基于扩散模型的视频生成发展简述

李林泽　　范孙奇

摘　要

基于扩散模型的图像生成已经发展得风生水起，以 CVPR2023 为例，有 99 篇基于扩散模型的工作被接收，随着 Stable Diffusion、Lora 等工作的开源，扩散模型的浪潮更是席卷到非该研究领域的群众中。考虑电影电视、游戏、短视频、广告等视觉制作领域，视频生成也许是这个浪潮的下一波。对比图像生成，视频生成任务与其有哪些联系呢？又面临哪些更大的挑战？

本文致力于梳理清楚基于扩散模型的视频生成发展的脉络，我们将回顾视频生成与图像生成的联系，从多个角度描述该领域的发展，阐述视频生成领域的难点以及解决方案，介绍该领域的 Sota 方法，展示各个公司在该方向上作品的效果。

关键词： 视频生成，扩散模型

图 1: 扫码观看各厂 Demo 效果

1　视频生成难点与挑战

在视频生成领域，研究人员和开发者面临诸多挑战。首先，大量、多样化且高清的带标注的**视频资源**目前是缺乏的，这可能会限制模型的性能和泛化能力；其次，相比于图像任务，视频生成需要成几十倍的**计算成本**，需要在提高采样效率和加速训练过程方面取得进展，以便更快地生成高质量视频；**空间连续性**难以保障也是一个挑战，研究人员必须考虑生成视频的帧间一致性，以及是否符合**运动学规律**，或者说怎样去除帧间的闪烁现象；此外，准确的**视频编辑**是另一个关键问题，如何对场景、物体和光线进行精细控制，以完成 prompt 的生成要求？另外，针对**长视频**的训练和扩展视频长度也是一个值得关注的问题，处理较长视频往往需要大量的计算资源，开发可扩展且高效的方法以处理更长的视频序列是一个重要研究方向。最后，保证生成视频的**高清程度**是另一个关键挑战，生成的细节逼真丰富的同时还要保证帧间的一致性就更是难上加难了。

2　视频生成发展脉络

2.1　从文生图到文生视频

相比于图像生成，视频生成是一个更年轻的领域，而基于扩散模型的视频生成就更加年幼了，最早的工作可以追溯到 2022 年 4 月谷歌出品的 Video Diffusion Models（VDM）[1]，通过将 UNet 网络架构修改为时空分离注意力机制，使文生图模型能够适应文生视频任务。具体的，原 UNet 中的 2D 卷积替换成 Space-Only 空间注意力 3D 卷积，每个空间注意力模块后面新接一个时间注意力模块；该方法还提出了对视频和图片进行联合训练，通过掩码的方式来阻止视频帧与图像帧在时间注意力模块上信息混淆；最后针对空间和时间视频扩展提出了一种新的重构指导条件采样，从而可以自回归地扩展视频到更长的时间步和更高的分辨率。VDM[1] 身先士卒提出了针对视频资源缺乏、空间连续性难以保障、生成视频长度受限这三个问题的解决方法，那么这几个问题有更好的解法吗？其他挑战又该如何应对呢？

2.2 从大规模训练到 Zero-Shot

2022 年 10 月谷歌出品了 Imagen Video[2]。Imagen Video[2] 采用级联的思维，包含 1 个文本编码器、1 个基础视频扩散模型以及 3 个空间超分辨率模型和 3 个时间超分辨率模型，这 7 个视频扩散模型共有 116 亿参数。训练集使用了谷歌内部未公开的 1400 万视频文本对以及 6000 万图像文本对数据集，此外还使用 LAION-400M 图像文本对数据集。虽然有钱确实可以为所欲为，然而并非所有研究人员都可以获取大规模高质量带标注的视频数据，也缺乏如此昂贵的计算资源，因此很多方法研究如何将一个现有的图像生成模型以最小的代价迁移到视频生成模型。2022 年 12 月腾讯出品了 Tune-A-Video[3]，利用已经训练好的文生图模型，仅仅需要一个文本视频对来微调文生图模型的一小部分参数就可以训练出 One-Shot 的文生视频模型。从文生图模型到文生视频模型的迁移，需要解决帧间内容一致性以及运动的连续性问题，Tune-A-Video[3] 通过扩展空间自注意力到时空注意力解决了帧间内容一致性，通过使用 DDIM 反衍提供每帧的结构信息解决了运动连续性问题。2023 年 3 月 Text2Video-Zero[4] 沿着这个方向又迈进了一步，**无需任何数据无需任何训练与优化**，论文里展现出来的效果甚至超过了 Tune-A-Video[3]。Text2Video-Zero[4] 有两个关键改动：首先用运动动态来丰富潜在编码从而保证全局场景与背景在时间上的一致性，然后使用跨帧注意力机制来保证每帧前景的内容、外观以及属性一致。

2.3 从 12 帧到 3376 帧

回首视频生成工作，大部分都是 3 到 5 秒钟的超短视频，然而在现实应用中几乎都是分钟级别的，即使是这几年很火的短视频也是几十秒级别，生成长视频在现实应用中是一个非常强的需求。然而生成长视频往往需要大量计算资源，为了应对这个挑战，一部分工作（如 VDM[1], Latent Video Diffusion Models for High-Fidelity Long Video Generation，Flexible Diffusion Modeling of Long Videos）使用自回归的方式在推理阶段通过滑窗的方式扩展视频长度，然而这类方法在短视频上做训练在长视频上做推理会造成训练推理间的差距，比如不切实际的镜头变化以及长序列的不连贯性，另外因为串行生成导致序列的生成过程也是低效的。2023 年 3 月微软提出了 NUWA-XL[5]，输入 16 句简单描述就能生成一段长达 **11 分钟**的动画片，NUWA-XL[5] 采用 coarse-to-fine 的思路，可以并行高效的生成视频。具体的，首先用一个基础的扩散模型来生成关键帧，然后局部模型负责递归的填充附近帧的内容，这种做法可以直接训练 3376 帧长度的视频，消除了训练和推理不一致造成的差异，保障生成视频的质量与连续性，并且生成 1024 帧视频只需要 26s。Gen-L-Video[6] 这篇工作将长视频分割成彼此之间有重叠的短视频片段，每一个短视频片段都利用现有短视频扩散模型进行生成，重叠部分则根据二范数最小距离对多个备选答案做了融合，最后拼在一起就形成了长视频。

2.4 从文生图到多种引导控制手段

由于时序动态上的巨大变化以及对跨帧一致性的要求，视频生成的可控性要比图像生成困难很多。正如文生图领域 Stable Diffusion 向 ControlNet 的发展一样，视频生成领域也不再限于只用文字作为 prompt。2023 年 2 月 Runway 发布了 Gen-1[7]，整合了深度图通过跨注意力机制来提供结构指导信息。MCDiff[8] 基于初始帧，结合人体各部位运动情况，生成流畅自然的人体运动视频。LaMD[9] 试图将视频的运动部分与静态背景分离开来，压缩成潜码表示，来控制扩散模型生成。VideoComposer[10] 通过组合式生成范式同时实现视频在时间和空间两个维度上的可控性，将视频分解成三种引导条件，即文本条件、空间条件、视频特有的时序条件。2023 年 3 月 Runway 发布了 Gen-2 并在最近开放用户使用，Gen-2 主打的就是视频生成 prompt 或者说编辑手段的多样化，除了基础的文生视频功能，还包括：（1）**文加图生成视频**：输入一幅图加一句文字描述 prompt 生成视频（2）**图生视频**：输入一幅图生成视频，比如一张静止的沙滩照变嬉戏打闹的沙滩场景视频（3）**风格化**：输入视频加想要达到的风格化图像生成视频，比如把你的自拍视频风格化成绿巨人（4）**故事板**：将输入视频转变为新的风格化并完成动画的渲染，比如将堆起的书本渲染成摩天大楼（5）**目标编辑**：分离视频中的目标并使用简单的文本提示对目标进行修

改，比如给视频中的小狗加上斑点。（6）**渲染**：将无纹理的渲染转换为真实的输出，比如将无纹理人体模型渲染成真实的水下游泳视频（7）**定制化**：定制化训练模型以获得更高的保真度。Prompt 手段的多样化又催生了视频编辑这个研究方向，除了和视频生成一样的挑战，视频编辑还需要考虑如何定位编辑的区域、如何保持编辑后的部分与原始图像整体语义一致等难点。

2.5 帧间一致性与运动连续性

视频生成的重大挑战是如何保持视频的帧间一致性与运动连续性。视频生成扩散模型几乎都考虑了这个挑战，但近期仍有不少工作从扩散模型的不同的角度入手，试图加强生成视频的时序一致性。PYoCo[11] 发现从图像扩展到视频的过程中，帧与帧之间的噪声标签并不是独立的，而是具有很强的关联。作者试图用数学建模构建这种关联，从而加强生成视频的帧间一致性。Latent-Shift[12] 在 UNet 中将相邻 3 帧的特征拼接起来进行融合，相当于增加了时序维度上的卷积，同时兼顾计算效率与时序一致性。VIDM[13] 先生成视频第一帧，根据第一帧和最新一帧得到光流编码，进而生成下一帧，迭代自回归地生成最终视频。

3 细化到特定领域

用于视频生成的扩散模型在大规模数据上训练过，具有一定的普适性，但其在特定领域的性能仍有提升的空间，比如通用的视频生成模型在人像视频生成上常常会出现“鬼脸”。最常见的方法是使用特定领域的数据去微调扩散模型，近来也涌现出了一些无需微调的方法。Video-Adapter[14] 不微调原先预训练好的大规模扩散模型，而是用特定领域（如机器人、动画、自视角视频等领域）的数据训练了一个小规模的视频生成扩散模型；利用扩散模型的能量表示，将这一大一小两个模型组合起来，同时保留了大小模型的优点。

在特定领域中，人的面部动作丰富、表情灵活，深受研究者重视；人脸动画（即让原面部具有目标面部的动作与表情）、面部编辑、说话头合成（talking head synthesis）、妆容迁移等任务具有较高实用价值。这些任务原先主流方式是基于条件生成对抗网络，近年来扩散模型也被应用于这些任务。DiffTalk[15] 使用语音、图片、面部标识作为条件，引导扩散模型生成说话头（talking head）。FADM[16] 试图用扩散模型解决人脸动画任务，对目标面部的姿态和外观抽取特征作为扩散模型的条件，指导模型生成精细的人脸动画。Diffusion Video Autoencoders[17] 针对面部编辑任务，在扩散模型加噪过程中就使用原面部抽取的多种特征作为条件，再根据编辑要求，利用分类器或者 CLIP Loss 对这些特征进行编辑，在去噪过程中以新的特征作为条件。

4 各公司视频生成作品对比

表 1按照发布时间顺序汇总了各大厂的代表方法，彼此都各有特色，如 NUWA-XL 可以生成长达 11 分钟的情节与风格都连贯的视频，Align Your Latent 生成的视频相当高清和真实，最近大火的 Gen-2 则给用户带来了更多新奇的视频生成与编辑体验。

公司	方法	亮点
Google	VDM	首篇基于扩散模型视频生成工作
智源	CogVideo	Transformer-based 方法，用作对比
Meta	Make-A-Video	无需视频数据，有图像广博性
ByteDance	MagicVideo	在隐空间中生成视频
Tencent	Tune-A-Video	One shot
Runway	Gen2	多种编辑手段
Ali	VideoFusion	将扩散过程解耦来完成帧间建模
Microsoft	NUWA-XL	超长视频
NVIDIA	AlignYourLatent	高清视频，效果出色
Meta	Latent-shift	隐空间，无需添加额外参数
Sensetime	Gen-L-Video	扩展短视频生成模型到数百帧
NVIDIA	Pyoco	描述主角、动作和位置生成视频

表 1: 各公司视频生成方法汇总

我们汇总了各公司视频生成方法的 Demo 效果，比较粗糙的分为了自然景观和人像两个类别，感兴趣的同学可以扫图 1二维码观看。需要说明的是，在实际测试每个方法时，我们发现大家对自身生成结果都有择优挑选再展示，用我们的样例测试也不见得公平，因此收集的都是各公司展示的 Demo 效果。

5 数据集与评价指标

视频生成模型使用的数据集大多由文字视频（图片）对组成，常用的大规模数据集包括：

- WebVid-10M：10.7M 视频文字对，来源于网络，种类丰富
- MSR-VTT：20 个大类，10000 条视频，每个视频配备 20 条文字描述
- HD-VILA-100M：15 个大类，100M 视频与文字对
- LAION-400M: 400M 图片文字对, 利用 CLIP 进行图片文字匹配度进行筛选过
- DAVIS: 纯视频数据集，需要自己标注文字描述

另外还有一些特定领域的数据集：

- 风景：Landscape, Cityscapes, Sky Timelapse
- 人体：Taichi-HD, Human3.6M, MPII Human Pose
- 物体：BAIR Robot Pushing，CATER-GEN
- 动作识别：UCF-101，KTH Action, MHAD

以下指标常用于评判视频生成的好坏：

- 与 groundtruth 视频逐帧比较：Fréchet Image Distance, LPIPS, PSNR, SSIM
- 与 groundtruth 视频整体比较：Fréchet Video Distance, Kernel Video Distance, Continuous Ranked Probability Score
- 无标签逐帧评价：Inception Score
- 检验视频的时序一致性：视频各帧之间 CLIP Embedding 的相似度
- 检验视频与文本的一致性：视频各帧与文本的 CLIP Embedding 的相似度
- 用户调查（User Study）

6 总结

短短的一年时间，基于扩散模型的视频生成工作层出不穷，生成内容越来越真实、连贯、高清、可控的同时，新颖的视频生成与编辑手段也令人眼花缭乱，但目前视频生成与编辑仍处于探索期，每个方法都有自己的局限性，期待生成式模型在视频领域有更大的研究进展，也期待视频生成的落地工作能像 ChatGPT 和 Stable Diffusion 一样迈入千门万户。

参考文献

[1] “Video diffusion models.”

[2] “Imagen video: High definition video generation with diffusion models.”

[3] “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation.”

[4] “Text2video-zero: Text-to-image diffusion models are zero-shot video generators.”

[5] “Nuwa-xl: Diffusion over diffusion for extremely long video generation.”

[6] “Gen-l-video: Multi-text to long video generation via temporal co-denoising.”

[7] “Structure and content-guided video synthesis with diffusion models.”

[8] “Motion-conditioned diffusion model for controllable video synthesis.”

[9] “Lamd: Latent motion diffusion for video generation.”

[10] “Videocomposer: Compositional video synthesis with motion controllability.”

[11] “Preserve your own correlation: A noise prior for video diffusion models.”

[12] “Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation.”

[13] “Vidm: Video implicit diffusion models.”

[14] “Probabilistic adaptation of text-to-video models.”

[15] “Difftalk: Crafting diffusion models for generalized talking head synthesis,”

[16] “Face animation with an attribute-guided diffusion model.”

[17] “Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding.”