Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

LDMs: text-to-image pretraining, video pretraining, and high-quality video finetuning. Furthermore, we demonstrate the necessity of a well-curated pretraining dataset for generating high-quality videos and present a systematic curation process to train a strong base model. 3D-prior and can serve as a base to finetune a multi-view diffusion model that jointly generates multiple views of objects in a feedforward fashion, outperforming image-based methods at a fraction of their compute budget.

Latent Video Diffusion Models Video-LDMs train the main generative model in a latent space of reduced computational complexity . Most related works make use of a pretrained text-to-image model and insert temporal mixing layers of various forms into the pretrained architecture. Ge et al. additionally relies on temporally correlated noise to increase temporal consistency and ease the learning task. In this work, we follow the architecture proposed in Blattmann et al. and insert temporal convolution and attention layers after every spatial convolution and attention layer. In contrast to works that only train temporal layers are completely training-free we finetune the full model. For textto-video synthesis in particular, most works directly condition the model on a text prompt or make use of an additional text-to-image prior In our work, we follow the former approach and show that the resulting model is a strong general motion prior, which can easily be finetuned into an image-to-video or multi-view synthesis model. Additionally, we introduce micro-conditioning on frame rate. We also employ the EDM-framework and significantly shift the noise schedule towards higher noise values, which we find to be essential for high-resolution finetuning.

Stage I: Image Pretaining

Stage II: Curating a Video Pretraining Dataset

A systematic approach to video data curation. For multimodal image modeling, data curation is a key element of many powerful discriminative

Stage III: High-Quality Finetuning

Training Video Models at Scale

Pretrained Base Model

High-Resolution Text-to-Video Model

High Resolution Image-to-Video Model

Camera Motion LoRA