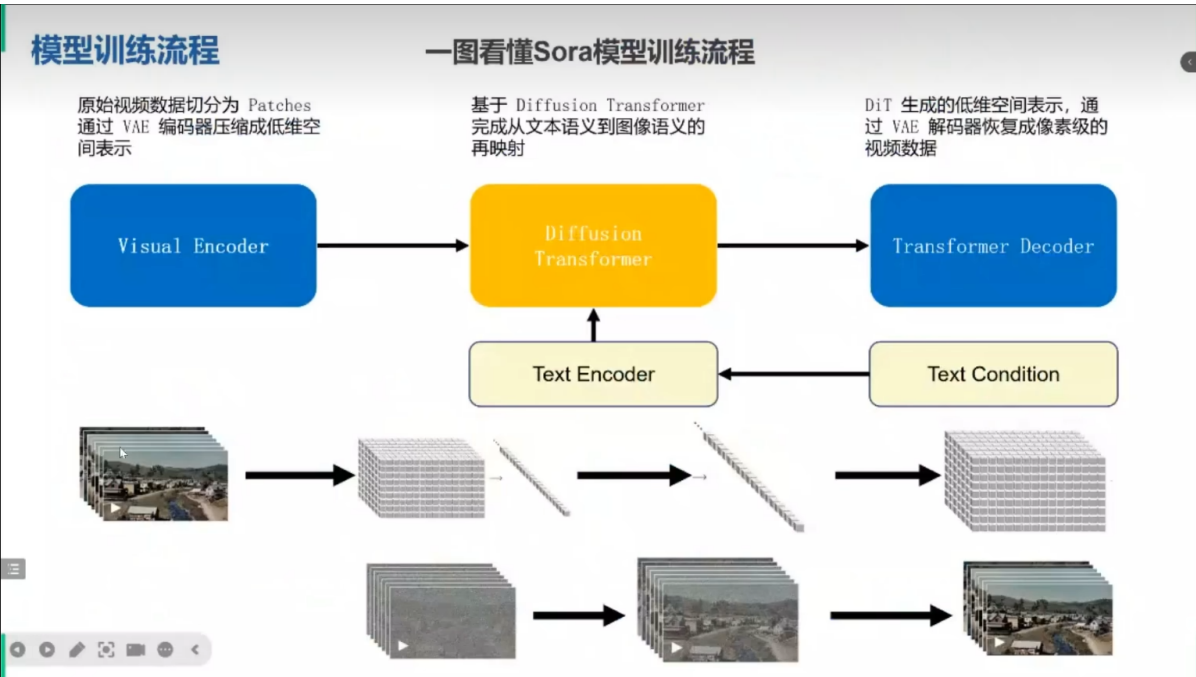


Sora技术路线详解

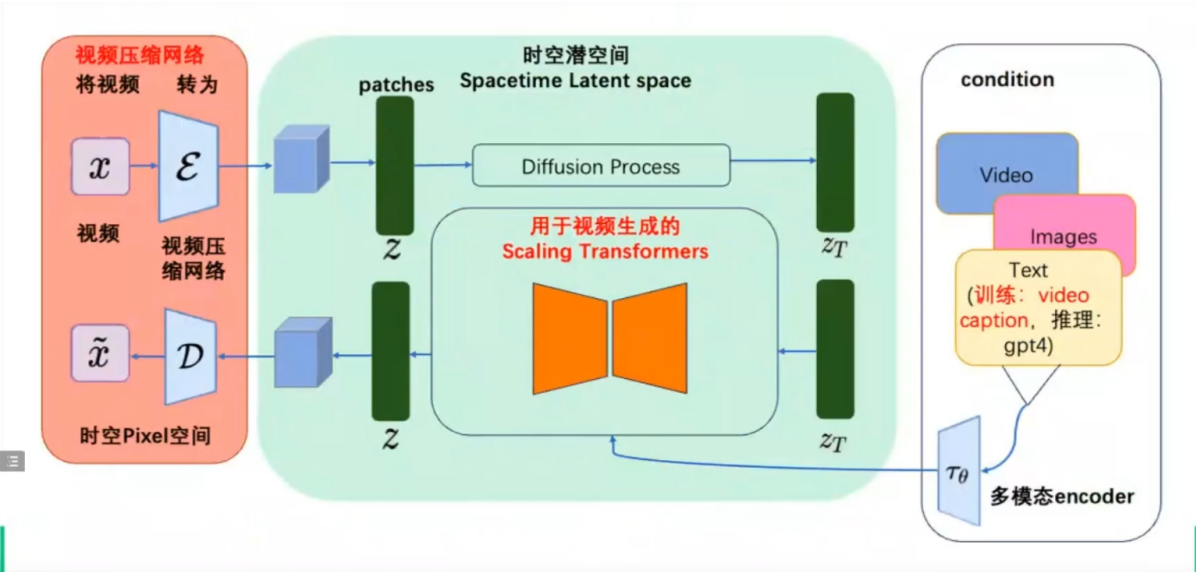
Sora模型不仅是一个视频生成模型，也算是一个世界模拟器。

模型训练流程



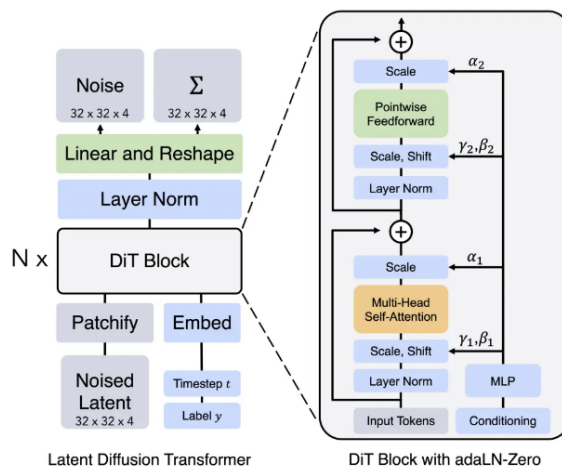
Sora是一个在不同时长、分辨率和宽高比的视频及图像上训练生成的**Diffusion模型**，同时采用了**transformer架构**。

其中Diffusion模型是基于非平衡热力学概念，通过马尔可夫链构建数据样本，示意图如下：

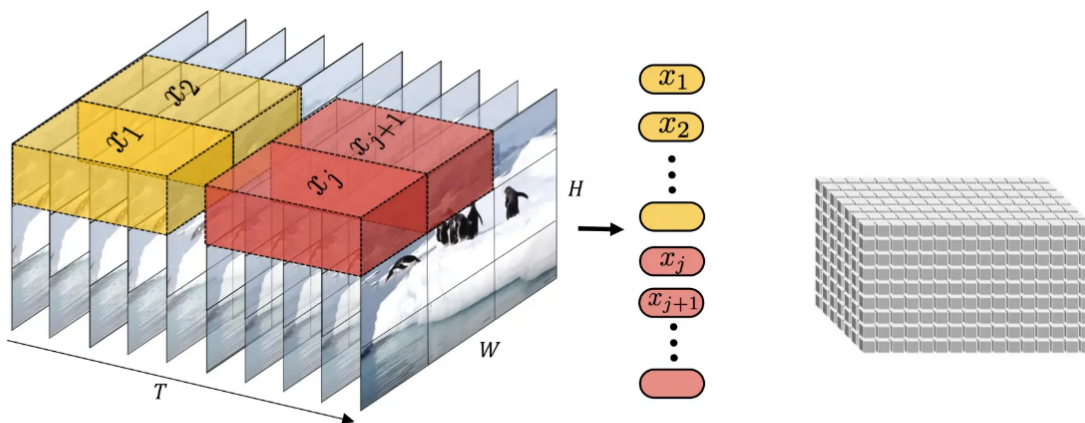


网络结构: Diffusion Transformer, DiT

- DiT 首先将每个 patch 空间表示 Latent 输入到第一层网络, 以此将空间输入转换为 tokens 序列。
- 将标准基于 ViT 的 Patch 和 Position Embedding 应用于所有输入 token, 最后将输入 token 由 Transformer 处理。
- DiT 还会处理额外信息, e.g. 时间步长、类别标签、文本语义等。



含时空编码的样本取样



将输入的视频划分为若干tuple, 每个tuple会变成token

经过Spatial Temporal Attention进行空间/时间建模获得有效的视频表征token, 即上图灰色block

Sora内部通过图上的切块法, 使得采样能包含时间和空间信息。

SORA 支持不同长度、不同分辨率的输入

NaViT: 多个patches打包成一个单一序列实现可变分辨率



使用不同分辨率、不同时长的视频进行训练
保证推理时在不同长度和分辨率上的效果

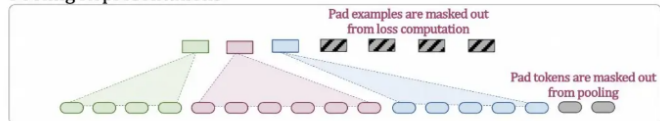


带来大量的计算负载不均衡

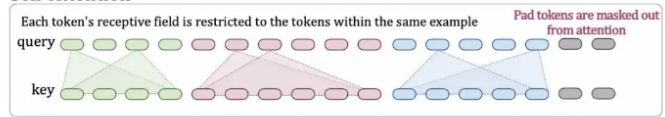


可能使用Google的NaViT相关技术降低计算量
支持动态输入

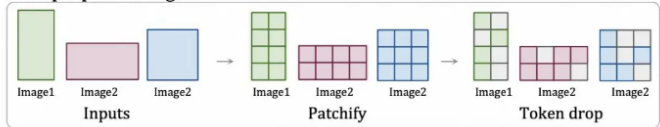
Pooling Representations



Self-Attention



Data preprocessing



Dehghani, Mostafa, et al.

Patch n 'Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution.
arXiv preprint arXiv:2307.06304 (2023)

同时使用了NaViT技术，可以降低计算量并支持动态输入。

(看的不是很懂，后面还会更新)