# 论文学习——Video LDM (Align your Latents)

## Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models

### 0. 来源

本文是阅读论文后的个人笔记，适应于个人水平，叙述顺序和细节详略与原论文不尽相同，并不是翻译原论文。

如果想了解所有细节，建议移步 arxiv 。

论文地址：https://arxiv.org/abs/2304.08818
项目地址：https://research.nvidia .com/labs/toronto-ai/VideoLDM/

### 1. 整体结构

本文基于 图像生成 领域的典型框架LDM，扩展到视频生成领域，且保留了其以低计算成本，在压缩后的低维潜空间内生成高质量图片的优点，最终能够局连贯的高分辨率长视频。

模型 整体训练流程可表述如下：
a. 在图像数据上 预训练 LDM，或是使用available的图像LDM预训练模型；
b. 通过在LDM（latent space diffusion model）中引入时间层，固定空间层（spatial layers）参数，并在视频数据上进行微调，以实现将图像 生成器 转换分为两部分，1.将图像的**潜向量**生成器转换为视频的**潜向量**生成器。2.将自编码器在像素空间上进行时序对齐。
c. 改造为预测模型以获得较长视频的生成（Long-Term Generation）
d. 使用时序的插值以获得高帧率
e. 在时间上对齐 图像DM 的上采样器，将其转换为时序一致的视频超分模型。（仅在需要合成高分辨率视频时选择进行该项，video upsampler仅在局部上留较低的计算成本）

要生成连续的长视频，其流程如下图所示

1 . 首先生成离散的关键帧；
2+3. 分两步使用相同的插值模型，在关键帧之间进行时序插值，以实现较高的帧率；
（以上三步均基于LDM模型，且它们共享相同的image backbone，分别进行微调）
4. 将潜向量解码到像素空间
5. （可选）使用视频上采样DM得到更高的分辨率

该模型的训练过程其实就是对同一款预训练的图片LDM（及DM上采样器）的不同微调过程，下面将逐步对以上a,b,c三步进行解释

### 2. 具体实现

#### b.1 将潜向量生成器由图像领域转到视频领域

现有的空间LDM能够高质量地生成独立帧，但难以生成多张连续的视频帧，这是由于其没有时间的概念。

如上左图所示，在原有LDM的空间层中交错插入时间层，时间层包括3d卷积层和时间注意力层，以时序一致性的方式对齐独立的帧，而在优化过程中需要而仅优化时间层。

上右图，则取出了一套“空间层+时间层”对运算过程进行了解释，在空间层和时间层，它们对（TxCxHxW）视频的理解不一样，空间层将时间维度并入ba做batch里的一小串无关的照片，B*T成为了新的batch size而已，故而输入空间层的张量格式为 (b t) c h w；而时间层将整个视频按时间维度排列，是一个层的张量格式为b c t h w。张量通过时间层前后需要进行如下变形：

在每一步运算之后，需要对输入时间层前后的张量z和z’进行加权。

$$\alpha_\phi^i z + (1 - \alpha_\phi^i) z', \text{其中} \alpha_\phi^i \in [0, 1]$$

关于图中其他可疑的点，如$c_s$，表示在训练预测模型时使用的对上下文帧的掩码

### b.2 自编码器的时序微调

直接将图像领域的LDM的自编码器用于时序上连贯的视频，会导致生成的视频出现闪烁的假象（flickering artifacts）。

为了保证在潜向量空间上预训练的LDM模型能够被复用，故而保留编码器不变，仅对解码器进行微调。

以视频为微调数据集，微调采样的是3d卷积搭建出的（patch-wise的）时序判别器。

需要注意的是，我们b.1中生成的是图像或视频的潜向量，是以特征的形式存在的，如上图底部，不同的潜特征对于不同的分布峰值，将通过decoder解码于不同的区间上的潜向量将decode出不同的图像。通过观察这个特性，可以解释对framework进行视频微调时帧之间的时序一致性

可以看见，不原本散落在不同峰值附近的图像内容各自独立，经过视频微调之后，它们处于同一峰值附近，而图像也呈现出内容上的连贯性。

### c. 改造为预测模型以获得长期的生成结果

b章节中使用的方法难以生成长视频序列，故而我们喂入S个上下文帧，训练模型成为一个预测模型。这是通过时序的二进制掩码来实现的，在长度为T的视频帧，而掩盖住T-S个要预测的视频帧。将视频帧编码后，乘以掩码，再经过已经学到的下采样操作（learned downsampling operation：resize+conv2学到的下采样，可以看看代码）并喂入到时序层中。

在推理过程中，为了生成场视频可以迭代地进行采样过程，复用最新的预测作为新的上下文。第一个初始序列的生成方式：从base image model生成单一于此生成初始序列；其余序列的生成办法：使用两个上下文帧来编码移动。为稳定这一过程，本文使用无分类器的扩散引导来引导采样过程，如下式。

本章的目的是生成关键帧，虽然较少的帧节省了内存，但不同帧之间仍然存在较大的语义变化，为实现高帧率，同时实现连贯性，下一章将对帧之间进行

### d. 时序插值以获得更高的帧率

沿用c章节中提到的条件掩码机制（masking-conditioning mechanism）在两两关键帧之间插值，不过掩码的对象是要被插值的帧，否则就和c章节一样了化为一个视频插值模型。在实验中，单次插值的结果是视频长度从T转为4T，可迭代使用两次，转为16T。

### e. 对超分模型进行时序微调

受级联DM的启发，作者尝试再使用一层上采样器如pixel -space DM 或LDM upsampler来增加单张图片的分辨率，但对各帧独立地上采样会导致时间一超分模型也需要具有时间上的视野，选择如同b章节中介绍的方法一样来微调上采样器。由于上采样器是仅在聚不上进行操作，所以可以高效地在patch上样器，然后卷积式地应用到模型上。

## 3. 不同任务领域及各自细节

本模型应用于户外驾驶数据的仿真、基于text2video的创造性内容生成

另外，该模型可以对现有的图像LDM仅训练出一套在时间上对齐的模型，而解锁不同的（对image LDM进行微调而得到的变种）LDMs的视频版本。