西北工业大学

生成图像动态–论文翻译

原论文标题：Generative Image Dynamics

李明泽

教育实验学院

计算机科学与技术

2024 年 11 月

学号：2022302574

Northwestern Polytechnical University

郑奇李理查德 • 塔克诺亚 • 斯纳夫利亚历山大 • 霍林斯基
谷歌研究

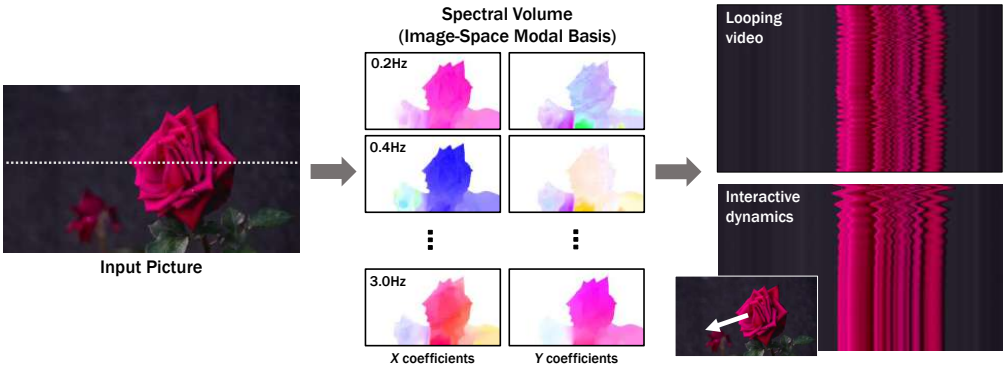

图 1. 我们对场景运动建模了生成图像空间先验: 从单个 RGB 图像出发，我们的方法生成一个光谱体积 [23]，这是一种在傅里叶域中建模密集、长期像素轨迹的运动表示。我们学习到的运动先验可以用于将单张图片转换为无缝循环的视频，或转换为响应用户输入 (如拖动和释放点) 的动态交互模拟。在右侧，我们将输出视频可视化为时空 X - t 切片 (沿左侧显示的输入扫描线)。

摘要

我们提出了一种对场景运动建模图像空间先验的方法。我们的先验是从一组提取自真实视频序列的运动轨迹中学习而来的，这些序列描绘了自然的、振荡的物体动态，如树木、花朵、蜡烛和在风中摇曳的衣物。我们将傅里叶域中的密集、长期运动建模为光谱体积，我们发现这非常适合与扩散模型进行预测。给定一张单独的图像，我们训练的模型使用频率协调的扩散采样过程来预测光谱体积，这可以转换为覆盖整个视频的运动纹理。结合基于图像的渲染模块，预测的运动表示可以用于多种下游应用，例如将静态图像转换为无缝循环的视频，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)。有关更多结果，请参见我们的项目页面:generative-dynamics.github.io。

1 引言

自然界始终处于运动之中，即使是看似静态的场景也因风、水流、呼吸或其他自然节奏而包含微妙的波动。模拟这种运动对于视觉内容合成至关重要——人类对运动的敏感性可能使得没有运动 (或运动略显不真实) 的图像显得怪异或不真实。

虽然人类很容易解读或想象场景中的运动，但训练一个模型来学习或生成真实的场景运动却远非易事。我们在世界中观察到的运动是场景潜在物理动态的结果，即施加在物体上的力，这些物体根据其独特的物理属性 (如质量、弹性等) 作出反应——这些量很难在大规模上进行测量和捕捉。幸运的是，在某些应用中，测量这些量并不是必要的: 例如，通过简单分析一些观察到的二维运动，可以在场景中模拟出合理的动态 [23]。

这种观察到的运动也可以作为学习场景动态的监督信号——因为尽管观察到的运动是多模态的，并且根植于复杂的物理效应，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。对于人类而言，这种可预测性深植于我们的感知系统中: 通过观看静态图像，我们可以想象出合理的运动——或者，由于可能存在许多这样的运动，可以想象出基于该图像的自然运动的分布。考虑到人类能够轻松建模这些分布，一个自然的研究问题就是如何在计算上对其进行建模。

最近在生成模型方面的进展，特别是条件扩散模型 [44, 85, 87]，使我们能够建模丰富的分布，包括基于文本的真实图像分布 [73, 74, 75]。这种能力催生了多个新应用，例如基于文本的多样且真实的图像内容生成。在这些图像模型成功的基础上，最近的研究将这些模型扩展到其他领域，如视频 [7, 43] 和 3D 几何 [77, 100, 101, 103]。

在本文中，我们为图像空间场景运动建模生成先验，即单幅图像中所有像素的运动。该模型是在从大量真实视频序列中自动提取的运动轨迹上训练的。具体而言，我们从每个训练视频中计算以光谱体积 [22, 23] 形式表示的运动，这是密集的、长范围像素轨迹的频域表示。光谱体积非常适合表现振荡动态的场景，例如在风中摇摆的树木和花朵。我们发现这种表示在作为扩散模型输出以建模场景运动时也非常有效。我们训练了一个生成模型，条件是单幅图像，可以从其学习的分布中采样光谱体积。然后，预测的光谱体积可以直接转化为运动纹理——一组长范围的、逐像素的运动轨迹——可用于为图像动画。光谱体积还可以被解释为用于模拟交互动态的图像空间模态基 [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，但通过共享注意模块在频率带之间协调这些预测。预测的运动可以用于合成未来帧 (通过基于图像的渲染模型)——将静态图像转变为真实的动画，如图 1 所

示。

与原始 RGB 像素的先验相比，运动捕捉的先验更为基础，具有更低维度的结构，能够有效解释像素值的长程变化。因此，生成中间运动可以导致更连贯的长期生成，并对动画实现更细致的控制。我们展示了我们训练模型在多个下游应用中的使用，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。最近在生成模型方面的进展使得基于文本提示的图像照片级合成成为可能 [16, 17, 24, 73, 74, 75]。这些文本到图像模型可以通过沿时间维度扩展生成的图像张量来增强，以合成视频序列 [7, 9, 43, 62, 84, 106, 106, 111]。虽然这些方法可以生成捕捉真实镜头时空统计特征的视频序列，但这些视频往往存在诸如运动不连贯、纹理的非现实时间变化以及违反物理约束 (如质量保持) 等伪影问题。

动画图像。与完全从文本生成视频不同，其他技术以静态图片为输入并对其进行动画处理。许多最近的深度学习方法采用 3D-Unet 架构直接生成视频体积 [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为条件)，并表现出与上述提到的相似伪影。克服这些限制的一种方法是，不直接生成视频内容，而是通过基于图像的渲染来动画化输入源图像，即根据来自外部源 (如驱动视频 [51, 80, 81, 82, 99])、运动或 3D 几何先验 [8, 29, 46, **?**, 67, 90, 97, 101, 102, 104, 109]，或用户注释 [6, 18, 20, 33, 38, 98, 105, 108]) 推导的运动来移动图像内容。根据运动场动画化图像可以产生更大的时间一致性和真实感，但这些先前的方法要么需要额外的指导信号或用户输入，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、振荡的三维运动 (例如，水波荡漾或树木在风中摇曳) 可以通过在傅里叶域中塑造的噪声进行建模，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]。这些谱技术被 Chuang 等人 [20] 适应用于从单个二维图像中动画植物、水和云，前提是有用户注释。我们的工作特别受到 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交互动态。我们采用了 Davis 等人的频率空间谱体积运动表示法，从大量训练视频中提取该表示，并展示了谱体积适合使用扩散模型从单幅图像预测运动。

其他方法在预测任务中使用了各种运动表示，其中图像或视频用于提供确定性的未来运动估计 [34, 71]，或更丰富的可能运动分布 [94, 96, 104]。然而，这些方法中的许多预测的是光流运动估计 (即每个像素的瞬时运动)，而不是完整的运动轨迹。此外，之前的许多工作集中在活动识别等任务上，而不是合成任务。最近的研究表明，在一些封闭域设置中，如人类和动物，使用生成模型建模和预测运动的

优势 [2, 19, 28, 72, 91, 107]。

视频作为纹理。某些动态场景可以被视为一种称为动态纹理的纹理 [26] - 它将视频建模为随机过程的时空样本。动态纹理可以表示平滑的自然运动，如波浪、火焰或摇动的树木，并已广泛用于视频分类、分割或编码 [12, 13, 14, 15, 76]。一种相关的纹理，称为视频纹理，将动态场景表示为一组输入视频帧以及任意一对帧之间的转移概率 [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，旨在生成无缝循环或无限变化的输出视频 [1, 21, 32, 58, 59, 78]。与大多数相关工作不同，我们的方法提前学习先验知识，然后可以应用于单幅图像。

3 概述

给定一幅单一图像 I_0 ，我们的目标是生成一段视频 $\left\{\widehat{I}_1, \widehat{I}_2 \cdot, \ldots, \widehat{I}_T\right\}$ ，展现如树木、花朵或在微风中摇曳的蜡烛火焰等振荡运动。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。我们的流程首先使用潜在扩散模型 (LDM) 为输入 I_0 预测一个光谱体积 $\mathcal{S}=\left(S_{f_0}, S_{f_1}, \ldots, S_{f_{K-1}}\right)$ 。然后，通过逆离散傅里叶变换将预测的光谱体积转换为运动纹理 $\mathcal{F}=\left(F_1, F_2, \ldots, F_T\right)$ 。这个运动决定了每个输入像素在未来每个时间步的位置。

给定预测的运动纹理，我们使用神经基于图像的渲染技术 (第 5 节) 来动画化输入的 RGB 图像。我们在第 6 节探讨了该方法的应用，包括生成无缝循环动画和模拟交互动态。

4 运动预测

4.1 运动表示

正式地说，运动纹理是时间变化的二维位移图的序列 $\mathcal{F}=\left\{F_t \mid t=1, \ldots, T\right\}$ ，其中输入图像 I_0 中每个像素坐标 $\mathbf{p}$ 的位移向量 $F_t(\mathbf{p})$ 定义了该像素在未来时间 t 的位置 $2D$ [20]。要在时间 t 生成未来帧，可以使用相应的位移图 D_t 从 I_0 中投影像素，从而得到一个前向扭曲的图像 I_t' :

$$I_t'\left(\mathbf{p}+F_t(\mathbf{p})\right)=I_0(\mathbf{p}) . \tag{1}$$

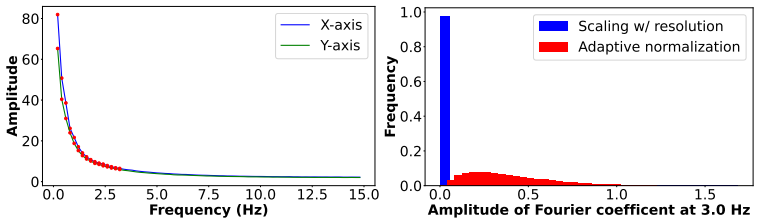

图 2. 左侧: 我们可视化了从真实视频中提取的 X 和 Y 运动分量的平均功率谱，显示为蓝色和绿色曲线。自然振荡运动主要由低频分量组成，因此我们使用前 $K = 16$ 项，用红点标记。右侧: 我们展示了在 3.0 Hz 处傅里叶项幅度的直方图，经过 (1) 按图像宽度和高度缩放幅度 (蓝色)，或 (2) 频率自适应归一化 (红色)。我们的自适应归一化防止系数集中在极端值。

如果我们的目标是通过运动纹理生成视频，那么一个选择是直接从输入图像预测时间域运动纹理。然而，运动纹理的大小需要与视频的长度成比例: 生成 T 输出帧意味着预测 T 位移场。为了避免为长视频预测如此大的输出表示，许多先前的动画方法要么自回归地生成视频帧 [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，这两种策略都无法确保生成视频的长期时间一致性。

幸运的是，许多自然运动可以被描述为少数谐振子的叠加，这些谐振子具有不同的频率、幅度和相位 [20, 23, 25, 50, 69]。由于这些基础运动是准周期性的，因此在频域中对其建模是自然的。因此，我们采用 Davis 等人 [23] 提出的高效频域运动表示，称为光谱体积，如图3 所示。光谱体积是从视频中提取的每个像素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光谱体积。我们采用潜在扩散模型 (LDMs) 来生成由 4 K 通道的 2D 运动谱图组成的光谱体积，其中 $K << T$ 是建模的频率数量，并且在每个频率下，我们需要四个标量来表示 x 和 y 维度的复傅里叶系数。请注意，未来时间步 $\mathcal{F}(\mathbf{p}) = \{F_t(\mathbf{p}) \mid t = 1, 2, \ldots T\}$ 的像素运动轨迹及其作为光谱体积的表示 $\mathcal{S}(\mathbf{p}) = \{S_{f_k}(\mathbf{p}) \mid k = \quad 0, 1, \cdot \frac{T}{2} - 1\}$ 通过快速傅里叶变换 (FFT) 相关联:

$$\mathcal{S}(\mathbf{p}) = FFT(\mathcal{F}(\mathbf{p})). \tag{2}$$

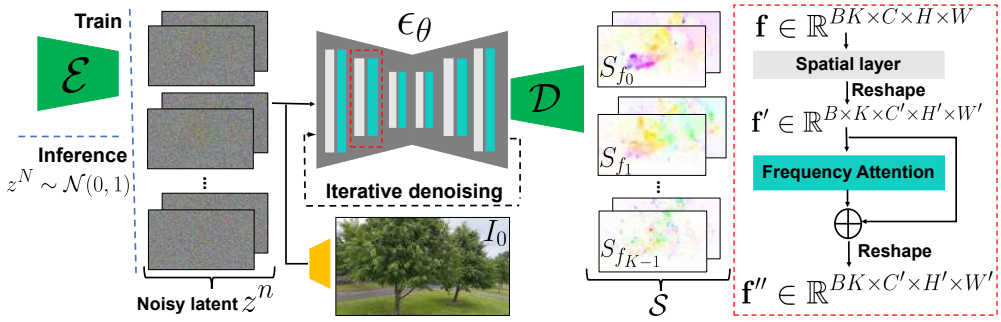

图 3. 运动预测模块。我们通过频率协调去噪模型预测光谱体积 $\mathcal{S}$ 。扩散网络的每个块 ϵ_θ 将 2D 空间层与注意力层 (红框，右侧) 交错，并迭代去噪潜在特征 z^n 。去噪特征被输入到解码器 $\mathcal{D}$ 中以生成 $\mathcal{S}$ 。在训练期间，我们将下采样的输入 I_0 与通过编码器 $\mathcal{E}$ 从真实运动纹理编码的噪声潜在特征连接，并在推理过程中用高斯噪声 z^N 替换噪声特征 (左侧)。

我们应该如何选择 K 输出频率？以往的实时动画研究观察到，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 1,000 个随机抽样的 5 秒真实视频片段中提取的运动的平均功率谱。如图2左侧所示，运动的功率谱随着频率的增加而呈指数下降。这表明大多数自然振荡运动确实可以通过低频项很好地表示。在实践中，我们发现前 $K = 16$ 个傅里叶系数足以在一系列真实视频和场景中真实地重现原始自然运动。

4.2 使用扩散模型预测运动

我们选择潜在扩散模型 (LDM)[74] 作为我们的运动预测模块的基础，因为 LDM 在计算效率上优于像素空间扩散模型，同时保持合成质量。标准的 LDM 由两个主要模块组成:(1) 一个变分自编码器 (VAE)，通过编码器将输入图像压缩到潜在空间 $z = E(I)$ ，然后通过解码器从潜在特征重建输入 $I = D(z)$；(2) 一个基于 U-Net 的扩散模型，学习从高斯噪声开始迭代去噪特征。我们的训练过程并不是针对 RGB 图像，而是针对来自真实视频序列的光谱体积，这些光谱体积被编码后经过预定义的方差调度进行 n 步的扩散，以产生噪声潜变量 z^n 。2DU-Net 被训练以通过迭代估计用于在每一步 $n \in (1, 2, \ldots, N)$ 更新潜在特征的噪声 $\epsilon_\theta\left(z^n; n, c\right)$ 来去噪这些噪声潜变量。LDM 的训练损失写为

$$\mathcal{L}_{LDM} = \mathbb{E}_{n\in\mathcal{U}[1,N],\epsilon^n\in\mathcal{N}(0,1)}\left[\left\|\epsilon^n - \epsilon_\theta\left(z^n; n, c\right)\right\|^2\right] \tag{3}$$

在这里，c 是任何条件信号的嵌入，例如文本，或者在我们的案例中，是训练视频序列的第一帧，I_0 。然后，干净的潜在特征 z^0 被传递通过解码器以恢复光谱体积。

频率自适应归一化。我们观察到的一个问题是，运动纹理在频率上具有特定的分布特征。如图2左侧的图所示，光谱体积的幅度范围从 0 到 100，并且随着频率的增加大约呈指数衰减。由于扩散模型要求输出的绝对值在-1 和 1 之间，以实现稳定的训练和去噪 [44]，我们必须在使用从真实视频中提取的 $\mathcal{S}$ 系数进行训练之前对其进行归一化。如果我们根据图像尺寸将这些系数的幅度缩放到 $[0,1]$ ，如先前的工作 [29, 77] 所示，几乎所有高频率的系数最终都会接近零，如图2右侧的图所示。基于这样的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，即使是小的预测误差也会在去归一化后导致较大的相对误差。

为了解决这个问题，我们采用了一种简单但有效的频率自适应归一化方法: 首先，我们根据从训练集计算的统计数据独立归一化每个频率的傅里叶系数。即，对于每个单独的频率 f_j ，我们计算所有输入样本中傅里叶系数幅度的 95^{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，我们对每个缩放后的傅里叶系数应用幂变换，以将其从极端值中拉开。在实践中，我们观察到平方根的表现优于其他非线性变换，例如对数或倒数。总之，光谱体积 $\mathcal{S}(\mathbf{p})$ 在频率 f_j (用于训练我们的 LDM) 的最终系数值计算为

$$S'_{f_j}(\mathbf{p}) = sign\left(S_{f_j}\right)\sqrt{\left|\frac{S_{f_j}(\mathbf{p})}{s_{f_j}}\right|}. \tag{4}$$

如图2右侧所示，应用频率自适应归一化后，光谱体积系数分布更加均匀。

频率协调去噪。预测具有 K 频带的光谱体积 $\mathcal{S}$ 的直接方法是从单个扩散 U-Net 输出一个 $4K$ 通道的张量。然而，正如之前的研究 [7] 所观察到的，训练一个模型以产生大量通道可能会导致过度平滑和不准确的输出。另一种选择是通过将额外的频率嵌入注入到 LDM[4] 中，独立预测每个单独的频率切片，但这种设计选择将导致频域中的无关预测，从而导致不现实的运动。

因此，受到最近视频扩散工作的启发 [7]，我们提出了一种频率协调去噪策略，如图3 所示。具体而言，给定输入图像 I_0 ，我们首先训练一个 LDM ϵ_θ 来预测一个单一的 4 通道频率切片的光谱体积 S_{f_j} ，在此过程中，我们将额外的频率嵌入与时间步嵌入一起注入到 LDM 中。然后，我们冻结该 LDM ϵ_θ 的参数，引入与 $2D$ 空间层交错的注意力层，并进行微调。具体来说，对于批量大小 B ，LDM ϵ_θ 的 $2D$ 空间层将通道大小为 C 的相应 $B\cdot K$ 噪声潜在特征视为形状为 $\mathcal{R}^{(B\cdot K)\times C\times H\times W}$ 的独立样本。然后，注意力层将这些特征解释为跨越频率轴的连续特征，我们将来自前面 $2D$ 空间层的潜在特征重塑为 $\mathcal{R}^{B\times K\times C\times H\times W}$ ，然后再输入到注意力层中。换句话说，频率注意力层经过微调，以协调所有频率切片，从而生成一致的光谱体积。在我们的实验中，我们观察到，当我们从单一的 2D U-Net 切换到频率协调去噪模块时，平均 VAE 重建误差从 0.024 改善到 0.018，这表明 LDM 预测精度的上限得到了改善；在第 7.3 节中，我们还展示了这一设计选择提高了视频生成质量。

5 基于图像的渲染

我们现在描述如何利用为给定输入图像 I_0 预测的光谱体积 $\mathcal{S}$ 渲染未来帧 $\widehat{I}_t$ 在时间 t 。我们首先使用应用于每个像素的逆时间 FFT 在时间域中推导运动纹理 $\mathcal{F}(\mathbf{p}) = FFT^{-1}(\mathcal{S}(\mathbf{p}))$ 。为了生成未来帧 $\widehat{I}_t$ ，我们采用基于深度图像的渲染技术，并使用预测的运动场 F_t 进行点云映射，以前向扭曲编码的 I_0 ，如图 4 所示。由于前向扭曲可能导致孔洞，并且多个源像素可能映射到相同的输出 2D 位置，我们采用了先前关于帧插值工作的特征金字塔 softmax 点云映射策略 [68]。

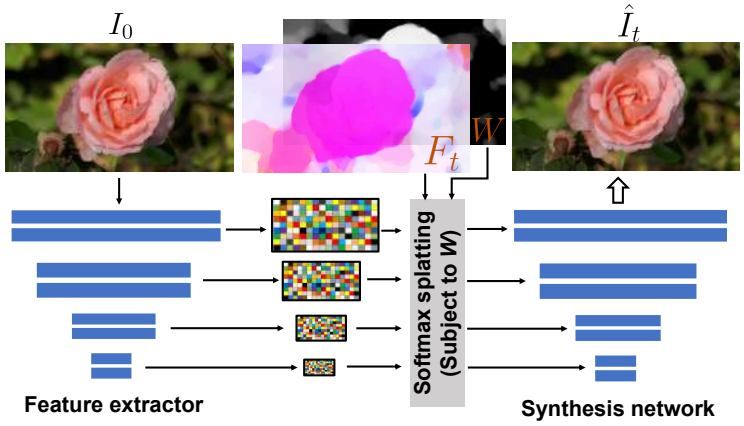

图 4. 渲染模块。我们使用深度基于图像的渲染模块填补缺失内容并细化扭曲的输入图像，其中从输入图像 I_0 中提取多尺度特征。然后，对具有从时间 0 到 t 的运动场 F_t 的特征应用 softmax 点云映射 (受权重 W 的影响)。扭曲的特征被输入到图像合成网络中，以生成渲染图像 $\widehat{I}_t$ 。

具体来说，我们通过特征提取网络编码 I_0 以生成多尺度特征图。对于每个尺度 j 的单独特征图，我们根据分辨率调整和缩放预测的二维运动场 F_t 。如 Davis 等人所述 [22]，我们使用预测的流动幅度作为深度的代理，以确定映射到其目标位置的每个源像素的贡献权重。特别地，我们计算每个像素的权重 $W(\mathbf{p}) = \frac{1}{T}\sum_t \left\| F_t(\mathbf{p}) \right\|_2$ ，作为预测运动纹理的平均幅度。换句话说，我们假设大幅度运动对应于移动的前景物体，而小幅度或零运动对应于背景。我们使用运动衍生的权重，而不是可学习的权重 [46]，因为我们观察到在单视图情况下，可学习的权重对于解决遮挡模糊并不有效。

利用运动场 F_t 和权重 W，我们应用软最大化溅射将每个尺度的特征图进行扭曲，以生成扭曲特征。然后将扭曲特征注入图像合成解码器的相应块中，以生成最终渲染的图像 $\widehat{I}_t$ 。

我们联合训练特征提取器和合成网络，使用从真实视频中随机采样的起始帧和目标帧 (I_0, I_t) ，利用从 I_0 到 I_t 的估计流场来扭曲来自 I_0 的编码特征，并使用

VGG 感知损失 [49] 监督预测 $\widehat{I}_t$ 与 I_t 之间的关系。

6 应用

图像到视频。我们的系统通过首先从输入图像预测运动光谱体积，然后应用我们的基于图像的渲染模块到从光谱体积转换而来的运动纹理，来实现单张静态图片的动画。由于我们明确建模场景运动，这使我们能够通过线性插值运动纹理来生成慢动作视频，或通过调整预测光谱体积系数的幅度来放大 (或缩小) 动画运动。

无缝循环。许多应用需要无缝循环的视频，其中视频的开始和结束之间没有不连续性。不幸的是，很难找到大量无缝循环视频用于训练。相反，我们设计了一种方法，利用我们的运动扩散模型，该模型在常规非循环视频剪辑上进行训练，以生成无缝循环视频。受到最近关于图像编辑指导的工作的启发，我们的方法是一种运动自我指导技术，利用明确的循环约束来指导运动去噪采样处理。特别是在推理过程中每次迭代的去噪步骤中，我们在标准的无分类器指导 [45] 旁边加入一个额外的运动指导信号，我们强制每个像素在开始和结束帧的位置和速度尽可能相似:

$$\widehat{\epsilon}^n = (1+w)\epsilon_\theta\left(z^n; n, c\right) - w\epsilon_\theta\left(z^n; n, \varnothing\right) + u\sigma^n\nabla_{z^n}\mathcal{L}_g^n \tag{5}$$

$$\mathcal{L}_g^n = \left\|F_T^n - F_1^n\right\|_1 + \left\|\nabla F_T^n - \nabla F_1^n\right\|_1 \tag{6}$$

其中 F_t^n 是时间 t 和去噪步骤 $n.w$ 的预测二维位移场，去噪步骤的 $n.w$ 是无分类器指导权重，而 u 是运动自我指导权重。在补充视频中，我们应用了一个基于外观的循环算法 [58] 来从我们的非循环输出生成循环视频，并展示我们的运动自我指导技术能够生成失真更小、伪影更少的无缝循环视频。

从单幅图像中获得的交互动态。Davis 等人 [22] 表明，在某些共振频率下评估的谱体积可以近似为图像空间模态基，这是一种对基础场景振动模式的投影 (或更一般地，捕捉振荡动态中的空间和时间相关性)，并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]，使我们能够将物体的物理响应的图像空间二维运动位移场写成运动谱系数的加权和 S_{f_j} ，并在每个模拟时间步 t 中通过复杂模态坐标的状态 $\mathbf{q}_{f_j}(t)$ 进行调制。

$$F_t\left(\mathbf{p}\right) = \sum_{f_j} S_{f_j}\left(\mathbf{p}\right)\mathbf{q}_{f_j}(t) \tag{7}$$

我们通过对在模态空间 [22, 23, 70] 中表示的解耦质量-弹簧-阻尼系统的运动方程应用显式欧拉方法来模拟模态坐标的状态 $\mathbf{q}_{f_j}(t)$。我们建议读者参考补充材料和原始工作以获取完整的推导。请注意，我们的方法从单幅图像生成交互场景，而这些先前的方法需要视频作为输入。

表 1. 测试集上的定量比较。我们报告图像合成和视频合成的质量。这里，KID 被缩放为 100。所有误差越低越好。有关基线和误差指标的描述，请参见第 7.1 节。

方法	图像合成		视频合成			
	FID	KID	FVD	FVD32	DTFVD	DTFVD32
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
Endo 等 [29]	10.4	0.19	166.0	231.6	5.35	65.1
Holynski 等 [46]	11.2	0.20	179.0	253.7	7.23	46.8
我们的方法	4.03	0.08	47.1	62.9	2.53	6.75

7 实验

实施细节。我们使用 LDM [74] 作为预测光谱体积的基础，对于此，我们使用一个维度为 4 的连续潜在空间的 VAE。我们使用 L_1 重建损失、多尺度梯度一致性损失 [?] 和 KL 散度损失，分别赋予权重 $1, 0.2, 10^{-6}$ 来训练 VAE。我们训练与原始 LDM 工作中使用的相同 $2D$ U-Net，以简单的 MSE 损失 [44] 进行迭代去噪，并采用 [41] 中的注意力层进行频率协调去噪。为了进行定量评估，我们从头开始在大小为 256×160 的图像上训练 VAE 和 LDM，以便进行公平比较，使用 16 个 Nvidia A100 GPU 大约需要 6 天才能收敛。对于我们的主要定量和定性结果，我们使用 DDIM [86] 运行运动扩散模型 250 步。我们还展示了生成的视频，分辨率高达 512×288 ，这些视频是通过在我们的数据集上微调预训练的图像修复 LDM 模型 [74] 创建的。

我们在 IBR 模块中采用 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修复的架构 [57, 110]。我们的渲染模块在推理过程中在 Nvidia V100 GPU 上以 25FPS 的实时速度运行。我们采用通用引导 [3] 来生成无缝循环视频，其中我们设置权重 $w = 1.75, u = 200$ ，并使用 500 个 DDIM 步骤和 2 次自我递归迭代。

数据。我们收集并处理了一组来自在线来源和我们自己捕获的 3,015 个自然场景视频，这些视频展示了振荡运动。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应用粗到细的光流方法 [10, 61]。作为训练数据，我们将每第 10 帧视频作为输

入图像，并使用计算出的运动轨迹推导出相应的真实光谱体积，涵盖接下来的 149 帧。总体而言，我们的数据包含超过 150 K 图像-运动对。

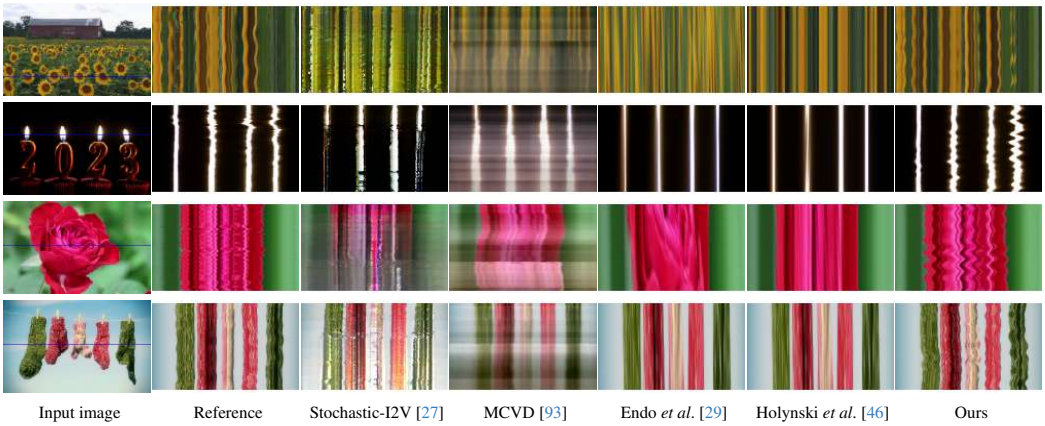

图 5. X - t 由不同方法生成的视频切片。从左到右: 输入图像及其对应的 X - t 来自真实视频的视频切片，来自三种基线生成的视频 [27, 29, 46, 93]，最后是我们的方法生成的视频。

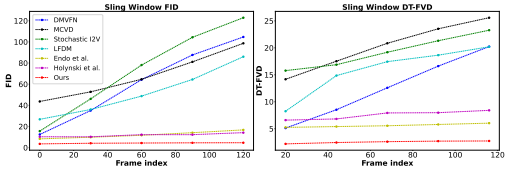

图 6. 滑动窗口 FID 和 DTFVD。我们展示了窗口大小为 30 帧的滑动窗口 FID，以及窗口大小为 16 帧的 DTFVD，用于由不同方法生成的视频。

基线。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比较。Endo 等人 [29] 和 DMVFN [48] 预测瞬时 2D 运动场并自回归地渲染未来帧。Holynski 等人 [46] 则通过单一静态欧拉运动描述来模拟运动。其他近期工作如随机图像到视频 (Stochastic-I2V) [27]、TATS [35] 和 MCVD [93] 采用变分自编码器 (VAEs)、变换器或扩散模型直接预测原始视频帧；LFDM [67] 通过预测流量体积并在扩散模型中扭曲潜变量来生成未来帧。我们使用各自的开源实现对上述所有方法在我们的数据上进行训练。[1]

我们通过两种方式评估我们的方法和先前基线生成的视频质量。首先，我们使用针对图像合成任务设计的指标评估单个合成帧的质量。我们采用 Fréchet Inception Distance (FID) [42] 和 Kernel Inception Distance (KID) [5] 来测量生成帧与真实帧分布之间的平均距离。

其次，为了评估合成视频的质量和时间一致性，我们采用 Fréchet 视频距离 [92]，窗口大小为 16 (FVD) 和 32 (FVD $_{32}$)，基于在 Human Kinetics 数据集

[52] 上训练的 I3D 模型 [11]。为了更真实地反映我们希望生成的自然振荡运动的合成质量，我们还采用动态纹理 Fréchet 视频距离 [27]，该距离使用窗口大小为 16 (DTFVD) 和 32 (DTFVD $_{32}$) 的视频，使用在动态纹理数据库 [37] 上训练的 I3D 模型，该数据集主要由自然运动纹理组成。

表 2. 消融研究。第 7.3 节描述了每个配置。

方法	图像合成		视频合成			
	FID	KID	FVD	FVD32	DTFVD	
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
不带自适应规范。	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。	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
完整 ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

我们进一步使用窗口大小为 30 帧的滑动窗口 FID，以及窗口大小为 16 帧的滑动窗口 DTFVD，如 [57, 60] 所示，以测量生成的视频质量随时间的退化情况。对于所有方法，我们通过中心裁剪在 256×128 分辨率下评估指标。

图 7. 我们展示了来自三个最近的大型视频扩散模型生成的未来帧 [31, 36, 98]。

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定量比较。我们的方法在图像和视频合成质量方面显著优于之前的单图像动画基线。具体而言，我们更低的 FVD 和 DT-FVD 距离表明，我们的方法生成的视频更具现实感且时间一致性更强。此外，图6显示了来自不同方法生成的视频的滑动窗口 FID 和滑动窗口 DT-FVD 距离。得益于全局光谱体积表示，我们的方法生成的视频不会随着时间的推移而退化。

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时空 $X-t$ 切片，这是一种可视化视频中小运动的标准方式 [95]。如图5所示，我们生成的视频动态与对应真实参考视频 (第二列) 中观察到的运动模式更为相似，相较于其他方法。基线方法如随机 I2V [27] 和 MCVD [93] 未能在时间上真实地建模外观和运动。Endo 等 [29] 和 Holynski 等 [46] 生成的视频帧具有较少的伪影，但在时间上表现出过于平滑或非振荡的运动。我们建议读者参考补充材料，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，以验证我们运动预测和渲染模块中的主要设计选择，比较我们的完整配置与不同变体。具体而言，我们使用不同数量的频带 $K = 4, 8, 16, 24$ 评估结果。我们观察到，增加频带数量可以提高视频预测质量，但在超过 16 个频率时，改善效果有限。接下来，我们从真实光谱体积中去除自适应频率归一化，而仅根据输入图像的宽度和高度进行缩放 (w/o 自适应归一化)。此外，我们去除了频率协调去噪模块 (独立预测)，或用一个更简单的 DM 替代，其中一个张量体积的 4 K 通道光谱体积通过单个 2DU-net 扩散模型共同预测 (体积预测)。最后，我们比较使用基线渲染方法的结果，该方法对单尺度特征应用 softmax splatting，并使用 [46] 所用的可学习权重 (基线 splat)。我们还增加了一个基线，其中生成的视频是通过重复输入图像 N 次而形成的 (重复 I_0)。从表2中，我们观察到所有更简单或替代的配置与我们的完整模型相比，性能更差。

图 8. 限制。我们展示了渲染未来帧的示例 (偶数)，以及输入图像和渲染图像的叠加 (奇数)。我们的方法在薄物体或大运动的区域，以及需要填充大量新内容的区域可能会产生伪影。

7.4 与大型视频模型的比较

我们进一步进行了一项用户研究，并将我们生成的动画与最近的大型视频扩散模型进行比较:AnimateDiff [36]、ModelScope [98] 和 Gen-2[31]，这些模型直接预测视频体积。在从测试集中随机选择的 30 个视频中，我们询问用户 “哪个视频更逼真？”。用户报告了对我们方法的 80.9% 偏好。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遵循输入图像内容，要么随着时间的推移表现出逐渐

的颜色漂移和失真。我们建议读者参考补充材料以获取完整的比较。

8 讨论与结论

限制。由于我们的方法仅预测光谱体积的低频部分，因此可能无法建模非振荡运动或高频振动——这可以通过使用学习的运动基来解决。此外，生成视频的质量依赖于基础运动轨迹的质量，这可能在场景中出现细小移动物体或位移较大的物体时下降。即使是正确的，要求生成大量新未见内容的运动也可能导致质量下降(图8)。

结论。我们提出了一种新的方法，用于从单张静态图片建模自然振荡动态。我们的图像空间运动先验通过频谱体积表示，频谱体积是每个像素运动轨迹的频率表示，我们发现这种方法在使用扩散模型进行预测时既高效又有效，并且我们从现实世界视频的集合中学习得来。频谱体积是使用频率协调的潜在扩散模型进行预测的，并通过基于图像的渲染模块用于动画未来的视频帧。我们展示了我们的方法能够从单张图片生成照片级真实感的动画，并显著优于之前的基准，并且它可以启用多个下游应用，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 的富有成效的讨论和有益的评论。

参考文献

[1] Aseem Agarwala, Ke Colin Zheng, Chris Pal, Maneesh Agrawala, Michael Cohen, Brian Curless, David Salesin, and Richard Szeliski. Panoramic video textures. In *ACM* Trans. Graphics (SIGGRAPH), pages 821-827. 2005.

[2] Hyemin Ahn, Esteve Valls Mascaro, and Dongheui Lee. Can we use diffusion probabilistic models for 3 d 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. Universal guidance for diffusion models. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 843-852, 2023.

[4] Hugo Bertiche, Niloy J Mitra, Kuldeep Kulkarni, Chun-Hao P Huang, Tuanfeng Y Wang, Meysam Madadi, Sergio Escalera, and Duygu Ceylan. Blowing

in the wind: Cy-clenet for human cinemagraphs from still images. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 459-468, 2023.

[5] Mikofaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. arXiv preprint arXiv:1801.01401, 2018.

[6] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In Proc. Int. Conf. on Computer Vision (ICCV), pages 14707-14717, 2021.

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dock-horn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 22563-22575, 2023.

[8] Richard Strong Bowen, Richard Tucker, Ramin Zabih, and Noah Snavely. Dimensions of motion: Monocular prediction through flow subspaces. In International Conference on $3D$ Vision (3DV), pages 454-464, 2022.

[9] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. Neural Information Processing Systems, 35:31769-31781, 2022.

[10] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In Proc. European Conf. on Computer Vision (ECCV), pages 25-36, 2004.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 6299-6308, 2017.

[12] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4 d video textures for interactive character appearance. In Computer Graphics Forum, volume 33, pages 371-380. Wiley Online Library, 2014.

[13] Antoni B Chan and Nuno Vasconcelos. Mixtures of dynamic textures. In Proc. Int. Conf. on Computer Vision (ICCV),

pages 641-647, 2005.

[14] Antoni B Chan and Nuno Vasconcelos. Classifying video with kernel dynamic textures. In Proc. Computer Vision and Pattern Recognition (CVPR), 2007.

[15] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. Trans. Pattern Analysis and Machine Intelligence, 30(5):909- 926, 2008.

[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: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704, 2023.

[17] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 11315-11325, 2022.

[18] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. arXiv preprint arXiv:2304.14404, 2023.

[19] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 18000-18010, 2023.

[20] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H Salesin, and Richard Szeliski. Animating pictures with stochastic motion textures. In ACM Trans. Graphics (SIGGRAPH), pages 853-860, 2005.

[21] Vincent C Couture, Michael S Langer, and Sebastien Roy. Omnistereo video textures without ghosting. In International Conference on 3D Vision, pages 64-70. IEEE, 2013.

[22] Abe Davis, Justin G Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. ACM Trans. Graphics (SIGGRAPH), 34(6):1–7, 2015.

[23] Myers Abraham Davis. Visual vibration analysis. PhD thesis, Massachusetts Institute of Technology, 2016.

[24] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Neural Information Processing Systems, 34:8780-8794, 2021.

[25] Julien Diener, Mathieu Rodriguez, Lionel Baboud, and Lionel Reveret. Wind projection basis for real-time animation of trees. In Computer graphics forum, volume 28, pages 533-540. Wiley Online Library, 2009.

[26] Gianfranco Doretto, Alessandro Chiuso, Ying Nian Wu, and Stefano Soatto. Dynamic textures. 51:91-109, 2003.

[27] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Om-mer. Stochastic image-to-video synthesis using cinns. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 3742-3753, June 2021.

[28] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 481-490, 2023.

[29] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: Self-supervised learning of decoupled motion and appearance for single-image video synthe-

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. Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986, 2023.

[31] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In Proc. Int. Conf. on Computer Vision (ICCV), pages 7346-7356, 2023.

[32] Matthew Flagg, Atsushi Nakazawa, Qiushuang Zhang, Sing Bing Kang, Young Kee Ryu, Irfan Essa, and James M Rehg. Human video textures. In Proceedings of the 2009 symposium on Interactive 3D graphics and games, pages 199-206, 2009.

[33] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In International Conference on Machine Learning, pages 3233-3246. PMLR, 2020.

[34] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2Flow: Motion hallucination from static images for action recognition. In Proc. Computer Vision and Pattern Recognition (CVPR), 2018.

[35] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. arXiv preprint arXiv:2204.03638, 2022.

[36] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

[37] Isma Hadji and Richard P Wildes. A new large scale dynamic texture dataset with application to convnet understanding. In Proc. European Conf. on Computer Vision (ECCV), pages 320-335, 2018.

[38] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 7854-7863, 2018.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 770- 778, 2016.

[40] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. arXiv preprint arXiv:2307.06940, 2023.

[41] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221, 2022.

[42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Neural Information Processing Systems, 30, 2017.

[43] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang,

Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022.

[44] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Neural Information Processing Systems, 33:6840-6851, 2020.

[45] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

[46] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with Eulerian motion fields. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 5810-5819, 2021.

[47] Tobias Hoppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. Trans. Mach. Learn. Res., 2022, 2022.

[48] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. ArXiv, abs/2303.09875, 2023.

[49] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proc. European Conf. on Computer Vision (ECCV), pages 694-711, 2016.

[50] Hitoshi Kanda and Jun Ohya. Efficient, realistic method for animating dynamic behaviors of 3 d botanical trees. In International Conference on Multimedia and Expo, volume 2, pages II-89. IEEE, 2003.

[51] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv preprint 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 human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[53] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523, 2018.

[54] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching

frozen people. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 4521-4530, 2019.

[55] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 2041- 2050, 2018.

[56] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 4273-4284, 2023.

[57] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In Proc. European Conf. on Computer Vision (ECCV), pages 515- 534. Springer, 2022.

[58] Jing Liao, Mark Finch, and Hugues Hoppe. Fast computation of seamless video loops. ACM Trans. Graphics (SIG-

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

[59] Zicheng Liao, Neel Joshi, and Hugues Hoppe. Automated video looping with progressive dynamism. ACM Transactions on Graphics (TOG), 32(4):1-10, 2013.

[60] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Maka-dia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In Proc. Int. Conf. on Computer Vision (ICCV), pages 14458-14467, 2021.

[61] Ce Liu. Beyond pixels: exploring new representations and applications for motion analysis. PhD thesis, Massachusetts Institute of Technology, 2009.

[62] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 10209-10218, 2023.

[63] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In Proc. Computer Vision and Pattern Recognition (CVPR), 2022.

[64] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. Text-guided synthesis of eulerian cinemagraphs. 2023.

[65] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. Neural Information Processing Systems, 35:22438-22450, 2022.

[66] Medhini Narasimhan, Shiry Ginosar, Andrew Owens, Alexei A Efros, and Trevor Darrell. Strumming to the beat: Audio-conditioned contrastive video textures. In Proc. Winter Conference on Applications of Computer Vision, pages 3761-3770, 2022.

[67] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 18444-18455, 2023.

[68] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 5437-5446, 2020.

[69] Shin Ota, Machiko Tamura, Kunihiko Fujita, T Fujimoto, K Muraoka, and Norishige Chiba. 1/f/sup/spl beta//noise-based real-time animation of trees swaying in wind fields. In Proceedings Computer Graphics International, pages 52-59. IEEE, 2003.

[70] Automne Petitjean, Yohan Poirier-Ginter, Ayush Tewari, Guillaume Cordonnier, and George Drettakis. Modalnerf: Neural modal analysis and synthesis for free-viewpoint navigation in dynamically vibrating scenes. In Computer Graphics Forum, volume 42, 2023.

[71] Silvia L. Pintea, Jan C. van Gemert, and Arnold W. M. Smeulders. Déjà vu: Motion prediction in static images. In Proc. European Conf. on Computer Vision (ECCV), 2014.

[72] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. arXiv preprint arXiv:2302.05905, 2023.

[73] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proc. Computer Vision 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. Neural Information Processing Systems, 35:36479-36494, 2022.

[76] Payam Saisan, Gianfranco Doretto, Ying Nian Wu, and Stefano Soatto. Dynamic texture recognition. In Proc. Computer Vision and Pattern Recognition (CVPR), 2001.

[77] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation, 2023.

[78] Arno Schödl, Richard Szeliski, David H Salesin, and Irfan Essa. Video textures. In ACM Trans. Graphics (SIGGRAPH), pages 489-498, 2000.

[79] Mikio Shinya and Alain Fournier. Stochastic motion-motion under the influence of wind. Computer Graphics Forum, 11(3), 1992.

[80] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 2377-2386, 2019.

[81] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. Neural Information Processing Systems, 32, 2019.

[82] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 13653-13662, 2021.

[83] Chen Qian Kwan-Yee Lin Hongsheng Li Siming Fan, Jing-tan Piao. Simulating fluids in real-world still images. arXiv preprint, arXiv:2204.11335, 2022.

[84] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elho-seiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 3626-3636, 2022.

[85] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pages 2256-2265. PMLR, 2015.

[86] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denois-ing diffusion implicit models. arXiv:2010.02502, October 2020.

[87] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Ab-hishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

[88] Jos Stam. Multi-scale stochastic modelling of complex natural phenomena. PhD thesis, 1995.

[89] Jos Stam. Stochastic dynamics: Simulating the effects of turbulence on flexible structures. Computer Graphics Forum,

16(3), 1997.

[90] Ryusuke Sugimoto, Mingming He, Jing Liao, and Pedro V Sander. Water simulation and rendering from a still photograph. 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. Human motion diffusion model. arXiv preprint arXiv:2209.14916, 2022.

[92] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717, 2018.

[93] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In Neural Information Processing Systems, 2022.

[94] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In Neural Information Processing Systems, 2016.

[95] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. ACM Trans. Graphics (SIGGRAPH), 32(4):1–10, 2013.

[96] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In Proc. European Conf. on Computer Vision (ECCV), 2016.

[97] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In Proc. Int. Conf. on Computer Vision (ICCV), pages 2443-2451, 2015.

[98] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018, 2023.

[99] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043, 2022.

[100] Frederik Warburg, Ethan Weber, Matthew Tancik, Alek-sander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. arXiv preprint arXiv:2304.10532, 2023.

[101] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628, 2022.

[102] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 5908-5917, 2019.

[103] Jamie Wynn and Daniyar Turmukhambetov. DiffusioN-eRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In Proc. Computer Vision and Pattern Recognition (CVPR), 2023.

[104] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman. Visual dynamics: Stochastic future generation via layered cross convolutional networks. Trans. Pattern 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: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089, 2023.

[106] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 18456-18466, 2023.

[107] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Re-modiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116, 2023.

[108] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077, 2023.

[109] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In Proc. Computer Vision and Pattern Recognition (CVPR), pages 3657-3666, 2022.

[110] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In International Conference on Learning Representations (ICLR), 2021.

[111] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022.

Generative Image Dynamics

Zhengqi Li Richard Tucker Noah Snavely Aleksander Holynski
Google Research

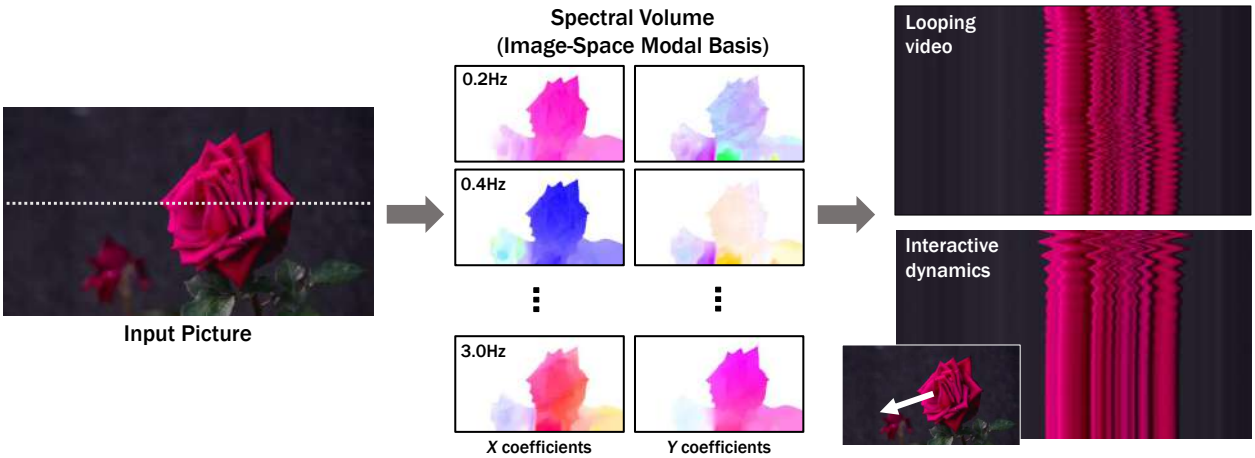

Figure 1. We model a generative image-space prior on scene motion: from a single RGB image, our method generates a *spectral volume* [23], a motion representation that models dense, long-term pixel trajectories in the Fourier domain. Our learned motion priors can be used to turn a single picture into a seamlessly looping video, or into an interactive simulation of dynamics that responds to user inputs like dragging and releasing points. On the right, we visualize output videos as space-time X-t slices (along the input scanline shown on the left).

Abstract

We present an approach to modeling an image-space prior on scene motion. Our prior is learned from a collection of motion trajectories extracted from real video sequences depicting natural, oscillatory dynamics of objects such as trees, flowers, candles, and clothes swaying in the wind. We model dense, long-term motion in the Fourier domain as spectral volumes*, which we find are well-suited to prediction with diffusion models. Given a single image, our trained model uses a frequency-coordinated diffusion sampling process to predict a spectral volume, which can be converted into a motion texture that spans an entire video. Along with an image-based rendering module, the predicted motion representation can be used for a number of downstream applications, such as turning still images into seamlessly looping videos, or allowing users to interact with objects in real images, producing realistic simulated dynamics (by interpreting the spectral volumes as image-space modal bases). See our project page for more results: generative-dynamics.github.io.*

1. Introduction

The natural world is always in motion, with even seemingly static scenes containing subtle oscillations as a result of wind, water currents, respiration, or other natural rhythms. Emulating this motion is crucial in visual content synthesis—human sensitivity to motion can cause imagery without motion (or with slightly unrealistic motion) to seem uncanny or unreal.

While it is easy for humans to interpret or imagine motion in scenes, training a model to learn or produce realistic scene motion is far from trivial. The motion we observe in the world is the result of a scene's underlying physical dynamics, i.e., forces applied to objects that respond according to their unique physical properties—their mass, elasticity, etc—quantities that are hard to measure and capture at scale. Fortunately, measuring them is not necessary for certain applications: *e.g.*, one can simulate plausible dynamics in a scene by simply analyzing some observed 2D motion [23].

This same observed motion can also serve as a supervisory signal in learning dynamics *across* scenes—because although observed motion is multi-modal and grounded in complex physical effects, it is nevertheless often predictable:

candles will flicker in certain ways, trees will sway, and their leaves will rustle. For humans, this predictability is ingrained in our systems of perception: by viewing a still image, we can imagine plausible motions— or, since there might have been many possible such motions, a *distribution* of natural motions conditioned on that image. Given the facility with which humans are able to model these distributions, a natural research problem is to model them computationally.

Recent advances in generative models, in particular conditional diffusion models [44, 85, 87], have enabled us to model rich distributions, including distributions of real images conditioned on text [73–75]. This capability has enabled several new applications, such as text-conditioned generation of diverse and realistic image content. Following the success of these image models, recent work has extended these models to other domains, such as videos [7, 43] and 3D geometry [77, 100, 101, 103].

In this paper, we model a generative prior for *image-space scene motion*, i.e., the motion of all pixels in a single image. This model is trained on motion trajectories automatically extracted from a large collection of real video sequences. In particular, from each training video we compute motion in the form of a *spectral volume* [22, 23], a frequency-domain representation of dense, long-range pixel trajectories. Spectral volumes are well-suited to scenes that exhibit oscillatory dynamics, e.g., trees and flowers moving in the wind. We find that this representation is also highly effective as an output of a diffusion model for modeling scene motion. We train a generative model that, conditioned on a single image, can sample spectral volumes from its learned distribution. A predicted spectral volume can then be directly transformed into a motion texture—a set of long-range, per-pixel motion trajectories—that can be used to animate the image. The spectral volume can also be interpreted as an *image-space modal basis* for use in simulating interactive dynamics [22].

We predict spectral volumes from input images using a diffusion model that generates coefficients one frequency at a time, but coordinates these predictions across frequency bands through a shared attention module. The predicted motions can be used to synthesize future frames (via an image-based rendering model)—turning still images into realistic animations, as illustrated in Fig. 1.

Compared with priors over raw RGB pixels, priors over motion capture more fundamental, lower-dimensional structure that efficiently explains long-range variations in pixel values. Hence, generating intermediate motion leads to more coherent long-term generation and more fine-grained control over animations. We demonstrate the use of our trained model in several downstream applications, such as creating seamless looping videos, editing the generated motions, and enabling interactive dynamic images via image-space modal bases, i.e., simulating the response of object dynamics to user-applied forces [22].

2. Related Work

Generative synthesis. Recent advances in generative models have enabled photorealistic synthesis of images conditioned on text prompts [16, 17, 24, 73–75]. These text-to-image models can be augmented to synthesize video sequences by extending the generated image tensors along a time dimension [7, 9, 43, 62, 84, 106, 106, 111]. While these methods can produce video sequences that capture the spatiotemporal statistics of real footage, these videos often suffer from artifacts like incoherent motion, unrealistic temporal variation in textures, and violations of physical constraints like preservation of mass.

Animating images. Instead of generating videos entirely from text, other techniques take as input a still picture and animate it. Many recent deep learning methods adopt a 3D-Unet architecture to produce video volumes directly [27, 36, 40, 47, 53, 93]. These models are effectively the same video generation models (but conditioned on image information instead of text), and exhibit similar artifacts to those mentioned above. One way to overcome these limitations is to not directly generate the video content itself, but instead animate an input source image through image-based rendering, i.e., moving the image content around according to motion derived from external sources such as a driving video [51, 80–82, 99], motion or 3D geometry priors [8, 29, 46, 63–65, 67, 90, 97, 101, 102, 104, 109], or user annotations [6, 18, 20, 33, 38, 98, 105, 108]. Animating images according to motion fields yields greater temporal coherence and realism, but these prior methods either require additional guidance signals or user input, or utilize limited motion representations.

Motion models and motion priors. In computer graphics, natural, oscillatory 3D motion (e.g., water rippling or trees waving in the wind) can be modeled with noise that is shaped in the Fourier domain and then converted to time-domain motion fields [79, 88]. Some of these methods rely on a modal analysis of the underlying dynamics of the system being simulated [22, 25, 89]. These spectral techniques were adapted to animate plants, water, and clouds from single 2D pictures by Chuang *et al*. [20], given user annotations. Our work is especially inspired by Davis [23], who connected modal analysis of a scene with the motions observed in a video of that scene, and used this analysis to simulate interactive dynamics from a video. We adopt the frequency-space *spectral volume* motion representation from Davis *et al*., extract this representation from a large set of training videos, and show that spectral volumes are suitable for predicting motion from single images with diffusion models.

Other methods have used various motion representations in *prediction* tasks, where an image or video is used to inform a deterministic future motion estimate [34, 71], or a more rich *distribution* of possible motions [94, 96, 104]. However,

many of these methods predict an optical flow motion estimate (i.e., the instantaneous motion of each pixel), not full motion trajectories. In addition, much of this prior work is focused on tasks like activity recognition, not on synthesis tasks. More recent work has demonstrated the advantages of modeling and predicting motion using generative models in a number of closed-domain settings such as humans and animals [2, 19, 28, 72, 91, 107].

Videos as textures. Certain moving scenes can be thought of as a kind of texture—termed *dynamic textures* [26]—that model videos as space-time samples of a stochastic process. Dynamic textures can represent smooth, natural motions like waves, flames, or moving trees, and have been widely used for video classification, segmentation or encoding [12–15, 76]. A related kind of texture, called a *video texture*, represents a moving scene as a set of input video frames along with transition probabilities between any pair of frames [66, 78]. A number of methods estimate dynamic or video textures through analysis of scene motion and pixel statistics, with the aim of generating seamlessly looping or infinitely varying output videos [1, 21, 32, 58, 59, 78]. In contrast to much of this work, our method learns priors in advance that can then be applied to single images.

3. Overview

Given a single picture I_0, our goal is to generate a video $\{\hat{I}_1, \hat{I}_2, ..., \hat{I}_T\}$ featuring oscillatory motions such as those of trees, flowers, or candle flames swaying in the breeze. Our system consists of two modules: a motion prediction module and an image-based rendering module. Our pipeline begins by using a latent diffusion model (LDM) to predict a spectral volume $\mathcal{S} = \left(S_{f_0}, S_{f_1}, ..., S_{f_{K-1}}\right)$ for the input I_0. The predicted spectral volume is then transformed to a motion texture $\mathcal{F} = (F_1, F_2, ..., F_T)$ through an inverse discrete Fourier transform. This motion determines the position of each input pixel at every future time step.

Given a predicted motion texture, we then animate the input RGB image using a neural image-based rendering technique (Sec. 5). We explore applications of this method, including producing seamless looping animations and simulating interactive dynamics, in Sec. 6.

4. Predicting motion

4.1. Motion representation

Formally, a motion texture is a sequence of time-varying 2D displacement maps $\mathcal{F} = \{F_t | t = 1, ..., T\}$, where the 2D displacement vector $F_t(\mathbf{p})$ at each pixel coordinate $\mathbf{p}$ from input image I_0 defines the position of that pixel at a future time t [20]. To generate a future frame at time t, one can splat pixels from I_0 using the corresponding displacement

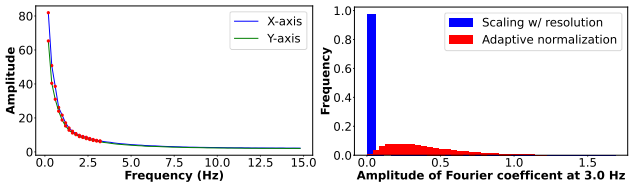

Figure 2. **Left:** We visualize the average power spectrum for the X and Y motion components extracted from real videos, shown as the blue and green curves. Natural oscillation motions are composed primarily of low-frequency components, and so we use the first $K = 16$ terms, marked with red dots. **Right:** we show a histogram of the amplitude of Fourier terms at 3.0 Hz after (1) scaling amplitude by image width and height (blue), or (2) frequency adaptive normalization (red). Our adaptive normalization prevents the coefficients from concentrating at extreme values.

map D_t, resulting in a forward-warped image I'_t:

$$I'_t(\mathbf{p} + F_t(\mathbf{p})) = I_0(\mathbf{p}). \quad (1)$$

If our goal is to produce a video via a motion texture, then one choice would be to predict a time-domain motion texture directly from an input image. However, the size of the motion texture would need to scale with the length of the video: generating T output frames implies predicting T displacement fields. To avoid predicting such a large output representation for long videos, many prior animation methods either generate video frames autoregressively [7, 29, 57, 60, 93], or predict each future output frame independently via an extra time embedding [4]. However, neither strategy ensures long-term temporal consistency of generated videos.

Fortunately, many natural motions can be described as a superposition of a small number of harmonic oscillators represented with different frequencies, amplitude and phases [20, 23, 25, 50, 69]. Because these underlying motions are quasi-periodic, it is natural to model them in the frequency domain. Hence, we adopt from Davis *et al.* [23] an efficient frequency space representation of motion in a video called a *spectral volume*, visualized in Fig. 3. A spectral volume is the temporal Fourier transform of per-pixel trajectories extracted from a video.

Given this motion representation, we formulate the motion prediction problem as a multi-modal image-to-image translation task: from an input image to an output motion spectral volume. We adopt latent diffusion models (LDMs) to generate spectral volumes comprised of a $4K$-channel 2D motion spectrum map, where $K << T$ is the number of frequencies modeled, and where at each frequency we need four scalars to represent the complex Fourier coefficients for the x- and y-dimensions. Note that the motion trajectory of a pixel at future time steps $\mathcal{F}(\mathbf{p}) = \{F_t(\mathbf{p}) | t = 1, 2, ...T\}$ and its representation as a spectral volume $\mathcal{S}(\mathbf{p}) = \{S_{f_k}(\mathbf{p}) | k = 0, 1, .. \frac{T}{2} - 1\}$ are related by the Fast Fourier transform (FFT):

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

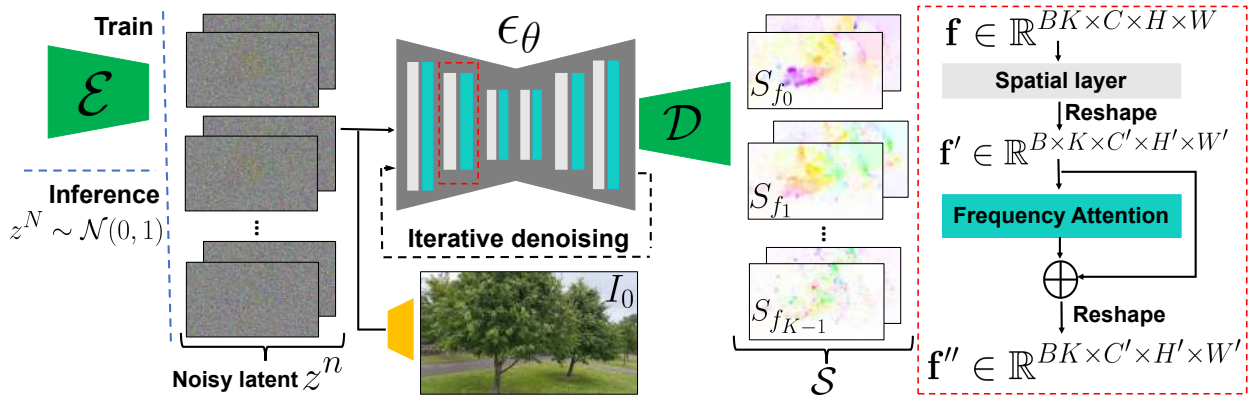

Figure 3. **Motion prediction module.** We predict a spectral volume $\mathcal{S}$ through a frequency-coordinated denoising model. Each block of the diffusion network ϵ_θ interleaves 2D spatial layers with attention layers (**red box, right**), and iteratively denoises latent features z^n. The denoised features are fed to a decoder $\mathcal{D}$ to produce $\mathcal{S}$. During training, we concatenate the downsampled input I_0 with noisy latent features encoded from a real motion texture via an encoder $\mathcal{E}$, and replace the noisy features with Gaussian noise z^N during inference (**left**).

How should we select the K output frequencies? Prior work in real-time animation has observed that most natural oscillation motions are composed primarily of low-frequency components [25, 69]. To validate this observation, we computed the average power spectrum of the motion extracted from 1,000 randomly sampled 5-second real video clips. As shown in the left plot of Fig. 2, the power spectrum of the motion decreases exponentially with increasing frequency. This suggests that most natural oscillation motions can indeed be well represented by low-frequency terms. In practice, we found that the first $K = 16$ Fourier coefficients are sufficient to realistically reproduce the original natural motion in a range of real videos and scenes.

4.2. Predicting motion with a diffusion model

We choose a latent diffusion model (LDM) [74] as the backbone for our motion prediction module, as LDMs are more computationally efficient than pixel-space diffusion models, while preserving synthesis quality. A standard LDM consists of two main modules: (1) a variational autoencoder (VAE) that compresses the input image to a latent space through an encoder $z = E(I)$, then reconstructs the input from the latent features via a decoder $I = D(z)$, and (2) a U-Net based diffusion model that learns to iteratively denoise features starting from Gaussian noise. Our training applies this process not to RGB images but to spectral volumes from real video sequences, which are encoded and then diffused for n steps with a pre-defined variance schedule to produce noisy latents z^n. The 2D U-Nets are trained to denoise the noisy latents by iteratively estimating the noise $\epsilon_\theta(z^n; n, c)$ used to update the latent feature at each step $n \in (1, 2, ..., N)$. The training loss for the LDM is written as

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

where c is the embedding of any conditional signal, such as text, or, in our case, the first frame of the training video sequence, I_0. The clean latent features z^0 are then passed through the decoder to recover the spectral volume.

Frequency adaptive normalization. One issue we observed is that motion textures have particular distribution characteristics across frequencies. As visualized in the left plot of Fig. 2, the amplitude of the spectral volumes spans a range of 0 to 100 and decays approximately exponentially with increasing frequency. As diffusion models require that the absolute values of the output are between -1 and 1 for stable training and denoising [44], we must normalize the coefficients of $\mathcal{S}$ extracted from real videos before using them for training. If we scale the magnitudes of these coefficients to [0,1] based on the image dimensions as in prior work [29, 77], nearly all the coefficients at higher frequencies will end up close to zero, as shown in the right plot of Fig. 2. Models trained on such data can produce inaccurate motions, since during inference, even small prediction errors can cause large relative errors after denormalization.

To address this issue, we employ a simple but effective frequency adaptive normalization method: First, we independently normalize Fourier coefficients at each frequency based on statistics computed from the training set. Namely, for each individual frequency f_j, we compute the 95^{th} percentile of Fourier coefficient magnitudes over all input samples and use that value as a per-frequency scaling factor s_{f_j}. We then apply a power transformation to each scaled Fourier coefficient to pull it away from extreme values. In practice, we observe that a square root performs better than other nonlinear transformations such as log or reciprocal. In summary, the final coefficient values of spectral volume $\mathcal{S}(\mathbf{p})$ at

frequency f_j (used for training our LDM) are computed as

$$S'_{f_j}(\mathbf{p}) = \text{sign}(S_{f_j}) \sqrt{\left| \frac{S_{f_j}(\mathbf{p})}{s_{f_j}} \right|}. \quad (4)$$

As shown on the right plot of Fig. 2, after applying frequency adaptive normalization, the spectral volume coefficients distribute more evenly.

Frequency-coordinated denoising. The straightforward way to predict a spectral volume $\mathcal{S}$ with K frequency bands is to output a tensor of $4K$ channels from a single diffusion U-Net. However, as in prior work [7], we observe that training a model to produce a large number of channels can yield over-smoothed, inaccurate outputs. An alternative would be to independently predict each individual frequency slice by injecting an extra frequency embedding into the LDM [4], but this design choice would result in uncorrelated predictions in the frequency domain, leading to unrealistic motion.

Therefore, inspired by recent video diffusion work [7], we propose a frequency-coordinated denoising strategy, illustrated in Fig. 3. In particular, given an input image I_0, we first train an LDM ϵ_θ to predict a single 4-channel frequency slice of spectral volume S_{f_j}, where we inject an extra frequency embedding along with the time-step embedding into the LDM. We then freeze the parameters of this LDM ϵ_θ, introduce attention layers interleaved with the 2D spatial layers of ϵ_θ across the K frequency bands, and fine-tune. Specifically, for a batch size B, the 2D spatial layers of ϵ_θ treat the corresponding $B \cdot K$ noisy latent features of channel size C as independent samples with shape $\mathcal{R}^{(B \cdot K) \times C \times H \times W}$. The attention layer then interprets these as consecutive features spanning the frequency axis, and we reshape the latent features from previous 2D spatial layers to $\mathcal{R}^{B \times K \times C \times H \times W}$ before feeding them to the attention layers. In other words, the frequency attention layers are fine-tuned to coordinate all frequency slices so as to produce coherent spectral volumes. In our experiments, we see that the average VAE reconstruction error improves from 0.024 to 0.018 when we switch from a single 2D U-Net to a frequency-coordinated denoising module, suggesting an improved upper bound on LDM prediction accuracy; in Sec. 7.3, we also show that this design choice improves video generation quality.

5. Image-based rendering

We now describe how we take a spectral volume $\mathcal{S}$ predicted for a given input image I_0 and render a future frame $\hat{I}_t$ at time t. We first derive a motion texture in the time domain using the inverse temporal FFT applied at each pixel $\mathcal{F}(\mathbf{p}) = \text{FFT}^{-1}(\mathcal{S}(\mathbf{p}))$. To produce a future frame $\hat{I}_t$, we adopt a deep image-based rendering technique and perform splatting with the predicted motion field F_t to forward-warp the encoded I_0, as shown in Fig. 4. Since forward warping can lead to holes, and multiple source pixels can map to the same output

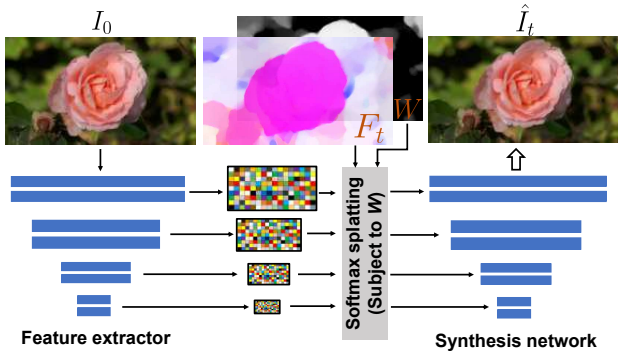

Figure 4. **Rendering module.** We fill in missing content and refine the warped input image using a deep image-based rendering module, where multi-scale features are extracted from the input image I_0. Softmax splatting is then applied over the features with a motion field F_t from time 0 to t (subject to the weights W). The warped features are fed to an image synthesis network to produce the rendered image $\hat{I}_t$.

2D location, we adopt the feature pyramid softmax splatting strategy proposed in prior work on frame interpolation [68].

Specifically, we encode I_0 through a feature extractor network to produce a multi-scale feature map. For each individual feature map at scale j, we resize and scale the predicted 2D motion field F_t according to the resolution. As in Davis *et al*. [22], we use predicted flow magnitude as a proxy for depth to determine the contributing weight of each source pixel mapped to its destination location. In particular, we compute a per-pixel weight, $W(\mathbf{p}) = \frac{1}{T} \sum_t ||F_t(\mathbf{p})||_2$ as the average magnitude of the predicted motion texture. In other words, we assume large motions correspond to moving foreground objects, and small or zero motions correspond to background. We use motion-derived weights instead of learnable ones as [46] because we observe that in the single-view case, learnable weights are not effective for addressing disocclusion ambiguities.

With the motion field F_t and weights W, we apply softmax splatting to warp the feature map at each scale to produce a warped feature. The warped features are then injected into the corresponding blocks of an image synthesis decoder to produce a final rendered image $\hat{I}_t$.

We jointly train the feature extractor and synthesis networks with start and target frames (I_0, I_t) randomly sampled from real videos, using the estimated flow field from I_0 to I_t to warp encoded features from I_0, and supervising predictions $\hat{I}_t$ against I_t with a VGG perceptual loss [49].

6. Applications

Image-to-video. Our system enables the animation of a single still picture by first predicting a motion spectral volume from the input image and generating an animation by applying our image-based rendering module to the motion

texture transformed from the spectral volume. Since we explicitly model scene motion, this allows us to produce slow-motion videos by linearly interpolating the motion texture, or to magnify (or minify) animated motions by adjusting the amplitude of the predicted spectral volume coefficients.

Seamless looping. Many applications require videos that loop seamlessly, where there is no discontinuity between the start and end of the video. Unfortunately, it is hard to find a large collection of seamlessly looping videos for training. Instead, we devise a method to use our motion diffusion model, trained on regular non-looping video clips, to produce seamless looping video. Inspired by recent work on guidance for image editing [3, 30], our method is a *motion self-guidance* technique that guides the motion denoising sampling processing using explicit looping constraints. In particular, at each iterative denoising step during inference, we incorporate an additional motion guidance signal alongside standard classifier-free guidance [45], where we enforce each pixel's position and velocity at the start and end frames to be as similar as possible:

$$\hat{\epsilon}^n = (1+w)\epsilon_\theta(z^n; n, c) - w\epsilon_\theta(z^n; n, \emptyset) + u\sigma^n \nabla_{z^n}\mathcal{L}_g^n$$
$$\mathcal{L}_g^n = ||F_T^n - F_1^n||_1 + ||\nabla F_T^n - \nabla F_1^n||_1 \quad (5)$$

where F_t^n is the predicted 2D displacement field at time t and denoising step n. w is the classifier-free guidance weight, and u is the motion self-guidance weight. In the supplemental video, we apply a baseline appearance-based looping algorithm [58] to generate a looping video from our non-looping output, and show that our motion self-guidance technique produces seamless looping videos with less distortion and fewer artifacts.

Interactive dynamics from a single image. Davis *et al*. [22] show that the spectral volume, evaluated at certain resonant frequencies, can approximate an *image-space modal basis* that is a projection of the vibration modes of the underlying scene (or, more generally, captures spatial and temporal correlations in oscillatory dynamics), and can be used to simulate the object's response to a user-defined force. We adopt this modal analysis method [22, 70], allowing us to write the image-space 2D motion displacement field for the object's physical response as a weighted sum of motion spectrum coefficients S_{f_j} modulated by the state of complex modal coordinates $\mathbf{q}_{f_j}(t)$ at each simulated time step t:

$$F_t(\mathbf{p}) = \sum_{f_j} S_{f_j}(\mathbf{p})\mathbf{q}_{f_j}(t) \quad (6)$$

We simulate the state of the modal coordinates $\mathbf{q}_{f_j}(t)$ via an explicit Euler method applied to the equations of motion for a decoupled mass-spring-damper system represented in modal space [22, 23, 70]. We refer readers to supplementary material and original work for a full derivation. Note that our method produces an interactive scene from a *single picture*, whereas these prior methods required a video as input.

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
TATS [35]	65.8	1.67	265.6	419.6	22.6	40.7
Stochastic 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
Endo *et al*. [29]	10.4	0.19	166.0	231.6	5.35	65.1
Holynski *et al*. [46]	11.2	0.20	179.0	253.7	7.23	46.8
Ours	**4.03**	**0.08**	**47.1**	**62.9**	**2.53**	**6.75**

Table 1. **Quantitative comparisons on the test set.** We report both image synthesis and video synthesis quality. Here, KID is scaled by 100. Lower is better for all error. See Sec. 7.1 for descriptions of baselines and error metrics.

7. Experiments

Implementation details. We use an LDM [74] as the backbone for predicting spectral volumes, for which we use a VAE with a continuous latent space of dimension 4. We train the VAE with an L_1 reconstruction loss, a multi-scale gradient consistency loss [54–56], and a KL-divergence loss with respective weights of $1, 0.2, 10^{-6}$. We train the same 2D U-Net used in the original LDM work to perform iterative denosing with a simple MSE loss [44], and adopt the attention layers from [41] for frequency-coordinated denoising. For quantitative evaluation, we train both VAE and LDM on images of size 256×160 from scratch for fair comparisons, and it takes around 6 days to converge using 16 Nvidia A100 GPUs. For our main quantitative and qualitative results, we run the motion diffusion model with DDIM [86] for 250 steps. We also show generated videos of up to a resolution of 512×288, created by fine-tuning a pre-trained image inpainting LDM model [74] on our dataset.

We adopt ResNet-34 [39] for the feature extractor in our IBR module. Our image synthesis network is based on an architecture for conditional image inpainting [57, 110]. Our rendering module runs in real-time at 25FPS on a Nvidia V100 GPU during inference. We adopt universal guidance [3] to produce seamless looping videos, where we set weights $w = 1.75, u = 200$, and use 500 DDIM steps with 2 self-recurrence iterations.

Data. We collect and process a set of 3,015 videos of natural scenes exhibiting oscillatory motions from online sources our own captures. We withhold 10% of videos for testing and use the remainder for training. To extract ground truth motion trajectories, we apply a coarse-to-fine flow method [10, 61] between each selected starting image and every future frame of the video. As training data, we take every 10th video frame as input images and derive the corresponding ground truth spectral volumes using the computed motion trajectories across the following 149 frames. In total, our data consists of over 150K image-motion pairs.

Figure 5. X-t **slices of videos generated by different approaches.** From left to right: input image and corresponding X-t video slices from the ground truth video, from videos generated by three baselines [27, 29, 46, 93], and finally videos generated by our approach.

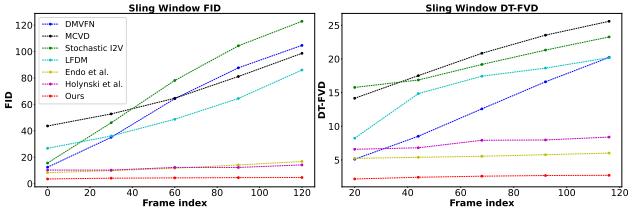

Figure 6. **Sliding window FID and DTFVD.** We show sliding window FID with window size 30 frames, and DTFVD with size 16 frames, for videos generated by different methods.

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
Repeat 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
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat [46]	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

Table 2. **Ablation study.** Sec. 7.3 describes each configuration.

Baselines. We compare our approach to recent single-image animation and video prediction methods. Endo *et al*. [29] and DMVFN [48] predict instantaneous 2D motion fields and render future frames auto-regressively. Holynski *et al*. [46] instead simulates motion through a single static Eulerian motion description. Other recent work such as Stochastic Image-to-Video (Stochastic-I2V) [27], TATS [35], and MCVD [93] adopt VAEs, transformers, or diffusion models to directly predict raw video frames; LFDM [67] generates future frames by predicting flow volumes and warping latents in a diffusion model. We train all the above methods on our data using their respective open-source implementations.[1]

We evaluate the quality of the videos generated by our approach and by prior baselines in two ways. First, we evaluate the quality of individual synthesized frames using metrics designed for image synthesis tasks. We adopt the Fréchet Inception Distance (FID) [42] and Kernel Inception Distance (KID) [5] to measure the average distance between the distributions of generated frames and ground truth frames.

Second, to evaluate the quality and temporal coherence of synthesized videos, we adopt the Fréchet Video Distance [92] with window size 16 (FVD) and 32 (FVD_{32}), based on an I3D model [11] trained on the Human Kinetics datasets [52]. To more faithfully reflect synthesis quality for the natural oscillation motions we seek to generate, we also adopt the Dynamic Texture Frechet Video Distance [27], which measures distance from videos with window size 16 (DTFVD) and size 32 ($DTFVD_{32}$), using a I3D model trained on the Dynamic Textures Database [37], a dataset consisting primarily of natural motion textures.

We further use a sliding window FID of window size 30 frames, and a sliding window DTFVD with window size 16 frames, as in [57, 60], to measure how generated video quality degrades over time. For all methods, we evaluate metrics at 256×128 resolution by center-cropping.

[1] We use the open-source implementation of [46] from Fan *et al*. [83].

Input AnimateDiff ModelScope GEN-2

Figure 7. We show generated future frames from three recent large video diffusion models [31, 36, 98].

Figure 8. **Limitations.** We show examples of rendered future frames (even), and overlay of input and rendered images (odd). Our method can produce artifacts in regions of thin objects or large motions, and regions requiring filling large amount of new contents.

7.1. Quantitative results

Table 1 shows quantitative comparisons between our approach and baselines on our test set. Our approach significantly outperforms prior single-image animation baselines in terms of both image and video synthesis quality. Specifically, our much lower FVD and DT-FVD distances suggest that the videos generated by our approach are more realistic and more temporally coherent. Further, Fig. 6 shows the sliding window FID and sliding window DT-FVD distances of generated videos from different methods. Thanks to the global spectral volume representation, videos generated by our approach do not suffer from degradation over time.

7.2. Qualitative results

We visualize qualitative comparisons between videos as spatio-temporal X-t slices of the generated videos, a standard way of visualizing small motions in a video [95]. As shown in Fig. 5, our generated video dynamics more strongly resemble the motion patterns observed in the corresponding real reference videos (second column), compared to other methods. Baselines such as Stochastic I2V [27] and MCVD [93] fail to model both appearance and motion realistically over time. Endo *et al*. [29] and Holynski *et al*. [46] produce video frames with fewer artifacts but exhibits oversmooth or non-oscillatory motion over time. We refer readers to supplementary material to assess the quality of generated video frames and estimated motion across different methods.

7.3. Ablation study

We conduct an ablation study to validate the major design choices in our motion prediction and rendering modules, comparing our full configuration with different variants. Specifically, we evaluate results using different numbers of frequency bands $K = 4, 8, 16, 24$. We observe that increasing the number of frequency bands improves video prediction quality, but the improvement is marginal with more than 16 frequencies. Next, we remove adaptive frequency normalization from the ground truth spectral volumes, and instead just scale them based on input image width and height (*w/o adaptive norm.*). Additionally, we remove the frequency coordinated-denoising module (*Independent pred.*), or replace it with a simpler DM where a tensor volume of $4K$ channel spectral volumes are predicted jointly via a single 2D U-net diffusion model (*Volume pred.*). Finally, we compare results where we use a baseline rendering method that applies softmax splatting over single-scale features subject to learnable weights as used by [46] (*Baseline splat*). We also add a baseline where the generated video is a volume by repeating input image N times (Repeat I_0). From Table 2, we observe that all simpler or alternative configurations lead to worse performance compared with our full model.

7.4. Comparing to large video models

We further perform a user study and compare our generated animations with ones from recent large video diffusion models: AnimateDiff [36], ModelScope [98] and Gen-2 [31], which predict video volumes directly. On a randomly selected 30 videos from the test set, we ask users "which video is more realistic?". Users report a 80.9% preference for our approach over others. Moreover, as shown in Fig. 7, we observe that the generated videos from these baselines are either unable to adhere to the input image content, or exhibit gradual color drift and distortion over time. We refer readers to the supplementary material for full comparisons.

8. Discussion and conclusion

Limitations. Since our approach only predicts lower frequencies of a spectral volume, it can fail to model non-oscillating motions or high-frequency vibrations—this may be resolved by using learned motion bases. Furthermore, the quality of generated videos relies on the quality of underlying motion trajectories, which may degrade in scenes with thin moving objects or objects with large displacements. Even if correct, motions that require generating large amounts of new unseen content may also cause degradations (Fig. 8).

Conclusion. We present a new approach for modeling natural oscillation dynamics from a single still picture. Our image-space motion prior is represented with spectral volumes, a frequency representation of per-pixel motion trajectories, which we find to be efficient and effective for prediction with diffusion models, and which we learn from collections of real world videos. Spectral volumes are predicted using frequency-coordinated latent diffusion model and are used to animate future video frames through an image-based rendering module. We show that our approach produces photo-realistic animations from a single picture and significantly outperforms prior baselines, and that it can enable several downstream applications such as creating seamlessly looping or interactive image dynamics.

Acknowledgements. We thank Abe Davis, Rick Szeliski, Andrew Liu, Boyang Deng, Qianqian Wang, Xuan Luo, and Lucy Chai for fruitful discussions and helpful comments.

References

[1] Aseem Agarwala, Ke Colin Zheng, Chris Pal, Maneesh Agrawala, Michael Cohen, Brian Curless, David Salesin, and Richard Szeliski. Panoramic video textures. In *ACM Trans. Graphics (SIGGRAPH)*, pages 821–827. 2005.

[2] Hyemin Ahn, Esteve Valls Mascaro, and 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. Universal guidance for diffusion models. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 843–852, 2023.

[4] Hugo Bertiche, Niloy J Mitra, Kuldeep Kulkarni, Chun-Hao P Huang, Tuanfeng Y Wang, Meysam Madadi, Sergio Escalera, and Duygu Ceylan. Blowing in the wind: Cyclenet for human cinemagraphs from still images. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2023.

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

[6] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 14707–14717, 2021.

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023.

[8] Richard Strong Bowen, Richard Tucker, Ramin Zabih, and Noah Snavely. Dimensions of motion: Monocular prediction through flow subspaces. In *International Conference on 3D Vision (3DV)*, pages 454–464, 2022.

[9] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Neural Information Processing Systems*, 35:31769–31781, 2022.

[10] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 25–36, 2004.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[12] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4d video textures for interactive character appearance. In *Computer Graphics Forum*, volume 33, pages 371–380. Wiley Online Library, 2014.

[13] Antoni B Chan and Nuno Vasconcelos. Mixtures of dynamic textures. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 641–647, 2005.

[14] Antoni B Chan and Nuno Vasconcelos. Classifying video with kernel dynamic textures. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[15] Antoni B Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *Trans. Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008.

[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: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[17] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 11315–11325, 2022.

[18] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023.

[19] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2023.

[20] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H Salesin, and Richard Szeliski. Animating pictures with stochastic motion textures. In *ACM Trans. Graphics (SIGGRAPH)*, pages 853–860, 2005.

[21] Vincent C Couture, Michael S Langer, and Sebastien Roy. Omnistereo video textures without ghosting. In *International Conference on 3D Vision*, pages 64–70. IEEE, 2013.

[22] Abe Davis, Justin G Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Trans. Graphics (SIGGRAPH)*, 34(6):1–7, 2015.

[23] Myers Abraham Davis. *Visual vibration analysis*. PhD thesis, Massachusetts Institute of Technology, 2016.

[24] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems*, 34:8780–8794, 2021.

[25] Julien Diener, Mathieu Rodriguez, Lionel Baboud, and Lionel Reveret. Wind projection basis for real-time animation of trees. In *Computer graphics forum*, volume 28, pages 533–540. Wiley Online Library, 2009.

[26] Gianfranco Doretto, Alessandro Chiuso, Ying Nian Wu, and Stefano Soatto. Dynamic textures. 51:91–109, 2003.

[27] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 3742–3753, June 2021.

[28] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2023.

[29] Yuki Endo, Yoshihiro Kanamori, and Shigeru Kuriyama. Animating landscape: Self-supervised learning of decoupled motion and appearance for single-image video synthe-

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. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023.

[31] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 7346–7356, 2023.

[32] Matthew Flagg, Atsushi Nakazawa, Qiushuang Zhang, Sing Bing Kang, Young Kee Ryu, Irfan Essa, and James M Rehg. Human video textures. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pages 199–206, 2009.

[33] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR, 2020.

[34] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2Flow: Motion hallucination from static images for action recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022.

[36] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[37] Isma Hadji and Richard P Wildes. A new large scale dynamic texture dataset with application to convnet understanding. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 320–335, 2018.

[38] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 7854–7863, 2018.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[40] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023.

[41] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Neural Information Processing Systems*, 30, 2017.

[43] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[44] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, 33:6840–6851, 2020.

[45] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[46] Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. Animating pictures with Eulerian motion fields. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5819, 2021.

[47] Tobias Hoppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *Trans. Mach. Learn. Res.*, 2022, 2022.

[48] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. *ArXiv*, abs/2303.09875, 2023.

[49] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 694–711, 2016.

[50] Hitoshi Kanda and Jun Ohya. Efficient, realistic method for animating dynamic behaviors of 3d botanical trees. In *International Conference on Multimedia and Expo*, volume 2, pages II–89. IEEE, 2003.

[51] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint 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 human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[53] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[54] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 4521–4530, 2019.

[55] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018.

[56] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 4273–4284, 2023.

[57] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 515–534. Springer, 2022.

[58] Jing Liao, Mark Finch, and Hugues Hoppe. Fast computation of seamless video loops. *ACM Trans. Graphics (SIG-*

GRAPH), 34(6):1–10, 2015.

[59] Zicheng Liao, Neel Joshi, and Hugues Hoppe. Automated video looping with progressive dynamism. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[60] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 14458–14467, 2021.

[61] Ce Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[62] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023.

[63] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2022.

[64] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. Text-guided synthesis of eulerian cinemagraphs. 2023.

[65] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Neural Information Processing Systems*, 35:22438–22450, 2022.

[66] Medhini Narasimhan, Shiry Ginosar, Andrew Owens, Alexei A Efros, and Trevor Darrell. Strumming to the beat: Audio-conditioned contrastive video textures. In *Proc. Winter Conference on Applications of Computer Vision*, pages 3761–3770, 2022.

[67] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 18444–18455, 2023.

[68] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 5437–5446, 2020.

[69] Shin Ota, Machiko Tamura, Kunihiko Fujita, T Fujimoto, K Muraoka, and Norishige Chiba. 1/f/sup/spl beta//noise-based real-time animation of trees swaying in wind fields. In *Proceedings Computer Graphics International*, pages 52–59. IEEE, 2003.

[70] Automne Petitjean, Yohan Poirier-Ginter, Ayush Tewari, Guillaume Cordonnier, and George Drettakis. Modalnerf: Neural modal analysis and synthesis for free-viewpoint navigation in dynamically vibrating scenes. In *Computer Graphics Forum*, volume 42, 2023.

[71] Silvia L. Pintea, Jan C. van Gemert, and Arnold W. M. Smeulders. Déjà vu: Motion prediction in static images. In *Proc. European Conf. on Computer Vision (ECCV)*, 2014.

[72] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023.

[73] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. Computer Vision 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. *Neural Information Processing Systems*, 35:36479–36494, 2022.

[76] Payam Saisan, Gianfranco Doretto, Ying Nian Wu, and Stefano Soatto. Dynamic texture recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2001.

[77] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation, 2023.

[78] Arno Schödl, Richard Szeliski, David H Salesin, and Irfan Essa. Video textures. In *ACM Trans. Graphics (SIGGRAPH)*, pages 489–498, 2000.

[79] Mikio Shinya and Alain Fournier. Stochastic motion—motion under the influence of wind. *Computer Graphics Forum*, 11(3), 1992.

[80] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 2377–2386, 2019.

[81] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Neural Information Processing Systems*, 32, 2019.

[82] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 13653–13662, 2021.

[83] Chen Qian Kwan-Yee Lin Hongsheng Li Siming Fan, Jingtan Piao. Simulating fluids in real-world still images. *arXiv preprint*, arXiv:2204.11335, 2022.

[84] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 3626–3636, 2022.

[85] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[86] Jiaming Song, Chenlin Meng, and Stefano Ermon. 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. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[88] Jos Stam. *Multi-scale stochastic modelling of complex natural phenomena*. PhD thesis, 1995.

[89] Jos Stam. Stochastic dynamics: Simulating the effects of turbulence on flexible structures. *Computer Graphics Forum*,

16(3), 1997.

[90] Ryusuke Sugimoto, Mingming He, Jing Liao, and Pedro V Sander. Water simulation and rendering from a still photograph. 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. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[92] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[93] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *Neural Information Processing Systems*, 2022.

[94] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Neural Information Processing Systems*, 2016.

[95] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Trans. Graphics (SIGGRAPH)*, 32(4):1–10, 2013.

[96] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Proc. European Conf. on Computer Vision (ECCV)*, 2016.

[97] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 2443–2451, 2015.

[98] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023.

[99] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[100] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. *arXiv preprint arXiv:2304.10532*, 2023.

[101] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

[102] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917, 2019.

[103] Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2023.

[104] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman. Visual dynamics: Stochastic future generation via layered cross convolutional networks. *Trans. Pattern 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: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[106] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 18456–18466, 2023.

[107] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023.

[108] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

[109] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 3657–3666, 2022.

[110] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[111] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.