MS-S2V

　　这篇文章要解决的是主体一致的视频生成。基于T2V模型，将主体信息作为reference image经过vae编码后与噪声seq拼接后一起送入DiT，通过self-attention进行信息交互。同时，本文提出了TS-RoPE，来解决主体信息与视频信息之间的位置关系问题，使得video和reference image中的subject和reference image保持一致。这里的subject指的是reference image中的主体，在训练的时候，通过随机采样的方式，从视频中采样出主体信息，作为reference image的输入。这样，就可以在推理的时候，通过随机采样的方式，从视频中采样出主体信息，作为reference image的输入，使得video和reference image。

　　下面这张图是pipeline的示意

2026-3-5	2026-3-4
WildActor	OmniTransfer