视频生成的两种范式

视频生成范式	基于SD逐帧生成	基于时空Patches生成
最新模型	Pika Labs, Runway Gen-2, I2VGEN-XL, DynamiCrafter, Stable Video Diffusion (SVD)	Sora, Video Diffusion Transformer (VDT)
模型优点	以SD作为初始化，模型更容易训练，训练成本可控	视频内容的一致性有保证，可以生成长视频
模型缺点	视频内容的一致性要差一些，长视频生成有困难	整个模型需要从头训练，训练成本很高

阿里I2VGEN-XL

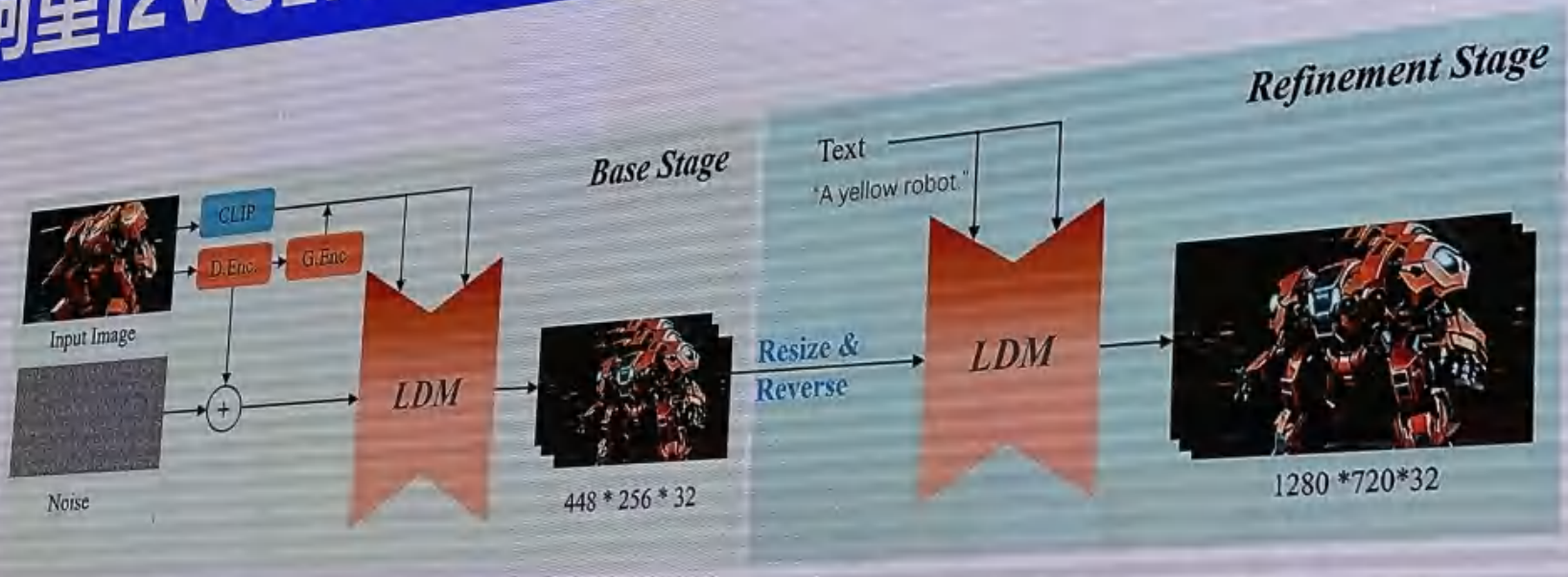


Figure 2. The overall framework of I2VGen-XL. In the *base stage*, two hierarchical encoders are employed to simultaneously capture high-level semantics and low-level details of input images, ensuring more realistic dynamics while preserving the content and structure of images. In the *refinement stage*, a separate diffusion model is utilized to enhance the resolution and significantly improve the temporal continuity of videos by refining details. "D.Enc." and "G.Enc." denote the detail encoder and global encoder, respectively.

代码和模型开源：https://github.com/ali-vilab/i2vgen-xl

腾讯DynamiCrafter

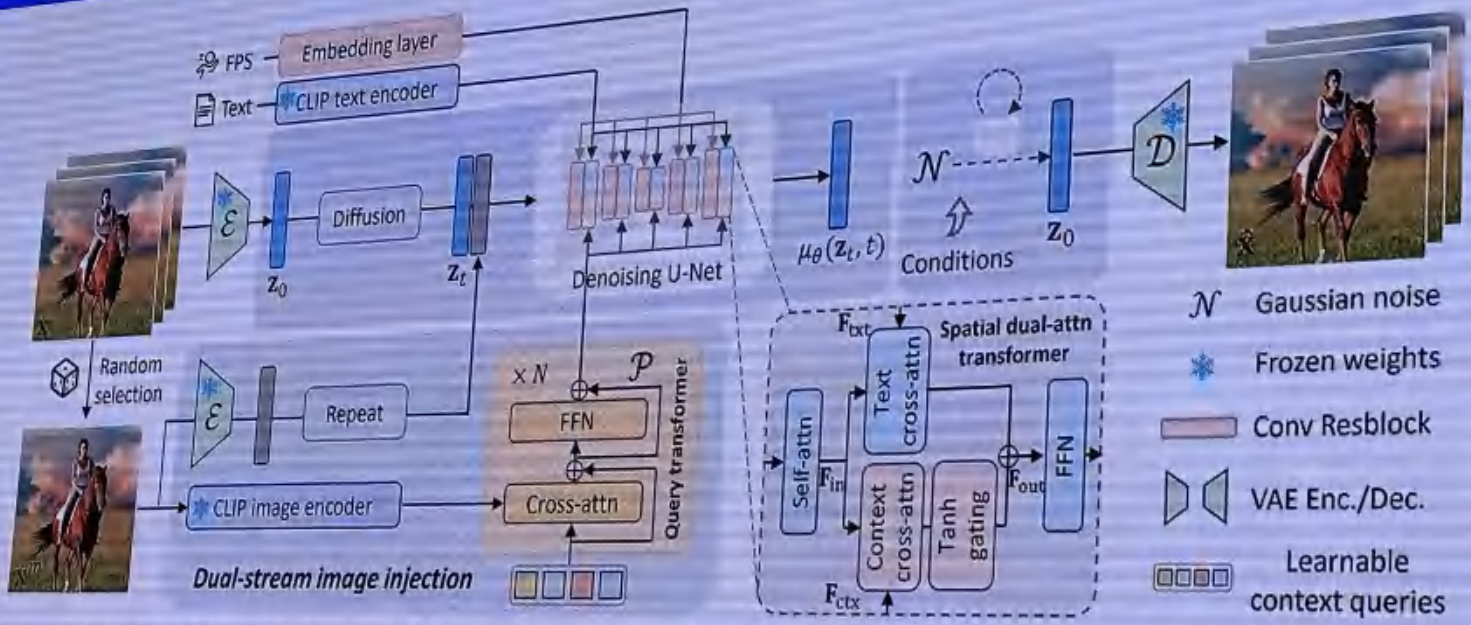


Figure 1. Flowchart of the proposed *DynamiCrafter*. During training, we randomly select a video frame as the image condition of the denoising process through the proposed dual-stream image injection mechanism to inherit visual details and digest the input image in a context-aware manner. During inference, our model can generate animation clips from noise conditioned on the input still image.

代码和模型开源：https://github.com/Doubiiu/DynamiCrafter

Stable Video Diffusion

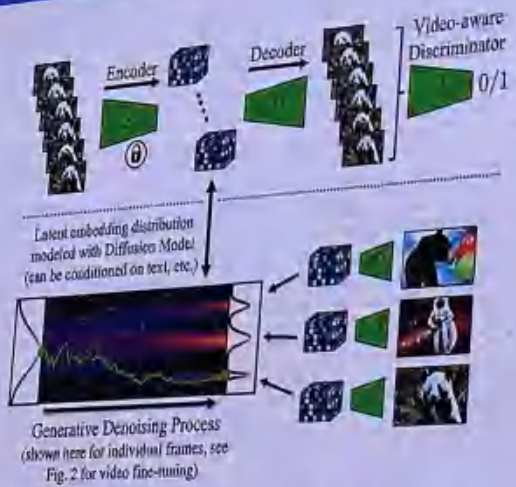


Figure 3. *Top:* During temporal decoder fine-tuning, we process video sequences with a frozen encoder, which processes frames independently, and enforce temporally coherent reconstructions across frames. We additionally employ a video-aware discriminator. *Bottom:* in LDMs, a diffusion model is trained in latent space. It synthesizes latent features, which are then transformed through the decoder into images. Note that the bottom visualization is for individual frames; see Fig. 2 for the video fine-tuning framework that generates temporally consistent frame sequences.

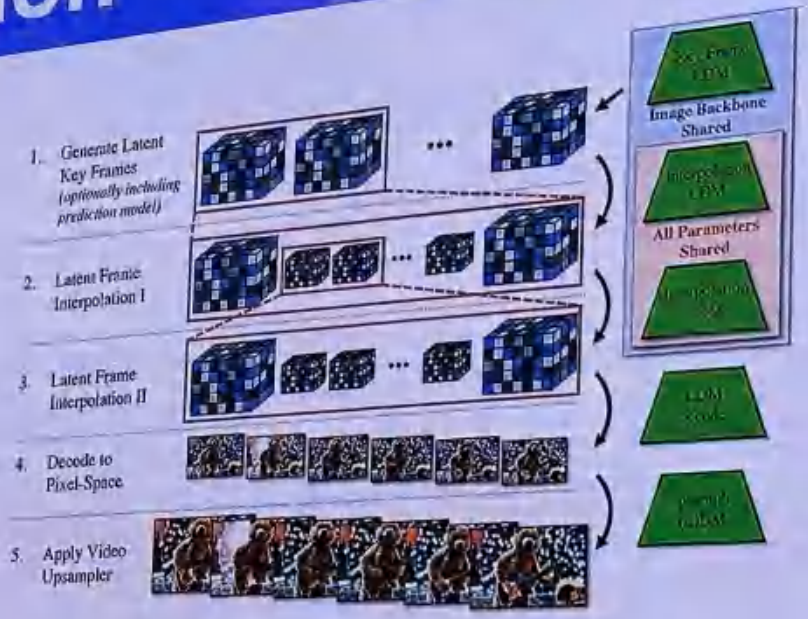
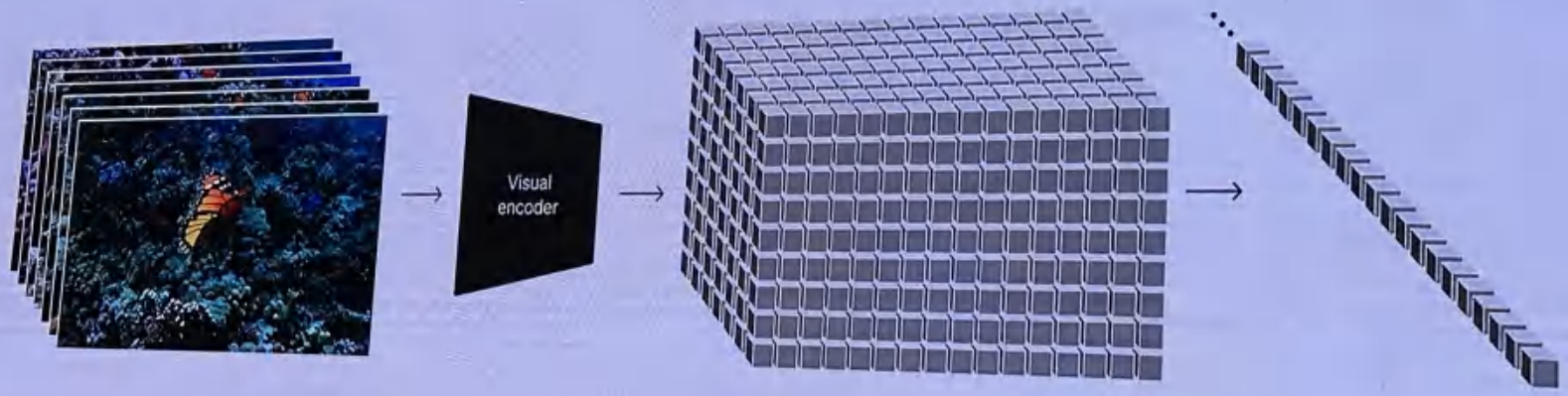


Figure 5. **Video LDM Stack.** We first generate sparse key frames. Then we temporally interpolate in two steps with the same interpolation model to achieve high frame rates. These operations are all based on latent diffusion models (LDMs) that share the same image backbone. Finally, the latent video is decoded to pixel space and optionally a video upsampler diffusion model is applied.

代码和模型开源：https://huggingface.co/stabilityai/stable-video-diffusion-img2vid

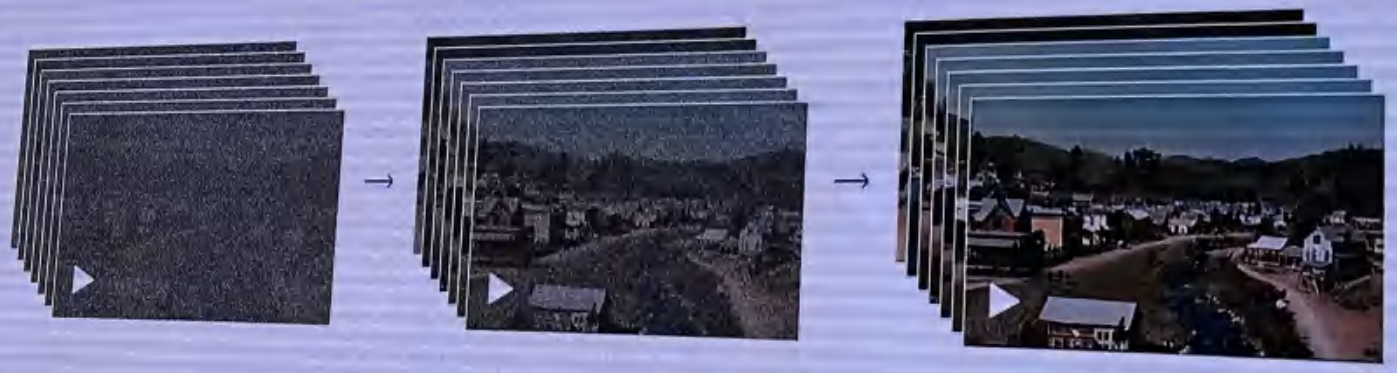
OpenAI Sora

- Sora将视频表示为时空patches：
  - 训练一个VQ-VAE网络，将视频压缩到较低维度的潜在空间
  - 将视频的潜在表示分解为时空patches，作为Transformer tokens



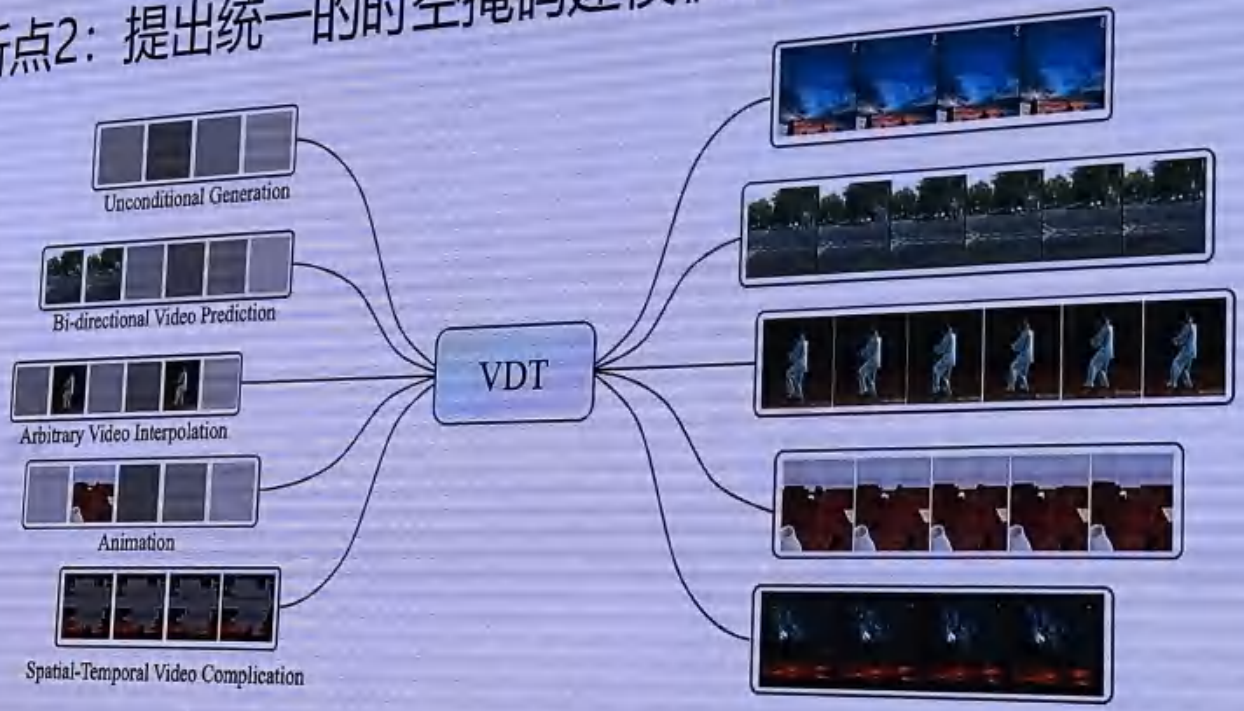
OpenAI Sora

- Sora是一个扩散模型，它接受输入的噪声patches（以及如文本提示等条件），然后被训练去预测原始的“干净”patches
- 重要的是，Sora采用Diffusion Transformer结构，这种模型已经在多个领域展现了显著的扩展性



VDT - 类Sora底座模型

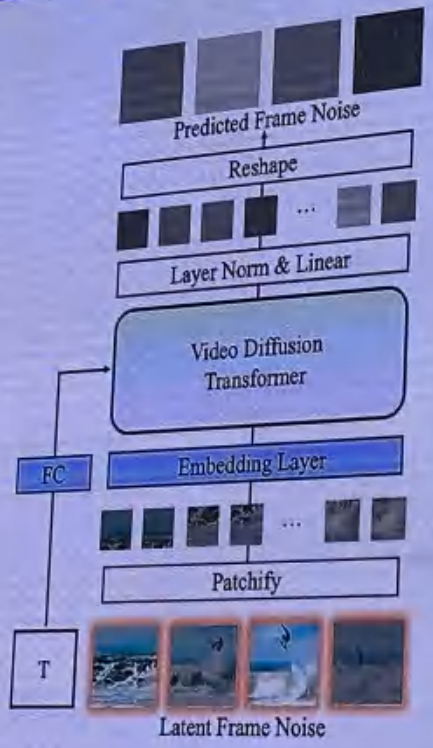
- 创新点1：将Transformer技术应用于基于扩散的视频生成
- 创新点2：提出统一的时空掩码建模机制



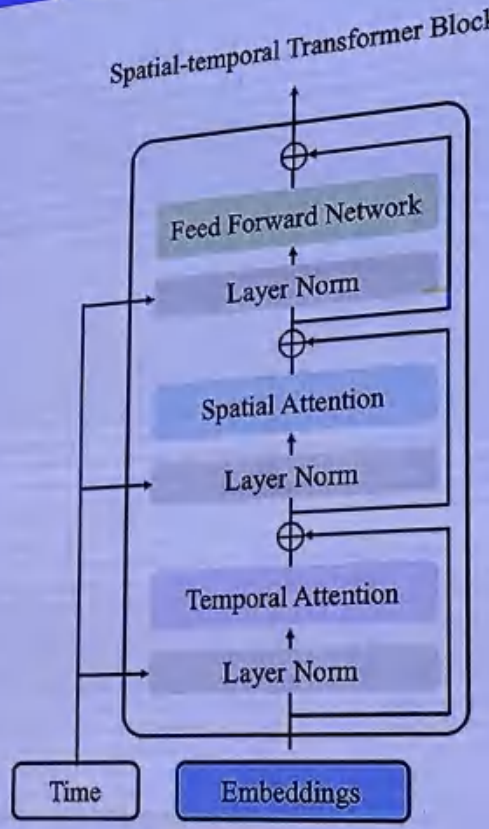
* Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu*, Ping Luo, and Mingyu Ding, VDT: General-purpose Video Diffusion Transformers via Mask Modeling, International Conference on Learning Representations (ICLR), 2024.

VDT - 类Sora底座模型

(a) 输入/输出视频压缩，表示为时空Tokens



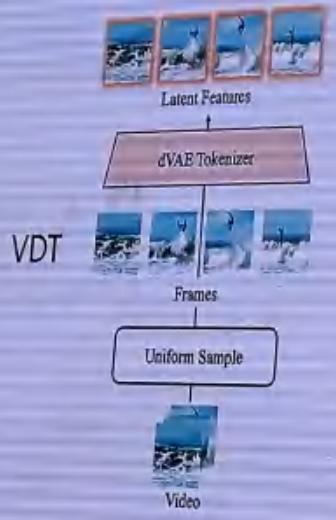
(b) 整体架构Video Diffusion Transformer (VDT)

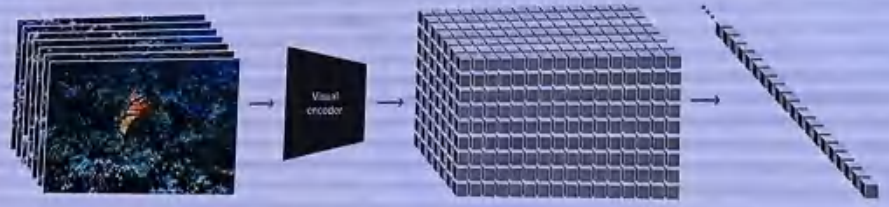


(c) 时空Transformer Block，考虑时空注意力

VDT - 与SORA的区别

- VDT采用的是在时空维度上分别进行注意力机制处理的方法
  - 分离注意力的做法在视频领域比较常见，是在显存限制下的trade-off
- Sora将时间和空间维度合并，通过统一时空注意力机制来处理
  - Sora强大的视频动态能力可能来自于时空整体的注意力机制





Sora

视频生成的发展趋势

- **视频生成加速**：为了更好的落地，视频生成的推理算法需要不断优化，单个视频生成耗时越少越好，对计算资源的占用越少越好
- **超长视频生成**：目前最好的Sora模型只能生成60秒的视频，如何生成超过60秒的超长视频值得在未来一年内重点关注
- **视频可控生成**：不同于图像可控生成，视频可控生成需要额外考虑精细运镜、复杂角色动作等要素，面临巨大的挑战