，训练视频序列的第一帧 I_0。然后，干净的潜特征 z^0将通过解码器来恢复频谱量。

频率自适应归一化。我们发现的一个问题是，运动纹理在不同频率之间具有特殊的分布特征。如图 2 左图所示，频谱量的振幅范围在 0 到 100 之间，随着频率的增加呈近似指数衰减。由于扩散模型要求输出的绝对值在-1 和 1 之间，以便进行稳定的训练和去噪[44]，因此我们必须在使用从真实视频中提取的 S 系数进行训练之前将其归一化。如果我们像之前的工作 [29, 77]那样，根据图像尺寸将这些系数的大小缩放为 [0,1]，那么几乎所有高频率的系数最终都会趋近于零，如图 2 右图所示。根据此类数据训练的模型可能会产生不准确的运动，因为在推理过程中，即使很小的预测误差也会在去规范化后造成很大的相对误差。

为了解决这个问题，我们采用了一种简单而有效的频率自适应归一化方法：首先，我们根据从训练集中计算出的统计数据，对每个频率的傅立叶系数进行独立归一化。也就是说，对于每个频率 f_j， 我们计算所有输入样本中傅立叶系数幅度的 95$^{\text{th}}$百分位数，并使用该值作为每个频率的缩放因子 $s_{(f)\,(j)}$。然后，我们对每个缩放的傅立叶系数进行幂变换，使其远离极端值。在实践中，我们发现平方根比对数或倒数等其他非线性变换效果更好。总之，光谱体积 S(**p**)的最终系数值为

频率 $f_{(j)}$（用于训练 LDM）的计算公式为

$$S'_{f}(\mathbf{p}) = \text{sign}(S_{f}(\mathbf{p})) \frac{|S_{f_{(j)}}(\mathbf{p})|^{\gamma}}{s_f}. \quad (4)$$

如图 2 右图所示，经过频率自适应归一化处理后，频谱体积系数的分布更加均匀。

频率协调去噪。预测具有 K 个频带的光谱体积 S 的直接方法是通过单个扩散 U-Net 输出 $4K$ 个通道的张量。然而，正如之前的研究[7]一样，我们发现训练模型以生成大量通道可能会产生过度平滑、不准确的输出结果。另一种方法是通过在 LDM 中注入额外的频率嵌入来独立预测每个频率片[4]，但这种设计选择会导致频域中的预测不相关，从而导致不切实际的运动。因此，受近期视频扩散工作[7]的启发，我们提出了一种频率协调去噪策略，如图 3 所示。具体来说，给定一幅输入图像 I_0，我们首先训练一个 LDM $\epsilon_{(\theta)}$，以预测频谱量 $S_{(f)(j)}$ 的单个 4 通道频率切片，在 LDM 中注入额外的频率嵌入和时间步长嵌入。然后，我们冻结该 LDM 的参数 $\epsilon_{(\theta)}$，在 K 个频段上引入与 $\epsilon_{(\theta)}$ 的二维空间层交错的注意力层，并进行微调。具体来说，对于批次大小 B，$\epsilon_{(}^{\theta}{}_{j)}$ 的二维空间层会将信道大小 C 的相应 $B \cdot K$ 噪声潜特征视为独立样本，其形状为 $R^{((B)(\cdot)(K)())\times(C)\times H\times(W)}$。然后，注意力层会将这些特征解释为跨越频率轴的连续特征，我们会重塑潜在特征的形状。在将之前二维空间层的特征输入注意力层之前，先将这些特征输入 $R^{B\times(K)\times C\times H\times(W)}$。换句话说，频率注意层经过微调，以协调所有频率切片，从而产生连贯的频谱卷。在我们的实验中，我们发现当我们从单一的二维 U-Net 转向频率协调去噪模块时，平均 VAE 重建误差从 0.024 减小到了 0.018，这表明 LDM 预测精度的上限得到了改善；在第 7.3 节中，我们还表明这种设计选择提高了视频生成质量。

5. 基于图像的渲染

我们首先利用应用于每个像素的反时间 FFT $\mathrm{F}_{(}\mathbf{p}) = \mathrm{FFT}^{-(1)}(S(\mathbf{p}))$，在时域中推导出运动纹理。为了生成未来帧 $\hat{I}_t$，我们采用了一种基于深度图像的渲染技术，并使用预测的运动场 $F_{(t)}$ 进行拼接，以向前翘曲编码后的 I_0，如图 4 所示。由于前向扭曲会导致
孔，而且多个源像素可以映射到同一个输出端。

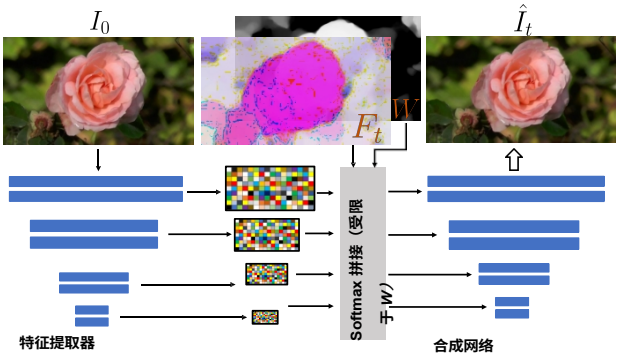

图 4 渲染模块**渲染模块。**我们使用基于深度图像的渲染模块来填充缺失的内容，并重新精细翘曲的输入图像。然后，在这些特征上应用 Softmax 拼接技术，再加上从 0 到 t 的运动场 F_t（取决于权重 W）。经扭曲的特征被输入图像合成网络，生成
渲染图像 I_t。

在二维位置上，我们采用了之前帧插值工作中提出的特征金字塔软最大拼接策略[68]。具体来说，我们通过特征提取网络对 $I_{(0)}$ 进行编码，生成多尺度特征图。对于尺度为 j 的每个单独特征图，

我们会根据分辨率调整预测二维运动场 F_t 的大小和尺度。与 Davis 等人的研究[22]一样，我们使用预测的流量大小作为深度代理，以确定映射到目的地位置的每个源像素的贡献权重。具体来说，我们计算每个像素的权重 $W(\mathbf{p}) = \frac{1}{T} \sum_{t} ||F_t(\mathbf{p})||_2$ 作为预测运动纹理的平均值。换句话说，我们假设大的运动对应于移动的前景物体，而小的或零的运动对应于背景。我们使用源自运动的权重，而不是像 [46] 那样使用可学习的权重，因为我们发现在单视角情况下，可学习的权重并不能有效地解决以下问题不确定性。

利用运动场 F_t 和权重 W，我们采用软最大拼接技术对每个尺度上的特征图进行翘曲，以生成翘曲特征。然后将翘曲特征注入图像合成解码器的相应块中
生成最终的渲染图像 $\hat{I}_t$。

我们从真实视频中随机抽取起始帧和目标帧（I_0, I_t），利用从 I_0 到 $I_{(t)}$ 的估计流场来联合训练特征提取器和合成网络。I_t 来翘曲 I_0 的编码特征，并利用 VGG 感知损失针对 I_t 监督预测 $\hat{I}_t$[49]。

6. 应用

图像到视频。我们的系统首先从输入图像中预测运动光谱体积，然后将基于图像的渲染模块应用于从光谱体积转换而来的运动纹理，从而生成动画。

生成动画。由于我们对场景运动进行了精确建模，因此可以通过对运动纹理进行线性插值来生成慢动作视频，或通过调整预测光谱体积系数的振幅来放大（或缩小）动画运动。

无缝循环。许多应用都需要无缝循环的视频，即视频的开始和结束之间没有间断。遗憾的是，很难找到大量的无缝循环视频来进行训练。相反，我们设计了一种方法，使用在常规非循环视频片段上训练的运动扩散模型来制作无缝循环视频。受近期图像编辑引导工作的启发[3, 30]，我们的方法是一种*运动自引导*技术，利用明确的循环约束引导运动去噪采样处理。具体来说，在推理过程中的每个迭代去噪步骤中，我们都会在标准的无分类器引导信号（classifier-free guidance）[45]之外加入一个额外的运动引导信号，强制要求每个像素在开始帧和结束帧的位置和速度尽可能相似：

$$\epsilon\hat{}^{n} = (1+w)\epsilon_\theta(z^{(n)}; n, c) - w\epsilon_\theta(z^{(n)}; n, \varnothing) + u\sigma^{(n)}\nabla_{z^n} L_g^n$$

$$L_g = \| F^n - \overleftarrow{F}^n \|_1 + \| \nabla F^{(n)} - \nabla \overleftarrow{F}^{(n)} \|_1 \quad (5)$$

其中，F^n是时间 t 和去噪步骤 n 时的预测二维位移场，w 是无分类器引导权重，u 是运动自引导权重。在补充视频中，我们采用基于外观的基线循环算法 [58]，从非循环输出生成循环视频，结果表明我们的运动自引导技术能生成失真更少、伪像更少的无缝循环视频。

单幅图像的互动动力学。Davis *等人*[22]的研究表明，在特定共振频率下评估的频谱体量可以近似得到*图像空间模态基础*，该基础是底层场景振动模式的投影（或者更广泛地说，捕捉振荡动力学中的空间和时间相关性），可用于模拟物体对用户定义的力的响应。我们采用这种模态分析方法[22, 70]，将物体物理响应的图像空间二维运动位移场写成运动频谱系数 $S_{(f)(j)}$的加权和，该系数由每个模拟时间步长 t 的复模态坐标 $\mathbf{q}_{(f)(j)(t)}$状态调制：

$$F_{(t)}(\mathbf{p}) = \sum_f s_{(f)(j)(p)} \mathbf{q}_{(f)(j)(t)} \quad (6)$$

我们通过明确的欧拉方法模拟模态坐标 $\mathbf{q}_{(f)(j)(t)}$的状态，该方法适用于模态空间中表示的解耦质量-弹簧-阻尼系统的运动方程[22, 23, 70]。有关完整推导过程，请读者参阅补充材料和原著。需要注意的是，我们的方法*只*需一张*图片*即可生成一个交互场景，而之前的方法则需要视频作为输入。

	图像合成		视频合成			
方法	fid	kid	fvd	fvd_{32}	dtfvd	$dtfvd_{32}$
TATS[35]	65.8	1.67	265.6	419.6	22.6	40.7
随机 I2V [27]	68.3	3.12	253.5	320.9	16.7	41.7
MCVD [93]	63.4	2.97	208.6	270.4	19.5	53.9
LFDM [67]	47.6	1.70	187.5	254.3	13.0	45.6
DMVFN [48]	37.9	1.09	206.5	316.3	11.2	54.5
远藤等人[29]	10.4	0.19	166.0	231.6	5.35	65.1
Holynski *et al.*	11.2	0.20	179.0	253.7	7.23	46.8
我们的	**4.03**	**0.08**	**47.1**	**62.9**	**2.53**	**6.75**

表 1.**测试集的定量比较。**我们同时报告了图像合成和视频合成的质量。此处，KID 按 100 的比例缩放。对所有误差而言，越低越好。有关基线和误差指标的说明，请参见第 7.1 节。

7. 实验

实施细节。我们使用 LDM [74] 作为预测频谱量的骨干，为此我们使用了维度为 4 的连续潜空间 VAE。我们训练 VAE 采用 L_1重建损失、多尺度梯度一致性损失 [54-56] 和 KL-发散损失，权重分别为 1、0.2、10^{-6}。我们训练相同的 2D 在最初的 LDM 工作中使用的 U-Net 以简单的 MSE 损失进行迭代去噪[44]，并采用[41]中的衰减层进行频率协调去噪。为了进行定量评估，我们在尺寸为 256× 160 的图像上对 VAE 和 LDM 进行了从头开始的训练，以进行公平的比较，使用 16 个 Nvidia A100 GPU，大约需要 6 天才能收敛。为了获得主要的定量和定性结果，我们使用 DDIM [86] 运行了 250 步运动扩散模型。我们还展示了生成的分辨率高达 512× 288 的视频，这些视频是通过在我们的数据集上微调预先训练好的图像绘制 LDM 模型[74]而生成的。

我们的 IBR 模块采用 ResNet-34 [39] 作为特征提取器。我们的图像合成网络基于条件图像内绘的架构[57, 110]。在推理过程中，我们的渲染模块在 Nvidia V100 GPU 上以 25FPS 的速度实时运行。我们采用通用引导法[3]来制作无缝循环视频，设置权重 $w = 1.75$，$u = 200$，并使用 500 DDIM 步数和 2 次自我递归迭代。

数据。我们从网上收集并处理了 3,015 个自然场景视频，这些视频都是由我们自己捕捉的。我们保留 10%的视频用于测试，其余用于训练。为了提取真实的运动轨迹，我们在每个选定的起始图像和视频的每个未来帧之间采用了一种从粗到细的流动方法 [10, 61]。作为训练数据，我们将每 10 个视频帧作为输入图像，并利用计算出的后续 149 个帧的运动轨迹得出相应的地面真实光谱卷。我们的数据总共包括超过 15 万个图像-运动对。

输入图像　　参考　　随机-I2V [27]　　MCVD [93]　　*远藤等人* [29]　　Holynski *等人* [46]　　我们的

图 5.**不同方法生成的 *X-t* 视频切片。**从左至右：输入图像和相应的 X-t 视频切片，分别来自地面实况视频、三种基线方法生成的视频 [27, 29, 46, 93]，以及我们的方法生成的视频。

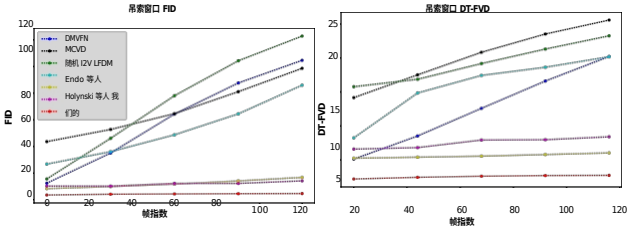

图 6.**滑动窗口 FID 和 DTFVD。**图中显示了针对不同方法生成的视频，窗口大小为 30 帧的滑动窗口 FID 和窗口大小为 16 帧的 DTFVD。

基准。我们将我们的方法与最近的单图像动画和视频预测方法进行了比较。Endo *等人* [29] 和 DMVFN [48] 预测瞬时二维运动场，并自动递归渲染未来帧。Holynski *等人*[46] 则通过单一的静态 Eule- rian 运动描述来模拟运动。其他最新研究，如 Stochas- tic Image-to-Video (Stochastic-I2V) [27]、TATS [35] 和 MCVD [93]，则采用 VAE、变换器或扩散模型来直接预测原始视频帧；LFDM [67] 则通过在扩散模型中预测流量和扭曲潜变量来生成未来帧。我们使用上述方法各自的开源实现对数据进行训练。

我们通过两种方法对我们的方法和之前的基线生成的视频质量进行评估。首先，我们使用为图像合成任务设计的指标来评估单个合成帧的质量。我们采用 Fre´chet Inception Distance（FID）[42] 和 Kernel Inception Disance（KID）[5] 来测量生成帧的分布与地面实况帧之间的平均距离。

第二，评估质量和时间一致性

[1] 我们使用 Fan *等人*[83]的开源实现[46]。

方法	图像合成 FID	KID	视频合成 FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
重复 I_0	-	-	237.5	316.7	5.30	45.6
$K=4$	3.92	0.07	60.3	78.4	3.12	8.59
$K=8$	3.95	0.07	52.1	68.7	2.71	7.37
$K=24$	4.09	0.08	48.2	65.1	2.50	6.94
无自适应规范。	4.53	0.09	62.7	80.1	3.16	8.19
独立预测。	4.00	0.08	52.5	71.3	2.70	7.40
体积预测值	4.74	0.09	53.7	71.1	2.83	7.79
基线溅射[46]	4.25	0.09	49.5	66.8	2.83	7.27
满员 ($K=16$)	4.03	0.08	47.1	62.9	2.53	6.75

表 2.**消融研究。**第 7.3 节介绍了每种配置。

我们采用窗口大小为 16（FVD）和 32（FVD_{32})的弗雷谢特视频间距[92]来测量合成视频的距离，该方法基于在人类动力学数据集[52]上训练的 I3D 模型[11]。为了更忠实地反映我们试图生成的自然摆动运动的合成质量，我们还采用了动态纹理弗雷谢特视频距离法[27]，该方法使用在动态纹理数据库[37]（一个主要由自然运动纹理组成的数据集）上训练的 I3D 模型，测量窗口大小为 16（DTFVD）和 32（$DTFVD_{32}$) 的视频的距离。

此外，我们还使用窗口大小为 30 帧的滑动窗口 FID 和窗口大小为 16 帧的滑动窗口 DTFVD（如文献 [57, 60] 所述）来衡量生成的视频质量随时间的变化情况。对于所有方法，我们通过中心裁剪在 256× 128 分辨率下进行评估。

输入 AnimateDiff ModelScope GEN-2

图 7.我们展示了最近三个大型视频扩散模型 [31, 36, 98] 生成的未来帧。

图 8.**局限性。**我们展示了渲染的未来帧（偶数）和输入图像与渲染图像的叠加（奇数）。我们的方法会在物体较薄或运动较大的区域，以及需要填充大量新内容的区域产生伪影。

7.1. 定量结果

表 1 显示了我们的方法与基线在测试集上的量化比较。在图像和视频合成质量方面，我们的方法明显优于之前的单图像动画基线。具体而言，我们的 FVD 和 DT-FVD 距离更小，这表明我们的方法生成的视频更逼真，时间上更连贯。此外，图 6 显示了不同方法生成的视频的滑动窗口 FID 和滑动窗口 DT-FVD 距离。由于采用了全局频谱体量表示法，我们的方法生成的视频不会随着时间的推移而质量下降。

7.2. 定性结果

我们将视频之间的定性比较可视化为生成视频的 X-t 时空切片，这是可视化视频中微小运动的标准方法[95]。如图 5 所示，与其他方法相比，我们生成的视频动态与相应真实参考视频（第二列）中观察到的运动模式更为相似。诸如 Stochastic I2V [27] 和 MCVD [93] 等基准方法无法同时对外观和运动进行随时间变化的真实建模。Endo 等人[29] 和 Holynski 等人[46] 所生成的视频帧具有较少的伪影，但随着时间的推移会表现出过度平滑或非振荡的运动。请读者参阅补充材料，以评估不同方法生成的视频帧和估计运动的质量。

7.3. 消融研究

我们进行了一项消融研究，以验证运动预测和渲染模块中的主要设计选择，并将完整配置与不同变体进行比较。具体来说，我们评估了使用不同频段数 $K = 4、8、16、24$ 的结果。我们观察到，增加频段数可提高视频预测质量，但频段数超过 16 个时，提高幅度微乎其微。接下来，我们从地面实况频谱卷中移除自适应频率正则化，取而代之的是根据输入图像的宽度和高度对其进行缩放（*无自适应正则化*）。此外，我们还移除了频率协调去噪模块（*独立预测*），或将其替换为更简单的 DM，即通过单个二维 U 网扩散模型联合预测 4K 信道光谱体积的张量体积（*体积预测*）。最后，我们比较了使用基线渲染方法的结果。

我们在单尺度特征上应用了软最大拼接（softmax splatting）技术，并采用了可学习权重（如 [46] 所用）（*基线拼接*）。我们还增加了一条基线，即通过重复输入图像 N 次（重复 $I_{(0)}$），生成的视频是一个卷。从表 2 中我们可以看出，与我们的完整模型相比，所有更简单的配置或替代配置都会导致性能下降。

7.4. 与大型视频模型的比较

我们进一步进行了用户研究，并将我们生成的动画与近期大型视频扩散模型的动画进行了比较：AnimateDiff [36]、ModelScope [98] 和 Gen-2 [31]，它们都能直接预测视频量。我们从测试集中随机抽取了 30 个视频，询问用户 "哪个视频更逼真？与其他方法相比，用户对我们的方法的偏好度高达 80.9%。此外，如图 7 所示，我们观察到这些基线生成的视频要么无法与输入的图像内容保持一致，要么随着时间的推移逐渐出现色彩漂移和失真。请读者参阅补充材料，了解全面的比较结果。

8. 讨论和结论

局限性。由于我们的方法只能预测频谱体的较低频率，因此可能无法模拟非摆动运动或高频振动--这可以通过使用学习的运动基础来解决。此外，生成视频的质量取决于底层运动轨迹的质量，而底层运动轨迹在有细小运动物体或大位移物体的场景中可能会降低质量。需要生成大量新的未见内容的运动即使正确，也可能导致质量下降（图 8）。

结论我们提出了一种从单张静态图片中建立自然振荡动态模型的新方法。我们的图像空间运动先验用频谱体积来表示，频谱体积是每像素运动轨迹的频率表示，我们发现它对于使用扩散模型进行预测非常有效，而且我们是从真实世界的视频集合中学习的。利用频率协调的潜在扩散模型对频谱体积进行预判，并通过基于图像的渲染模块将其用于未来视频帧的动画制作。我们的研究结果表明，我们的方法能从单张图片中生成逼真的动画，其效果明显优于之前的基线方法，而且还能支持多种下游应用，如创建无缝循环或交互式图像动态。

致谢。我们感谢 Abe Davis、Rick Szeliski、Andrew Liu、Boyang Deng、Qianqian Wang、Xuan Luo 和 Lucy Chai 富有成效的讨论和有益的意见。

参考文献

[1] Aseem Agarwala、Ke Colin Zheng、Chris Pal、Maneesh Agrawala、Michael Cohen、Brian Curless、David Salesin 和 Richard Szeliski。全景视频纹理*ACM Trans.Graphics (SIGGRAPH)*，第 821-827 页。2005.

[2] Hyemin Ahn、Esteve Valls Mascaro 和 Dongheui Lee。Can we use diffusion probabilistic models for 3d motion prediction? *arXiv preprint arXiv:2302.14503*, 2023.

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild、Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein.扩散模型的通用指导。在 *Proc.计算机视觉与模式识别》（CVPR）*，第 843-852 页，2023 年。

[4] Hugo Bertiche, Niloy J Mitra, Kuldeep Kulkarni, Chun-Hao P Huang, Tuanfeng Y Wang, Meysam Madadi, Sergio Escalera, and Duygu Ceylan.随风飘扬：从静态图像中提取人体电影镜头的 Cy-clenet。In *Proc.计算机视觉与模式识别》（CVPR）*，第 459-468 页，2023 年。

[5] Mikołaj Bin´kowski、Danica J Sutherland、Michael Arbel、和 Arthur Gretton.解密 MMD GANs。*arXiv preprint arXiv:1801.01401*, 2018.

[6] Andreas Blattmann、Timo Milbich、Michael Dorkenwald、ipoke：戳静态图像，实现可控随机视频合成。In *Proc.计算机 Vision (ICCV)*，第 14707-14717 页，2021 年。

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dock-horn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis.对齐你的潜影：使用帐篷扩散模型的高分辨率视频合成。在 *Proc.Computer Vision and Pattern Recognition (CVPR)*, pages 22563-22575, 2023.

[8] Richard Strong Bowen、Richard Tucker、Ramin Zabih 和诺亚-斯纳韦利运动的维度：通过流动子空间进行单目预测。*国际 3D 视觉会议（3DV）*，第 454-464 页，2022 年。

[9] Tim Brooks、Janne Hellsten、Miika Aittala、Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras.生成动态场景的长视频*神经信息处理系统》*，35:31769- 31781，2022。

[10] Thomas Brox、Andre´s Bruhn、Nils Papenberg 和 Joachim Weickert.基于扭曲理论的高精度光流估计。In *Proc.欧洲计算机会议 Vision (ECCV)*，第 25-36 页，2004 年。

[11] Joao Carreira 和 Andrew Zisserman.Quo vadis, action 新模型和动力学数据集。In *Proc.计算机视觉与模式识别》（CVPR）*，6299-6308 页，2017 年。

[12] Dan Casas、Marco Volino、John Collomosse 和 Adrian 希尔顿用于交互式角色外观的 4d 视频纹理。*计算机图形论坛》*，第 33 卷，第 371-380 页。威利在线图书馆，2014 年。

[13] Antoni B Chan 和 Nuno Vasconcelos。动态纹理混合物。纹理混合物。In *Proc.计算机视觉会议（ICCV）*、第 641-647 页，2005 年。

[14] Antoni B Chan 和 Nuno Vasconcelos.用内核动态纹理对视频进行分类在 *Proc.计算机视觉与 模式识别（CVPR）*，2007 年。

[15] Antoni B Chan 和 Nuno Vasconcelos。建模、聚类动态纹理混合物视频建模、聚类和分割。*Trans.模式分析与机器智能》*，30（5）：909- 926，2008 年。

[16] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot、Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse： *arXiv preprint arXiv:2301.00704*, 2023.

[17] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman.Maskgit：屏蔽生成式图像变换器。在 *Proc.计算机视觉与模式识别》（CVPR）*，第 11315-11325 页，2022 年。

[18] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang.用于可控视频合成的运动条件扩散模型》。*ArXiv 预印本 arXiv:2304.14404*, 2023.

[19] 陈昕、蒋彪、刘文、黄子龙、傅斌、陶涛、陈昕、蒋彪、刘文、黄子龙、傅斌、陶涛。Chen, and Gang Yu.通过潜空间运动扩散执行命令。In *Proc.计算机视觉与 模式识别（CVPR）*，第 18000-18010 页，2023 年。

[20] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H Salesin, and Richard Szeliski.用随机运动纹理制作动画图片。In *ACM Trans.Graphics (SIGGRAPH)*，第 853-860 页，2005 年。

[21] Vincent C Couture、Michael S Langer 和 Sebastien Roy。无重影的全立体视频纹理。 *3D Vision 国际会议*，第 64-70 页。IEEE，2013。

[22] Abe Davis、Justin G Chen 和 Fre´do Durand。图像空间视频中物体可信操作的模态基础。*ACM Trans.Graphics (SIGGRAPH)*, 34(6):1-7, 2015.

[23] 迈尔斯-亚伯拉罕-戴维斯*视觉振动分析*。 博士论文，麻省理工学院，2016 年。

[24] 普拉富拉-达里瓦尔和亚历山大-尼克尔。扩散模型图像合成中的 "节拍"。*神经信息处理 ing Systems*，34：8780-8794，2021。

[25] Julien Diener、Mathieu Rodriguez、Lionel Baboud 和 Li-onel Reveret.树木实时动画的风投影基础。*计算机图形论坛》*第 28 卷， 533-540 页。威利在线图书馆，2009 年。

[26] Gianfranco Doretto、Alessandro Chiuso、Ying Nian Wu 和 Stefano Soatto.动态纹理》。51:91-109, 2003.

[27] Michael Dorkenwald、Timo Milbich、Andreas Blattmann、Robin Rombach、Konstantinos G. Derpanis 和 Bjorn Om- mer。使用 cinns 的随机图像到视频合成。在 *Proc.计算机视觉与模式识别》（CVPR）*，第 3742-3753 页，2021 年 6 月。

[28] 杜玉明、罗宾-基普斯、阿尔伯特-普马罗拉、塞巴斯蒂安-斯塔克、阿里-塔贝特和阿尔乔姆-萨纳科耶乌、Ali Thabet 和 Artsiom Sanakoyeu。头像长腿：利用扩散模型从稀疏跟踪中生成平滑的人体运动。在 *Proc.Computer Vision and Pattern Recognition (CVPR)*，第 481-490 页，2023 年。

[29] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. 动画景观：用于单图像视频合成的运动和外观解码自监督学习。

sis.*ACM Trans.Graphics (SIGGRAPH Asia)*, 38(6):175:1- 175:19, 2019.

[30] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski.控制图像生成的扩散自引导。*arXiv 预印本 arXiv:2306.00986*, 2023.

[31] Patrick Esser, Johnathan Chiu, Parmida Atighehchian、Jonathan Granskog 和 Anastasis Germanidis.用扩散模型进行结构和内容引导的视频合成。In *Proc.计算机视觉会议（ICCV）*，第 7346- 7356 页，2023 年。

[32] Matthew Flagg、Atsushi Nakazawa、Qiushuang Zhang、Sing Bing Kang、Young Kee Ryu、Irfan Essa 和 James M Rehg、Sing Bing Kang, Young Kee Ryu, Irfan Essa, and James M Rehg.人类视频纹理*2009 年交互式 3D 图形和游戏研讨会论文集*》，199-206 页，2009 年。

[33] Jean-Yves Franceschi、Edouard Delasalles、Mickae¨l Chen、Sylvain Lamprier, and Patrick Gallinari.随机潜在残差视频预测。*机器学习国际会议*，第 3233-3246 页。PMLR, 2020.

[34] Ruohan Gao、Bo Xiong 和 Kristen Grauman。Im2Flow：用于动作识别的静态图像运动幻觉。In *Proc.计算机视觉与模式识别(CVPR)*，2018。

[35] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh.使用时间不可知的 vqgan 和时间敏感变换器生成长视频。*arXiv 预印本 arXiv:2204.03638*, 2022.

[36] 郭玉伟、杨采元、饶安义、王耀辉、于乔宇、林大华和戴波。动画扩散：无需特定 调整的个性化文本到图像扩散模型动画。*arXiv 预印本 arXiv:2307.04725*, 2023.

[37] Isma Hadji 和 Richard P Wildes.新的大规模动态纹理数据集namic texture dataset with application to convnet understand- ing.In *Proc.欧洲计算机视觉会议（ECCV）*，第 320-335 页，2018 年。

[38] Zekun Hao, Xun Huang, and Serge Belongie.可控稀疏轨迹视频生成在 *Proc.计算机视觉与模式识别》（CVPR）*，第 7854-7863 页，2018。

[39] 何开明、张翔宇、任少清和孙健。图像识别的深度残差学习In *Proc.Com- puter Vision and Pattern Recognition (CVPR)*, pages 770- 778, 2016.

[40] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun、Animate-a-story：用检索增强视频生成讲故事。*arXiv 预印本 arXiv:2307.06940*, 2023.

[41] 何颖清、杨天宇、张勇、单莹和陈奇峰。用于任意长度高保真视频生成的潜在视频扩散模型。*arXiv 预印本 arXiv:2211.13221*, 2022.

[42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bern-Hard Nessler 和 Sepp Hochreiter.用双时间尺度更新规则训练的甘斯收敛于局部纳什均衡*神经信息处理系统*》，2017年30期。

[43] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang、Imagen video：使用扩散模式生成高清视频 els. *arXiv preprint arXiv:2210.02303*, 2022.

[44] Jonathan Ho, Ajay Jain, and Pieter Abbeel.去噪差异sion probabilistic models.*神经信息处理 系统*，33:6840-6851，2020。

[45] Jonathan Ho 和 Tim Salimans.无分类器扩散 *arXiv preprint arXiv:2207.12598*, 2022.

[46] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski.用欧拉运动场制作动画图片在 *Proc.计算机视觉与模式识别(CVPR)*，第 5810-5819 页，2021 年。

[47] Tobias Hoppe、Arash Mehrjou、Stefan Bauer、Didrik Nielsen、和 Andrea Dittadi。用于视频预测的扩散模型 and infilling.*Trans.Mach.Learn.Res.*, 2022, 2022.

[48] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou.用于视频预测的动态多尺度体素流网络 *.ArXiv*，abs/2303.09875，2023。

[49] 贾斯汀-约翰逊、亚历山大-阿拉希和李菲菲。感知实时风格转移和超分辨率的损失。In *Proc.欧洲计算机视觉会议（ECCV）*， 694-711 页，2016 年。

[50] Hitoshi Kanda and Jun Ohya.高效、逼真的三维植物树动态行为动画制作方法。*In- ternational Conference on Multimedia and Expo*, volume 2, pages II-89.IEEE, 2003.

[51] Johanna Karras、Aleksander Holynski、Ting-Chun Wang、和 Ira Kemelmacher-Shlizerman.Dreampose：通过稳定扩散进行时尚图像到视频合成。*ArXiv 预印本 arXiv:2304.06025*, 2023.

[52] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang、Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu- man action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[53] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel、Chelsea Finn 和 Sergey Levine.随机对抗 视频预测。*arXiv preprint arXiv:1804.01523*, 2018.

[54] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker、Noah Snavely, Ce Liu, and William T Freeman.通过观察凝固的人学习移动人的深度。在 *Proc.计算机视觉与模式识别》（CVPR）*，第 4521-4530 页，2019 年。

[55] Zhengqi Li and Noah Snavely.Megadepth：学习单从网络照片中预测视图深度In *Proc.Com- puter Vision and Pattern Recognition (CVPR)*, pages 2041- 2050, 2018.

[56] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker、and Noah Snavely.Dynibar：基于神经的动态图像渲染。在 *Proc.Computer Vision and Pattern Recogni- tion (CVPR)*，第 4273-4284 页，2023 年。

[57] 李正奇、王向前、Noah Snavely 和 Angjoo金泽无限自然零：从单张图像学习自然场景的永久视图生成。In *Proc.欧洲计算机视觉会议（ECCV）*，第 515-534 页。534.Springer, 2022.

[58] Jing Liao, Mark Finch, and Hugues Hoppe.无缝视频循环的快速计算。*ACM Trans.Graphics (SIG-*

graph), 34(6):1-10, 2015.

[59] Zicheng Liao, Neel Joshi, and Hugues Hoppe.渐进动态的自动视频循环。*ACM Transac- tions on Graphics (TOG)*, 32(4):1-10, 2013.

[60] Andrew Liu、Richard Tucker、Varun Jampani、Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa.无限自然：从单张图像生成自然场景的永久视图。In *Proc.计算机视觉会议（ICCV）*，14458-14467 页，2021 年。

[61] 刘策。*超越像素：探索运动分析的新表征和应用。运动分析的新表征和应用。*博士论文，麻省理工学院，2009 年。

[62] 罗正雄、陈大有、张颖雅、黄艳、王亮、沈玉军、张颖雅、黄艳、王亮、沈玉军、赵德利、周敬仁、谭铁牛。视频融合：用于生成高质量视频的分解扩散模型。在 *Proc.计算机视觉与 模式识别》（CVPR）*，第 10209-10218 页，2023 年。

[63] Aniruddha Mahapatra 和 Kuldeep Kulkarni.可控静态图像中的流体元素动画。In *Proc.计算机视觉与模式识别（CVPR）*，2022年。

[64] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee、Sergey Tulyakov 和 Jun-Yan Zhu.文本引导合成的优勒电影胶片。2023.

[65] Arun Mallya、Ting-Chun Wang 和 Ming-Yu Liu.隐式图像集动画的内隐翘曲*神经信息 处理系统》*，35:22438-22450，2022。

[66] Medhini Narasimhan、Shiry Ginosar、Andrew Owens、Alexei A Efros, and Trevor Darrell.跟着节拍敲击：音频条件对比视频纹理。在 *Proc.计算机视觉应用大会》（Win- ter Conference on Applications of Computer Vision）*，第 3761-3770 页，2022 年。

[67] 倪浩淼、施昌皓、李凯、黄绍龙马丁-任强民利用潜流扩散模型生成条件图像到视频 In *Proc.Computer Vi- sion and Pattern Recognition (CVPR)*，第 18444-18455 页，2023。

[68] Simon Niklaus 和 Feng Liu.用于视频帧插值。在 *Proc.计算机视觉与模式 Recognition (CVPR)*，第 5437-5446 页，2020 年。

[69] Shin Ota, Machiko Tamura, Kunihiko Fujita, T Fujimoto、K Muraoka, and Norishige Chiba.基于1/f/sup/spl beta//噪声的风场树木摇摆实时动画。*国际计算机图形学论文集》*，第 52-59 页。IEEE, 2003.

[70] Automne Petitjean, Yohan Poirier-Ginter, Ayush Tewari、Guillaume Cordonnier 和 George Drettakis。Modalnerf：用于动态振动场景中自由视点导航的神经模态分析与合成。*计算机图形学论坛》（Computer Graph- ics Forum）*，第42卷，2023页。

[71] Silvia L. Pintea、Jan C. van Gemert 和 Arnold W. M. Smeulders.似曾相识：静态图像中的运动预测。在*欧洲计算机视觉会议（European Conf.欧洲计算机视觉会议（ECCV）*，2014 年。

[72] Sigal Raab、Inbal Leibovitch、Guy Tevet、Moab Arar、Amit H Bermano 和 Daniel Cohen-Or。单次运动扩散。 *arXiv preprint arXiv:2302.05905*, 2023.

[73] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu、and Mark Chen.带剪辑潜变量的分层文本条件图像生成。*ArXiv 预印本 arXiv:2204.06125*, 1(2):3, 2022.

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjo¨ rn Ommer.潜在扩散模型的高分辨率图像合成。In *Proc.Computer Vi- sion and Pattern Recognition (CVPR)*, pages 10684-10695, 2022.

[75] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *神经信息处理系统》（Neural Information Process- ing Systems）*，35:36479-36494，2022。

[76] Payam Saisan、Gianfranco Doretto、Ying Nian Wu 和 Stefano Soatto.动态纹理识别。在 *Proc. puter Vision and Pattern Recognition (CVPR)*, 2001.

[77] Saurabh Saxena、Charles Herrmann、Junhwa Hur、Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet.扩散模型在光学 流和单目深度估计中的惊人效果，2023.

[78] Arno Scho¨ dl, Richard Szeliski, David H Salesin, and Irfan 埃萨。视频纹理。In *ACM Trans.Graphics (SIGGRAPH)*, 第 489-498 页，2000 年。

[79] 随机运动--风力影响下的运动。风力影响下的运动。*计算机图形 论坛》*，11(3)，1992 年。

[80] Aliaksandr Siarohin, Ste´phane Lathuilie`re, Sergey Tulyakov、Elisa Ricci 和 Nicu Sebe.通过深度运动转移实现任意物体动画在 *Proc.计算机视觉与模式识别》（CVPR）*，第 2377-2386 页，2019 年。

[81] 阿利亚克山大-西亚罗欣、斯蒂芬-拉图伊列尔、谢尔盖-图利亚科夫、伊丽萨-里奇和尼库-塞贝Elisa Ricci 和 Nicu Sebe.图像动画的一阶运动模型*神经信息处理系统》*， 32, 2019.

[82] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 关节动画的运动表示。在 *Proc.Computer Vision and Pattern Recognition (CVPR)*, pages 13653-13662, 2021.

[83] 陈谦 Kwan-Yee Lin Hongsheng Li Siming Fan, Jing- tan Piao. tan Piao.模拟真实世界静态图像中的流体。*arXiv 预印本*，arXiv:2204.11335，2022。

[84] 伊万-斯科罗霍多夫、谢尔盖-图利亚科夫、穆罕默德-埃尔霍seiny。Stylegan-v：具有 stylegan2 的价格、图像质量和优点的连续视频生成器。In *Proc.计算机视觉与模式识别（CVPR）*，第 3626- 3636 页，2022 年。

[85] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan、和 Surya Ganguli. 使用非平衡热力学的深度无监督学习。In *International confer- ence on machine learning*, pages 2256-2265.PMLR，2015。

[86] 宋家明、孟晨霖和斯特凡诺-埃尔蒙。Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[87] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole.通过随机微分方程进行基于分数的生成建模 *tions. arXiv preprint arXiv:2011.13456*, 2020.

[88] Jos Stam.*复杂自然环境的多尺度随机建模现象。*博士论文，1995 年。

[89] Jos Stam.随机动力学：模拟湍流对柔性结构的影响。*计算机图形论坛》*、

16(3), 1997.

[90] Ryusuke Sugimoto, Mingming He, Jing Liao, and Pedro V Sander.从静态照片中模拟和渲染水。In *ACM Trans.Graphics (SIGGRAPH Asia)*, pages 1-9, 2022.

[91] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir、Daniel Cohen-Or, and Amit H Bermano.人类运动 扩散模型。*arXiv 预印本 arXiv:2209.14916*, 2022.

[92] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach、Raphael Marinier, Marcin Michalski, and Sylvain Gelly.实现准确的视频生成模型：新指标 & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[93] Vikram Voleti、Alexia Jolicoeur-Martineau 和 Christopher Pal.Mcvd：用于预测、生成和插值的屏蔽条件视频扩散。*神经信息处理系统*，2022 年。

[94] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba.生成场景动态视频。*神经信息处理系统》*，2016 年。

[95] Neal Wadhwa、Michael Rubinstein、Fre´do Durand 和 William T Freeman.基于相位的视频运动处理。*ACM Trans.Graphics (SIGGRAPH)*, 32(4):1-10, 2013.

[96] Jacob Walker、Carl Doersch、Abhinav Gupta 和 Martial Hebert。不确定的未来：使用变异自动编码器从静态图像进行预测。In *Proc.欧洲 计算机视觉会议（ECCV）*，2016 年。

[97] Jacob Walker、Abhinav Gupta 和 Martial Hebert。密集从静态图像进行光流预测。In *Proc.计算机视觉会议（ICCV）*，第 2443-2451 页，2015 年。

[98] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen、Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou.Videocomposer：*ArXiv preprint arXiv:2306.02018*, 2023.

[99] 王耀辉、杨迪、Francois Bremond 和 Antitza Dantcheva.潜在图像动画器：通过潜在空间导航学习动画图像。*ArXiv 预印本 arXiv:2203.09043*, 2022.

[100] Frederik Warburg、Ethan Weber、Matthew Tancik、Aleksander Holynski, and Angjoo Kanazawa.纳福克星：从随意捕捉的纳福中再移动幽灵般的人工制品。*arXiv 预印本 arXiv:2304.10532*, 2023。

[101] Daniel Watson、William Chan、Ricardo Martin-Brualla、Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi.新颖的视图合成与扩散模型。*ArXiv 预印本 arXiv:2210.04628*, 2022.

[102] 翁忠义、布莱恩-柯利斯和艾拉-凯梅尔马切尔-Shlizerman. 照片唤醒：单张照片的三维角色动画。In *Proc.Computer Vision and Pattern Recog- nition (CVPR)*, pages 5908-5917, 2019.

[103] Jamie Wynn 和 Daniyar Turmukhambetov.DiffusioN-eRF：利用去噪扩散模型对神经辐射场进行正则化。在 *Proc.计算机视觉与模式 Recognition (CVPR)*，2023 年。

[104] 薛天帆、吴佳俊、Katherine L Bouman 和威廉-T-弗里曼视觉动力学：通过分层交叉卷积网络的随机未来生成。*Trans.Pat- tern Analysis and Machine Intelligence*, 41(9):2236-2250, 2019.

[105] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan.Dragnuwa：通过整合文本、图像和 轨迹实现视频生成的细粒度控制。*arXiv preprint arXiv:2308.08089*, 2023.

[106] Sihyun Yu、Kihyuk Sohn、Subin Kim 和 Jinwoo Shin.投影潜空间中的视频概率扩散模型。在 *Proc.计算机视觉与模式识别 (CVPR)*，第 18456-18466 页，2023 年。

[107] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai、洪方舟、李慧荣、杨磊、刘紫薇。Re- modiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023.

[108] 张亚波、魏玉祥、蒋东升、肖鹏Zhang, Wangmeng Zuo, and Qi Tian.Controlvideo：免训练可控文本视频生成。*arXiv preprint arXiv:2305.13077*, 2023.

[109] 赵健、张辉。薄板样条运动模型图像动画的薄板样条运动模型。在 *Proc.计算机视觉与模式 Recognition (CVPR)*，第 3657-3666 页，2022 年。

[110] 赵胜宇、Jonathan Cui、盛一伦、董玥、肖Liang, Eric I Chang, and Yan Xu.通过共调制生成式对抗网络实现大规模图像合成。*(ICLR)*, 2021.

[111] 周大权、王为民、闫汉书、吕伟伟、朱一哲和贾石、朱一哲、冯嘉仕。Magicvideo：用潜在扩散模型高效生成视频。*arXiv preprint arXiv:2211.11018*, 2022.