Paper Reading 9

沙之洲 2020012408

**Align your latents:**

1 本文中实现 video 连续性的方式是，选择当前帧前边的 S 个帧作为 condition 指导后边帧的生成。但是实际上，可能更远处的帧保存的信息会对当前帧的生成起到作用。所以这种简单地向前选择 S 个帧的方法是否不太符合逻辑？是否可以考虑采取 attention 的方式，从前边选取对当前帧生成而言的重要的帧？

2 用 diffusion model 做 video 生成面临的一个很大的问题是 computational heavy，这对于 scaling 到小时级别的 video 是极大的阻碍，虽然本文中提出用 latent diffusion model 能够缓解这个问题，但是仍然不能 scale 到小时级别。该如何解决 diffusion model 在 video 生成上的 computational heavy 的问题呢？

**Scaling Autoregressive Models for Content-Rich Text-to-Image Generation**

1 latent diffusion model 和 transformer 在 text to image 过程中，都是先使用 VAE based model 把图片放到 latent space。从模型结构上而言，两者对于 latent vector 的处理，各有什么优劣？

2 作出更 fidelity 的图片究竟是否需要 world knowledge？先验的 world knowledge 究竟在 image generation 中起到了什么样的作用？应该怎样去衡量它？