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

沙之洲 2020012408

本篇文章提出了一种利用训练好的 image diffusion model 来做 video generation 的方法。具体来说，是在 image diffusion 的 spatial layer 中插入了一些 temporal layer。而且，image batch 在 spatial layer 和 temporal layer 之间会进行转换，对 spatial layer 来说，image batch 是不存在时序信息的。

本篇文章的优势是提出的 video generation model 可以生成 high resolution 和更长时间的 video。具体来说是通过超分的 upsampler 实现的，而 diffusion 只需要生成 resolution 更小的图片，减轻了 diffusion 生成的计算复杂度。

本文的缺点在于，生成的 video 在帧之间仍然具有较大的抖动，视频的连续型仍然不能够被很好地保证。同时，实际上在视频当中，是存在很多关键帧的，而关键帧的密度和视频中的画面变化幅度强相关。比如一个打斗场面的关键帧密度比一个静止场景的关键帧高很多。但是本篇工作中，video 所有的帧都是以关键帧的形式生成的，并不会根据画面进行侧重的生成。

以下是一些 insight 和 video diffusion 一些未来可能的方向。

首先，传统视频的帧补全方法可以生成连续的帧，而这能很好地弥补现有 video generation model 生成视频不连续的问题，是否可以将传统算法融合进来辅助模型的生成呢？

其次，现有 video 生成方法是基于 image 的，但是一种对于 video 更好的压缩方式是进行层次化压缩，比如说首先将相邻的五帧进行第一层压缩成一个 token，第二层压缩可以使一个 token 代表五秒的视频，依此类推，以 transformer 为架构进行层次化压缩，上述过程的逆过程可能可以带来更好的 video generation。但是，对于这个方法，video 的数据集是一个瓶颈，现在缺少带有高质量 caption 的 video 数据集，不确定现有 video 数据集是否能够训出来这样的层次式的模型。

video generation 需要很强的时序信息支撑，而这一点 transformer 比 diffusion 更擅长，diffusion 比 transformer 更擅长 image 生成。可以考虑将 transformer 和 diffusion 进行结合。具体来说，用 transformer 生成 image token，再让 diffusion conditioned on image token 进行 image 解码，最后用超分模型对 image 进行超分。这个 pipeline 需要让 transformer 的 image token 和 diffusion 的 image token 进行对齐，这可能是一个实现的难点。

Diffusion的 text to image 的生成中涉及到很多的信息补全工作，因为一个 text 只能描述 image 中的一小部分细节，而大部分细节都需要模型来补全，这就导致了一个 text 必然会对应着很多个 image，而 diffusion 是通过 random sample 来从众多 image 中选择一个对应的最好的，所以一个可能的思路是找到一种 random token 来描述这种细节填充的方法，这样能够更好地生成相似风格的图片，这种 random token 应该能有非常大的作用。