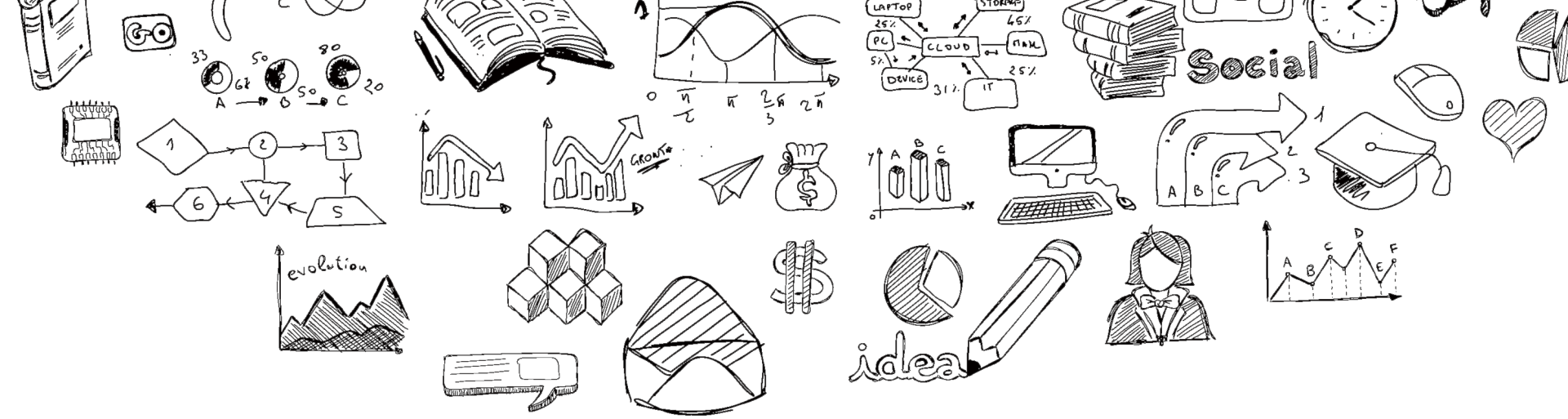


Text to Video Generation

汇报人：吴平

目录

Part 1

任务简介

任务简介

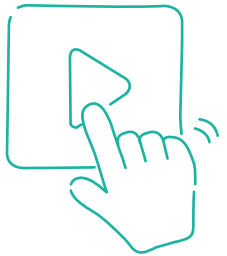
A bunch of autumn leaves falling on a calm lake to form the text 'Imagen Video'. Smooth.

A cat eating food out of a bowl, in style of van Gogh.

Balloon full of water exploding in extreme slow motion.

Part 2

自回归模型

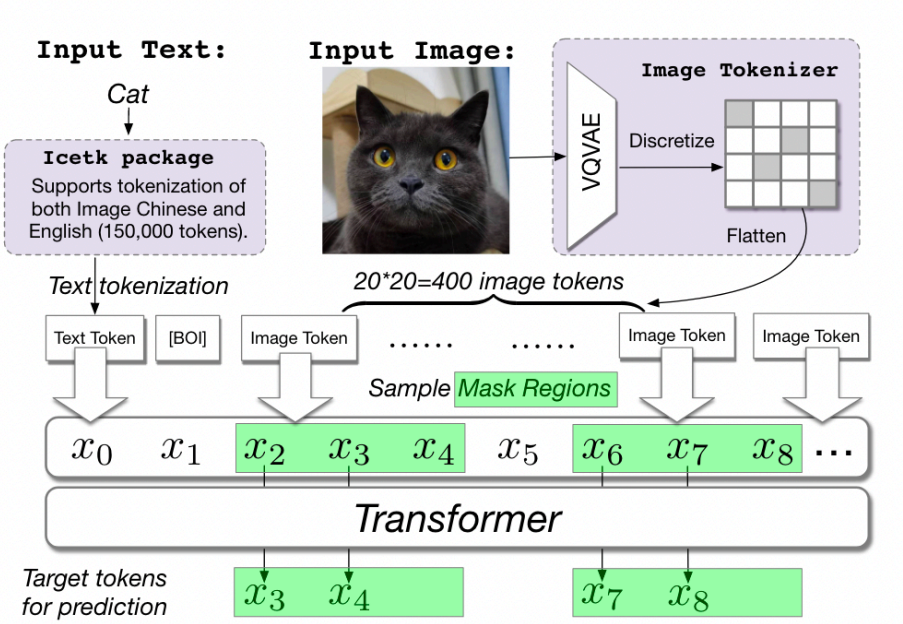


自回归模型

CogLM: 将文本和图像都编码为离散的token，在此基础上训练跨模态的语言模型。

ICETK:

130000 文本token
20000 图像token

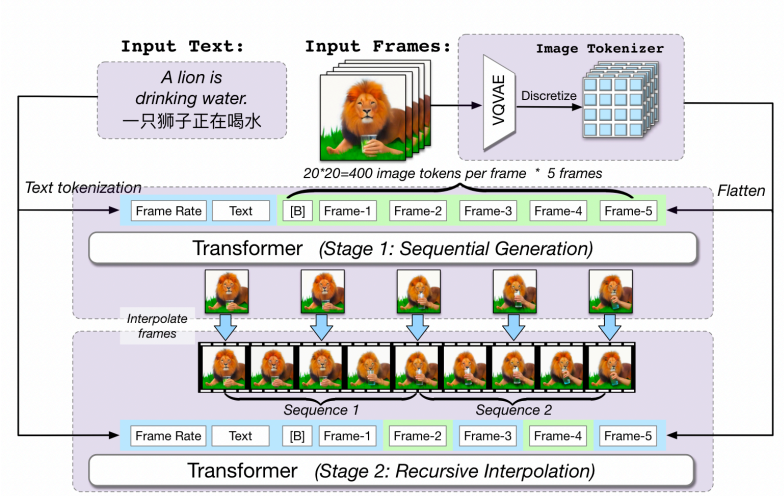


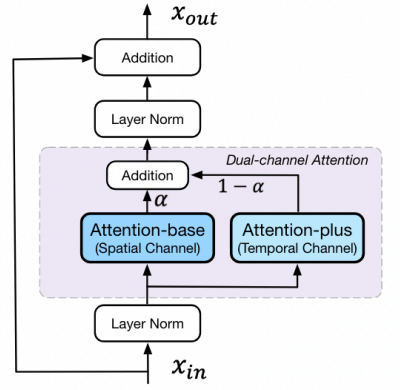
CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. arXiv 2022

自回归模型

CogVideo：

- 序列生成：将视频的图像token进行拼接，并使用CogLM进行预测
- 递归插帧：在已经预测出的图像间插入未知帧，进行未知帧的补全，实现视频的增长





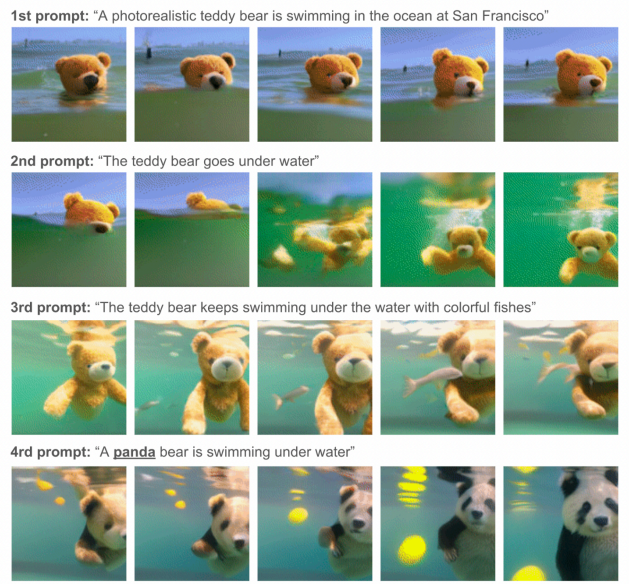
Dual channel attention：

为利用图像数据增强视频模型的效果，CogVideo的模型架构直接继承自CogLM，仅新加入了一个注意力模块，继承自CogLM的模块在训练的过程中不更新参数。

CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. arXiv 2022

自回归模型

Phenaki：使用因果注意力机制，基于多条prompt生成任意长度的视频



Prompts used:

Lots of traffic in futuristic city. An alien spaceship arrives to the futuristic city. The camera gets inside the alien spaceship. The camera moves forward until showing an astronaut in the blue room. The astronaut is typing in the keyboard. The camera moves away from the astronaut. The astronaut leaves the keyboard and walks to the left. The astronaut leaves the keyboard and walks away. The camera moves beyond the astronaut and looks at the screen. The screen behind the astronaut displays fish swimming in the sea. Crash zoom into the blue fish. We follow the blue fish as it swims in the dark ocean. The camera points up to the sky through the water. The ocean and the coastline of a futuristic city. Crash zoom towards a futuristic skyscraper. The camera zooms into one of the many windows. We are in an office room with empty desks. A lion runs on top of the office desks. The camera zooms into the lion's face, inside the office. Zoom out to the lion wearing a dark suit in an office room. The lion wearing looks at the camera and smiles. The camera zooms out slowly to the skyscraper exterior. Timelapse of sunset in the modern city

Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. arXiv 2022

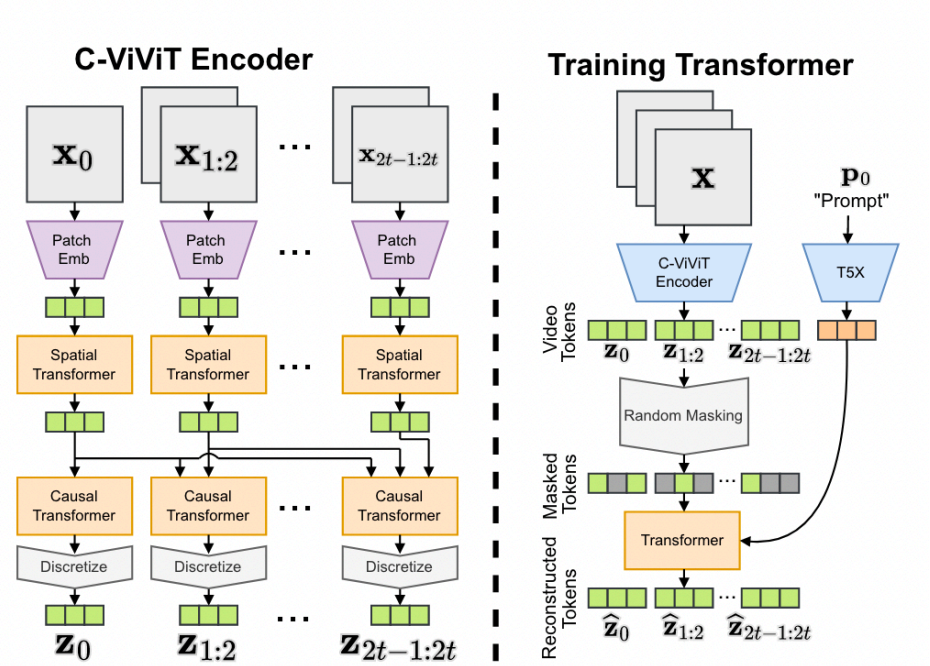
自回归模型

C-ViViT编码器：

- 将视频序列切分为patch序列，每个patch的维度是t * c * h * w
- 使用多层的空间-时间transformers对patch表示进行处理
- Transformers的输出被量化为码表中的向量

序列预测模型：

- 数据中的文本嵌入作为序列预测模型的上下文
- 训练过程中，采取随机mask的策略，利用双向transformer进行序列预测

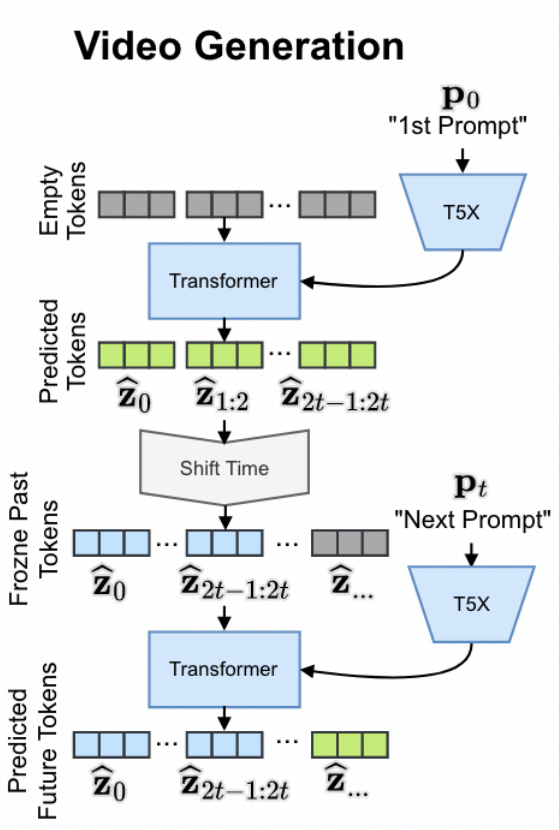


Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. arXiv 2022

自回归模型

视频生成：

采用移动窗口式的视频生成，将前一段的视频尾部帧的token作为下一段视频的开头。



Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. arXiv 2022

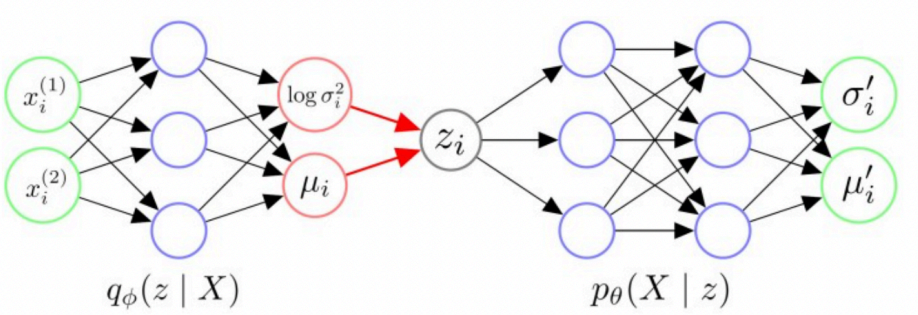
Part 3

扩散模型

扩散模型

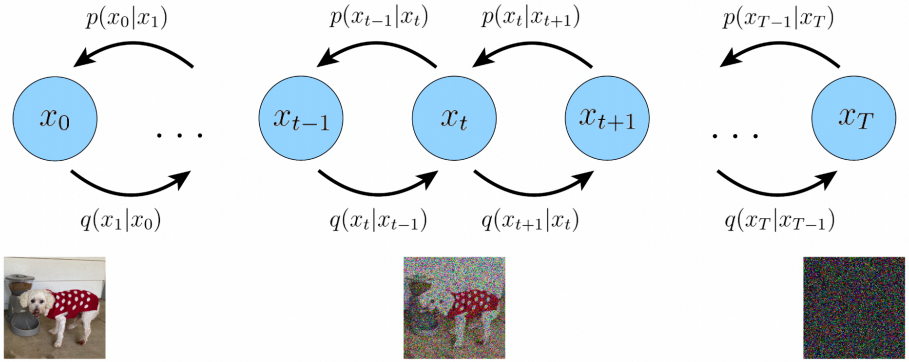
VAE：

将源数据映射到已知分布的隐变量上，通过对隐变量的随机采样，从而实现对源数据分布的采样。



扩散模型：

将VAE中一步生成样本的方式改为多步生成，并且限定中间的隐变量拥有和源数据一样的维度，最终生成的隐变量仍然满足高斯分布。



Understanding Diffusion Models: A Unified Perspective. arXiv 2022

扩散模型

VAE损失函数：

- 重构损失：解码器重构的图像与输入图像之间的距离
- 先验匹配损失：隐变量分布与正态分布之间的距离

$$
\begin{aligned}
\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{prior matching term}}
\end{aligned}
$$

扩散模型损失函数：

- 重构损失：从x1重构到x0之间的距离
- 先验匹配损失：最终隐变量分布与正态分布之间的距离
- 去噪损失：从x2到xT重构到x0之间的距离

$$
\log p(\boldsymbol{x}) \geq \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}
$$

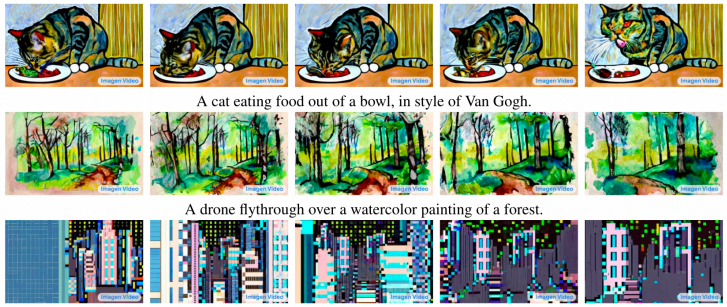
Understanding Diffusion Models: A Unified Perspective. arXiv 2022

扩散模型

Imagen Video：

- 生成高分辨率、长度可观的视频
- 可以生成各种文本指定的艺术风格
- 视频中可以体现出物体的3D结构
- 可以生成各种艺术文字视频

Figure 1: Imagen Video sample for the prompt: "*A bunch of autumn leaves falling on a calm lake to form the text 'Imagen Video'. Smooth.*" The generated video is at 1280×768 resolution, 5.3 second duration and 24 frames per second.



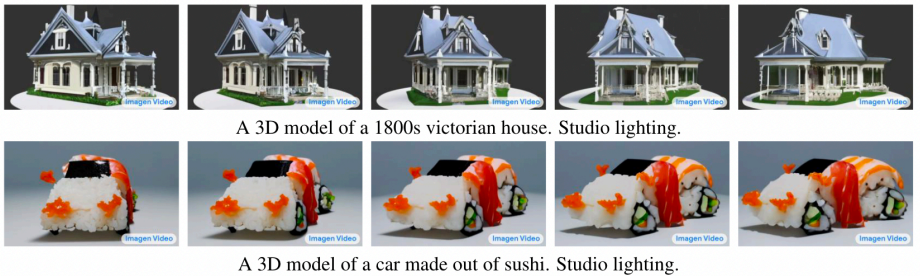
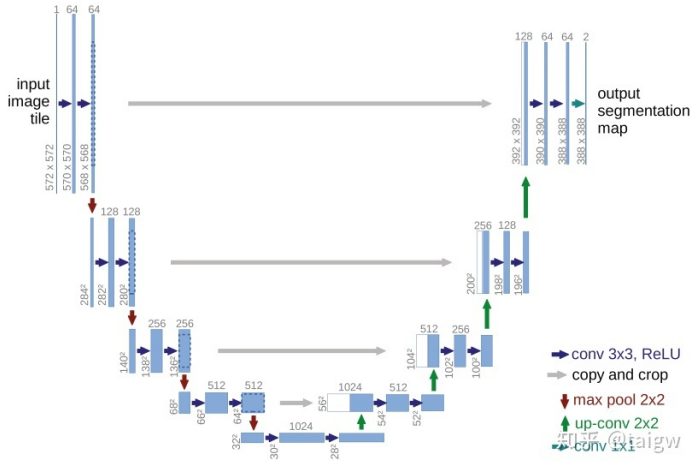
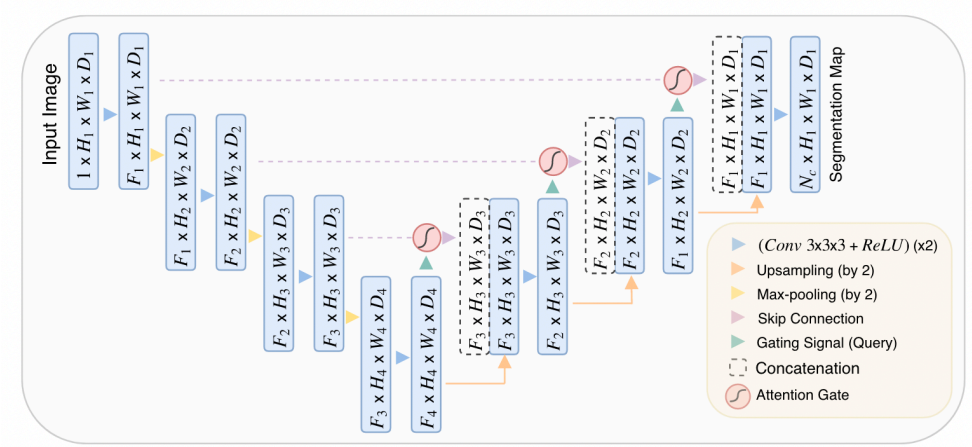
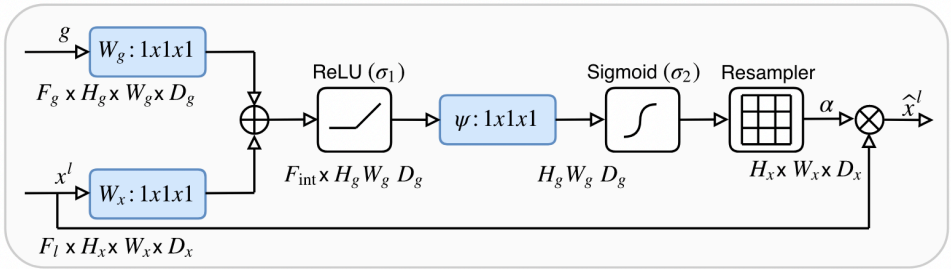


Imagen Video: High Definition Video Generation with Diffusion Models. arXiv 2022

扩散模型







文本-视频生成模型延续了之前工作中使用的Attention Unet架构。

Imagen Video: High Definition Video Generation with Diffusion Models. arXiv 2022

扩散模型

Imagen Video：采用了适用于视频数据的卷积与注意力机制。

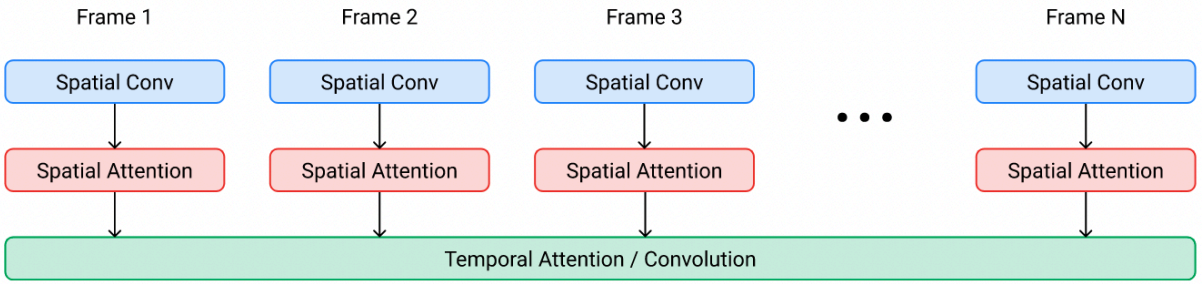


Figure 7: Video U-Net space-time separable block. Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Our base model uses spatial convolutions, spatial self-attention and temporal self-attention. For memory efficiency, our spatial and temporal super-resolution models use temporal convolutions instead of attention, and our models at the highest spatial resolution do not have spatial attention.

Imagen Video: High Definition Video Generation with Diffusion Models. arXiv 2022

扩散模型

使用级联模型来完成视频的生成：

- 文本-视频模型的输出是16x40x24的视频
- 使用多个超分辨率模型和插帧模型
- 最终得到128x1280x768的视频

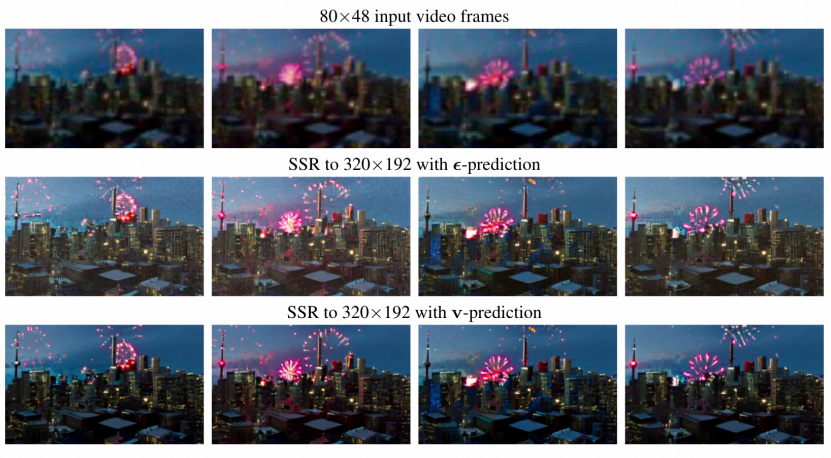


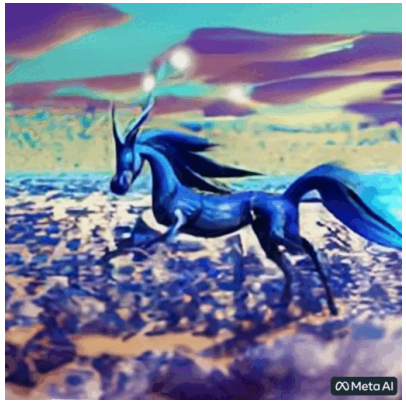
Figure 6: The cascaded sampling pipeline starting from a text prompt input to generating a 5.3-second, 1280×768 video at 24fps. "SSR" and "TSR" denote spatial and temporal super-resolution respectively, and videos are labeled as frames×width×height. In practice, the text embeddings are injected into all models, not just the base model.

所有的模型可以同时进行训练，并且SSR模型和TSR模型是通用模型，可以用在其它任务上。

Imagen Video: High Definition Video Generation with Diffusion Models. arXiv 2022

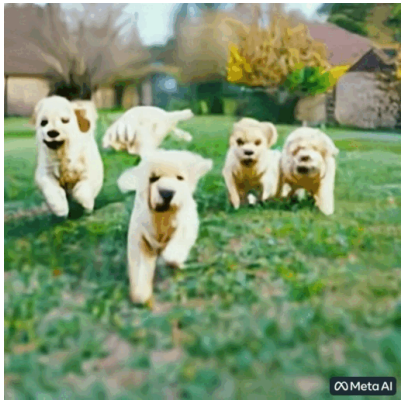
扩散模型

Make-A-Video：使用不含文本-视频对的数据训练文本-视频生成模型。



A blue unicorn flying over a mystical land.

A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up, highly detailed, studio lighting, screen reflecting in its eyes.



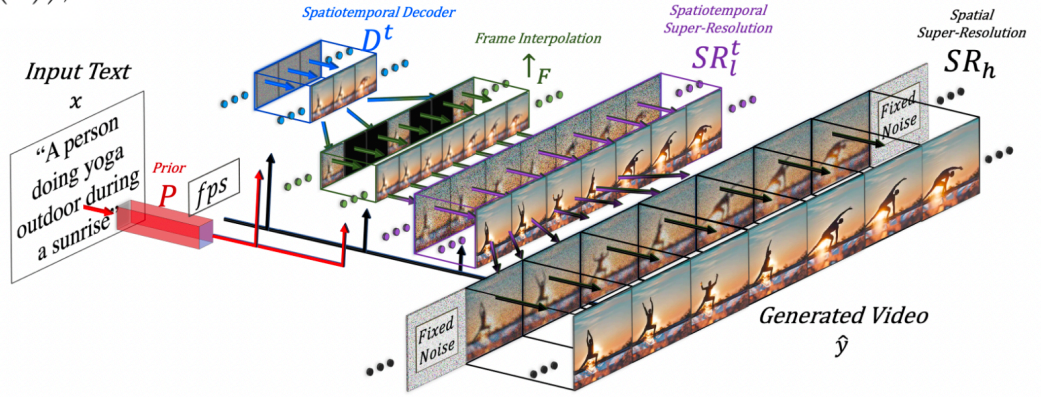
A litter of puppies running through the yard.

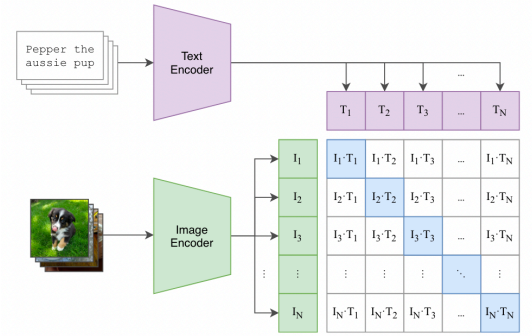
Make-A-Video: Text-to-Video Generation Without Text-Video Data. arXiv 2022

扩散模型

$$\hat{y}_t = \mathrm{SR}_h \circ \mathrm{SR}_l^t \circ \uparrow_F \circ \mathrm{D}^t \circ \mathrm{P} \circ (\hat{x}, \mathrm{C}_x(x)),$$

x： 文本输入
C_x： CLIP文本编码器
P： 先验计算模型
D^t： 时空编码器
$\uparrow_F$： 插帧模型
SR： 超分辨率模型





CLIP：

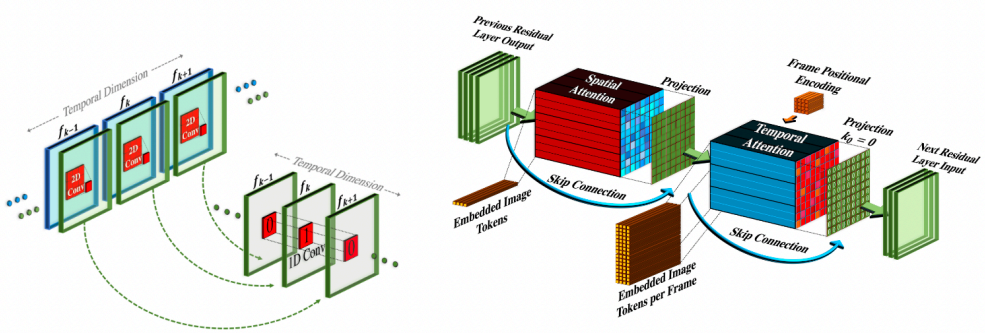
- 文本编码器
- 图像编码器

Make-A-Video: Text-to-Video Generation Without Text-Video Data. arXiv 2022

扩散模型

文本-图像（视频）模型：

- 伪3D卷积：在空间卷积层的末尾加入时间维度的卷积层，初始化为恒等函数。
- 伪3D注意力：在空间注意力机制后加入时间注意力机制，初始化为恒等函数。



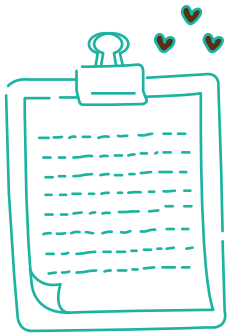
训练过程：

使用文本-图像数据训练网络中的空间卷积、注意力模块，然后加入时间卷积、注意力模块，进行初始化，使用无标注的视频数据进行训练。

Make-A-Video: Text-to-Video Generation Without Text-Video Data. arXiv 2022

Part 4

常用评价指标



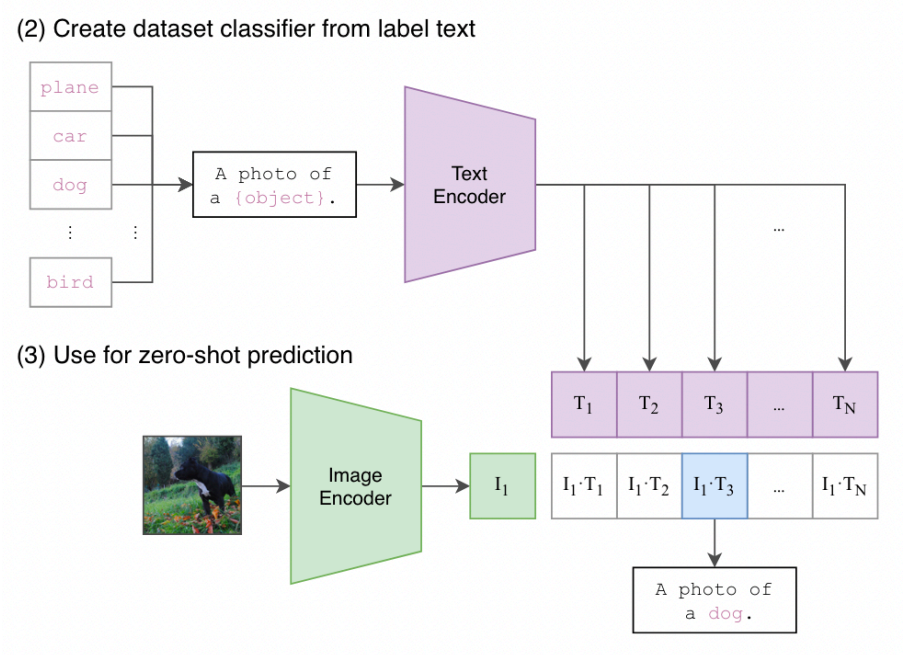
评价指标

沿用文本-图像生成指标：

Inception Score（IS）

$$\mathbf{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}\left(p(y|\mathbf{x})||p(y)\right)\right)$$

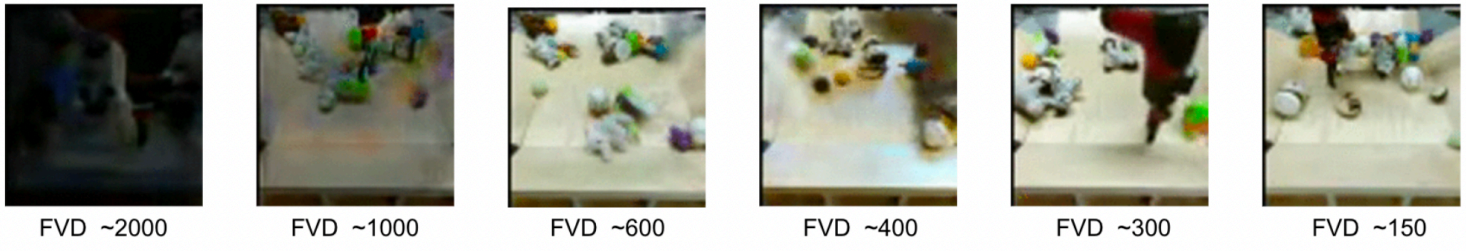
CLIP Score



评价指标

FVD（Fr´echet Video Distance）：

- 视频数据经过合适的特征提取后，呈现近似的高斯分布
- 使用Fr´echet Distance来计算两个高斯分布之间的距离

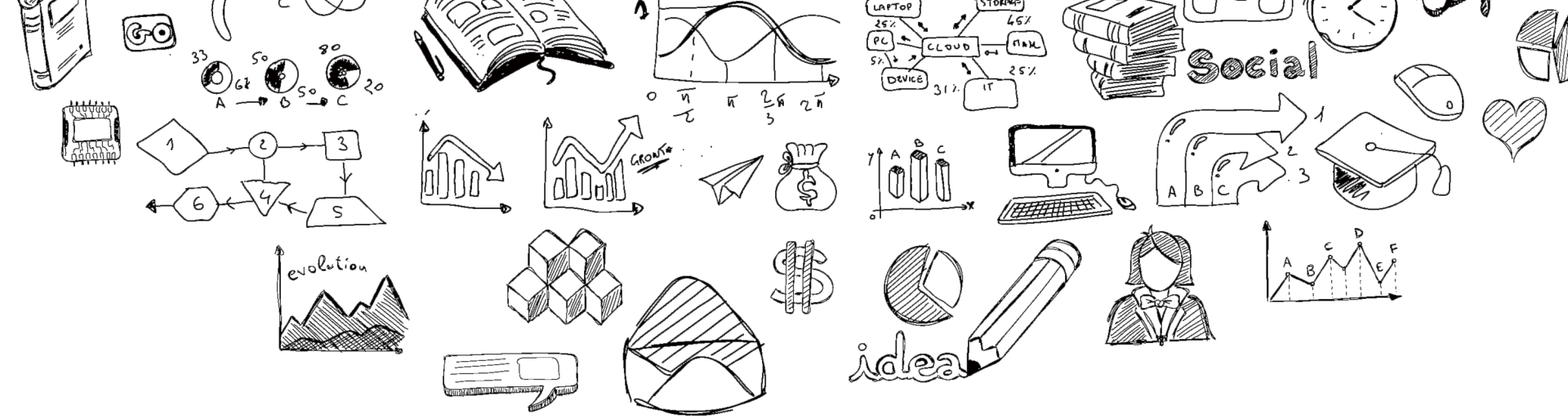


I3D：

在Kinetics数据集上预训练一个视频分类模型，用于视频特征的提取。

YouTube

Towards Accurate Generative Models of Video: A New Metric & Challenges. arXiv 2018



汇报完毕 谢谢您的观看

汇报人：吴平