



类SORA的多模态视觉文本大模型的文 生视频技术

范晨悠

2024秋季·公选课

华南师范大学人工智能学院 fanchenyou@scnu.edu.cn

“What I cannot create, I do not understand.”

Richard Feynman

A man skiing on the moon

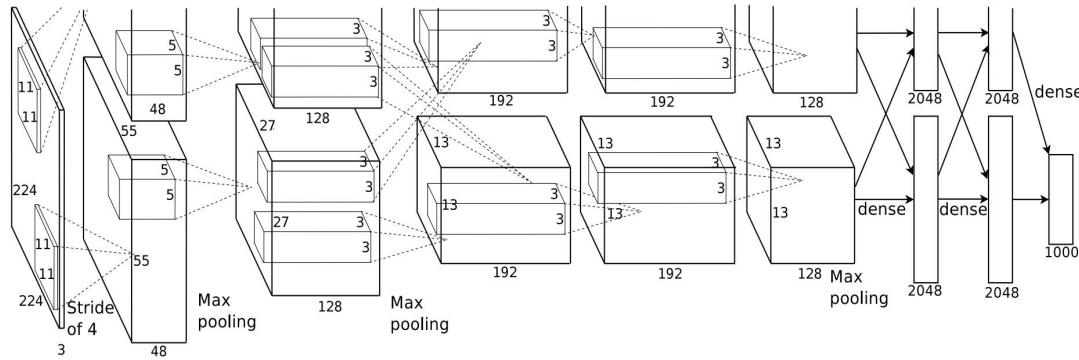


可灵

AI发展里程碑

AI发展里程碑 (2014年)

现代卷积神经网络的开创性工作，在2块GPU上，使用120万张图片进行训练，超出传统方法10%的准确率。

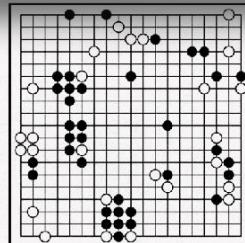


ImageNet Classification with Deep Convolutional Neural Networks



AI发展里程碑 (2016年)

围棋AI模型AlphaGo/Zero，
击败人类世界冠军并登上
N&S 杂志封面。



深度学习三巨头获图灵奖 (2018年)

“They led significant breakthroughs in AI technologies ...”

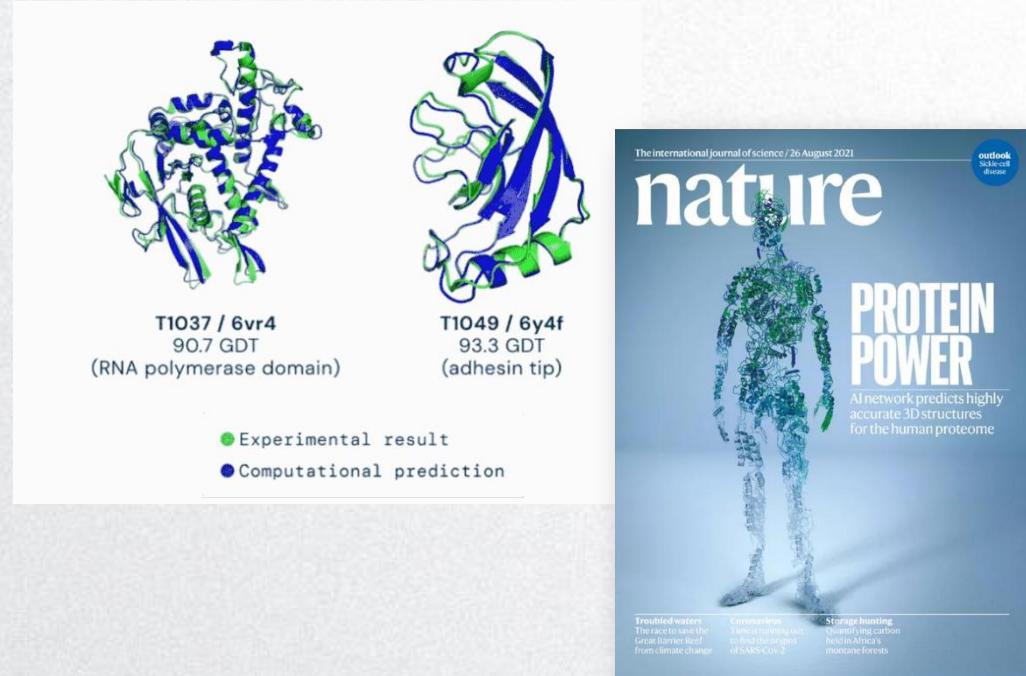
- 约书亚·本吉奥(Yoshua Bengio)
 - 加拿大蒙特利尔大学
- 杰弗里·辛顿(Geoffrey Hinton)
 - 多伦多大学, Google Brain
- 杨乐昆(Yann LeCun)
 - 纽约大学, Meta首席科学家



AI发展里程碑 (2021年)

蛋白质结构预测算法AlphaFold
登上 Nature 杂志封面。

-- “可准确预测98.5%的人类蛋白结构”

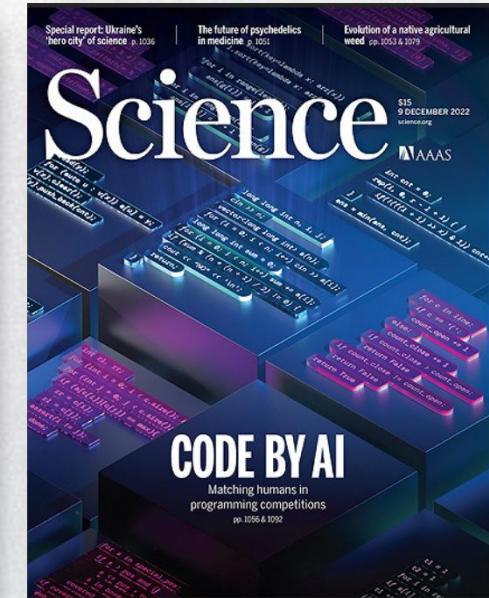
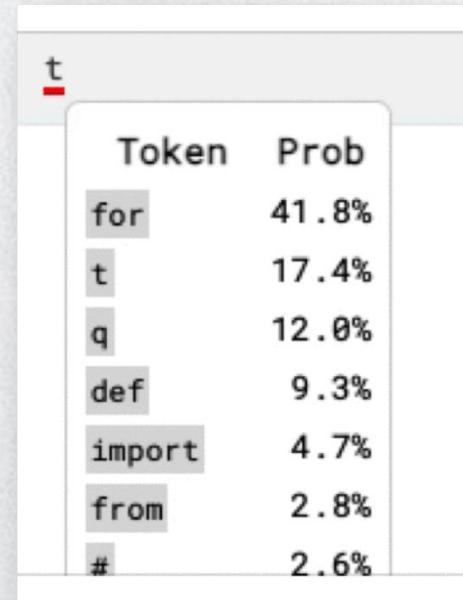


Highly accurate protein structure prediction with AlphaFold, Nature, 2021

AI发展里程碑 (2022年)

AI编程模型AlphaCode登上
Science封面。

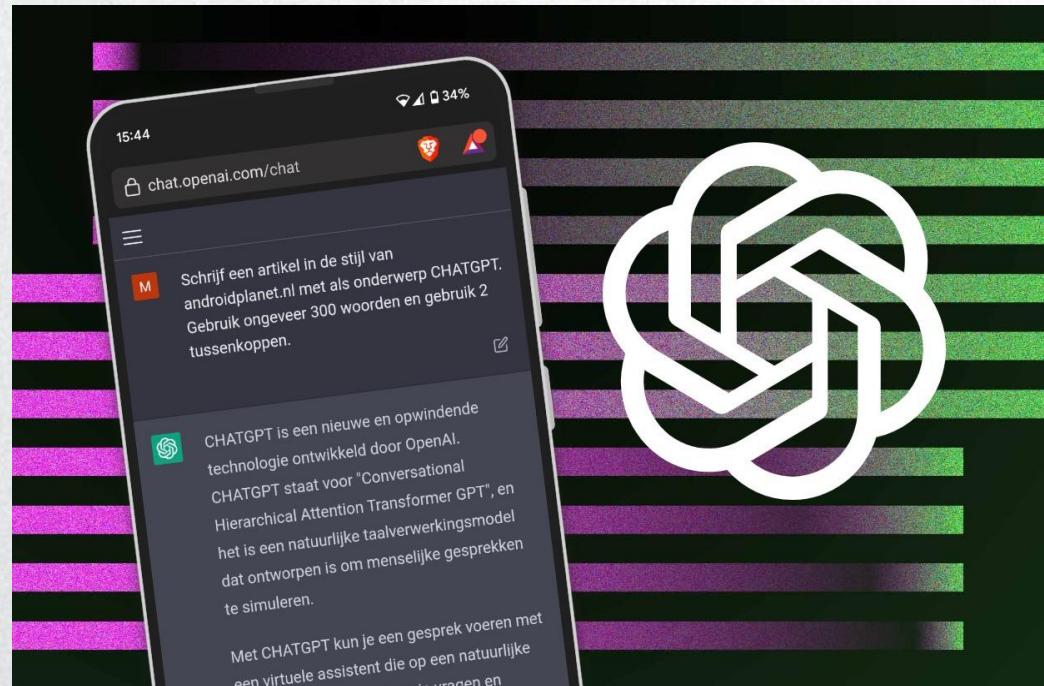
该模型在全球顶级计算机编程
评测系统Codeforces中击败了
半数的人类程序员。



AI learns to write computer code in ‘stunning’ advance, Science, 2022

AI发展里程碑 (2022年)

人工智能对话模型
ChatGPT 发布并突破
1亿用户。



• 国产大模型 百团大战

 科大讯飞	 昆仑万维	 华为
1+N认知智能大模型 <small>未发布</small> 预计2023年5月6日发布	天工3.5 <small>未发布</small> 预计2023年4月17日开始测试	盘古NLP模型 <small>未发布</small> 2023年4月10日举行发布会但未公布上线时间
 达观数据	 网易	 阿里巴巴
曹植 <small>未发布, 可试用</small> 2023年3月18日公布研发进度可申请试用	玉言 <small>未发布</small> 发布时间未知	通义千问 <small>2023年4月11日发布将接入所有阿里产品</small>
 商汤科技	 360	 清华大学
日日新 <small>2023年4月10日发布</small>	360智脑 <small>2023年4月10日发布</small>	ChatGLM-6B <small>2023年3月28日发布</small>
 百度	 澜舟科技	 复旦大学
文心 <small>2023年3月16日发布关键产品“文心一言”</small>	孟子 <small>2023年3月14日发布</small>	MOSS <small>2023年2月21日发布</small>
 腾讯	 中科院自动化所	 智源研究院
混元 <small>2022年12月发布, 预计关键产品“混元助手”近期上线</small>	紫东太初 <small>2021年9月27日发布</small>	悟道2.0 <small>2021年6月1日发布</small>

AI发展里程碑 (2024年2月)

OpenAI 公司发布**文本生成视频(Text-to-Video, T2V) 模型** Sora。

Sora 不仅仅是一个视频生成模型，它旨在探索 AI 如何在理解真实世界运动和交互，所以被认为是一个“世界模拟器”。



AI发展里程碑 (2024年9月)

推理大模型，非常擅长科学工程类学科，尤其是数学和编程。

-- “达到了PhD-level”的科学思考能力。



OpenAI o1

人工智能模型发展总结

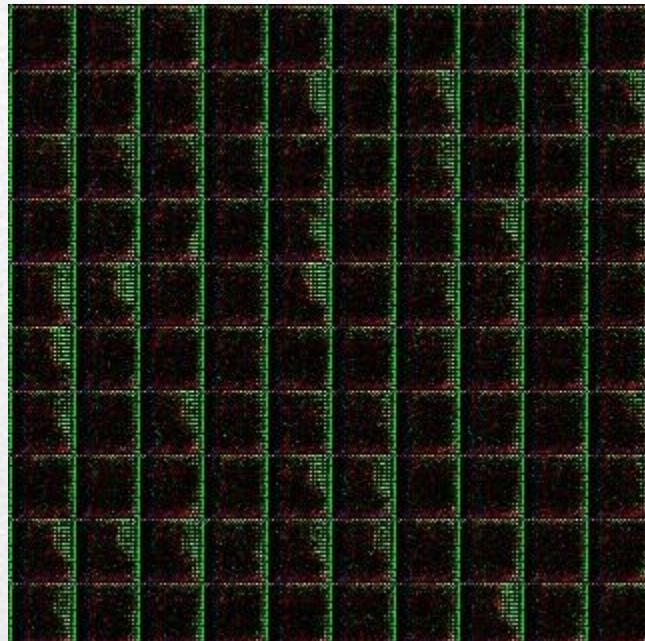
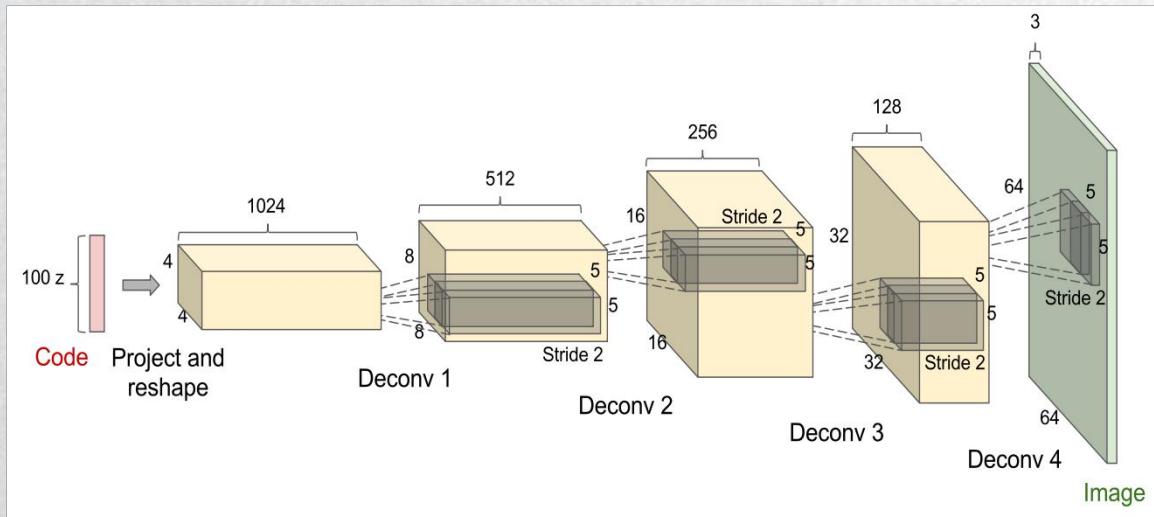
- 2012年，以AlexNet为代表的卷积神经网络(百万参数)
 - AI进入**深度学习**时代
- 2017年，以Transformer为代表的自然语言模型(>1亿参数)
 - AI进入**大模型**时代
- 2020~至今，以GPT-3为代表的预训练大模型(>1000亿参数)
 - 进入**超大模型**时代

生成式 AI 的愿景和实现之路

- OpenAI 在 2016 年提出: **生成模型是让计算机理解世界最有潜力的方向。**
*Generative models are one of the most promising approaches towards this goal.
- 引用物理学家 Richard Feynman的一句话: **我不能创作出来的，我就没有理解。** What I cannot create, I do not understand.
- 文本生成: 2020年出现GPT-3, 2022年提出GPT-3.5, 即 ChatGPT。
- 图像生成: 2016年提出GAN生成对抗模型, 2021年提出扩散模型。
- 语音和视频生成: 2023年出现语音生成网络。2024年出现Sora视频生成。

* <https://openai.com/index/generative-models/>

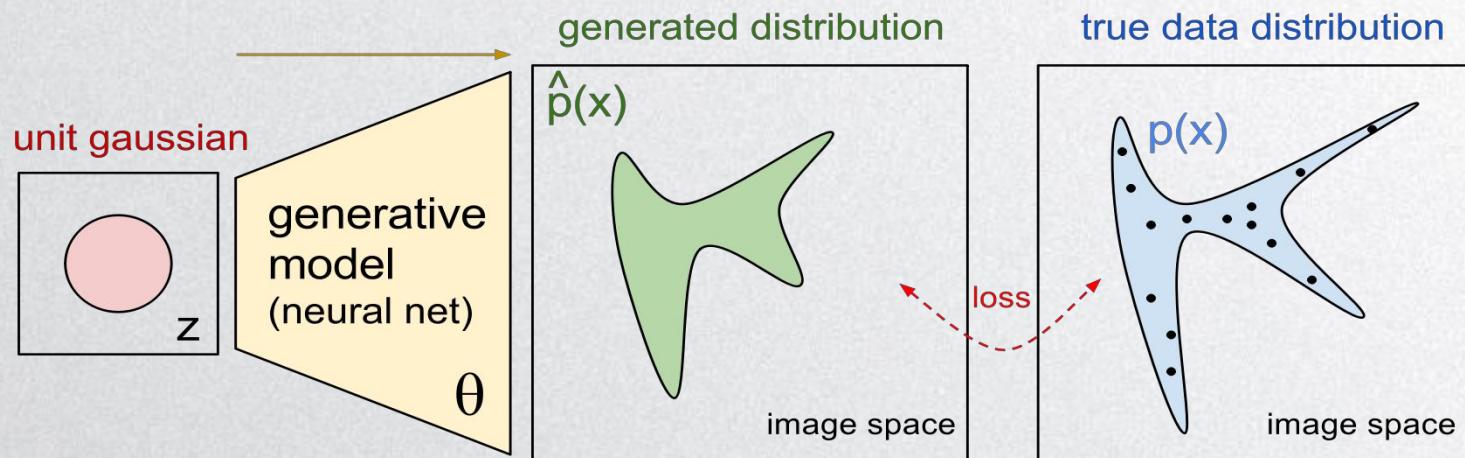
DCGAN (2015)



Radford et al., “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” 2015.

生成网络的本质

从容易抽样的空间，映射到真实数据空间的学习过程。



Reference

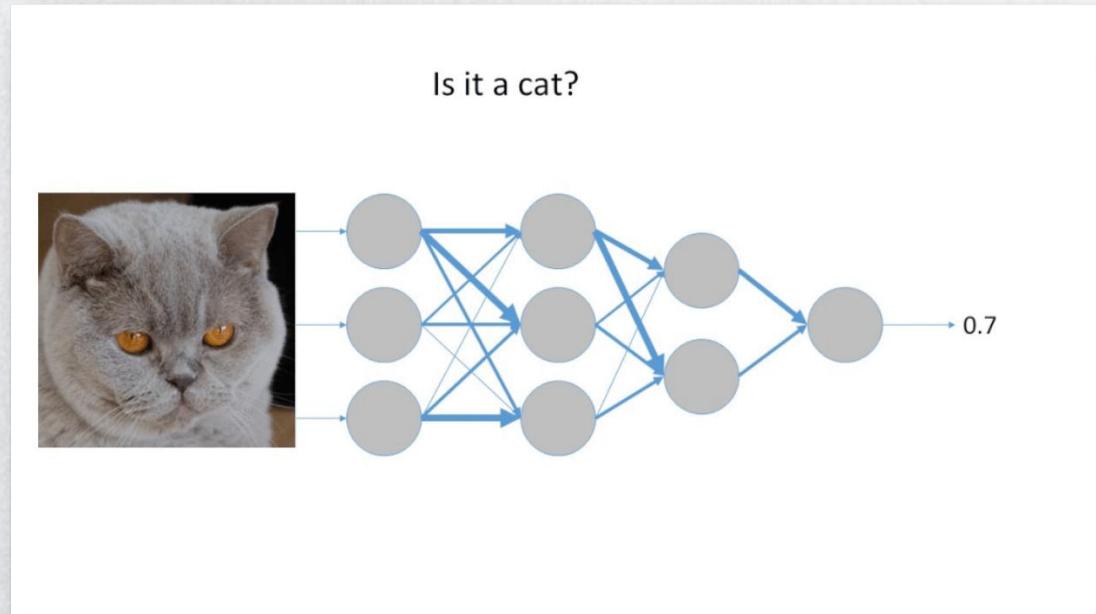
- Ho et al., “Denoising Diffusion Probabilistic Models,” 2020. Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” 2021. **开创性的Diffusion文章。**
- Peebles, W., & Xie, S. “Scalable Diffusion Models with Transformers,” 2023. **图像生成大模型的架构模型基础。**
- Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” 2022. **开创了图像生成大模型基于潜在扩散模型(LDM)文章。**
- Ho et al., “High Definition Video Generation with Diffusion Models,” 2022. **发展了LDM用于视频生成的文章。**
- OpenAI Sora “Video generation models as world simulators,” 2024. **Sora的技术文档和展示。**
- OpenAI O1 “Advancing cost-efficient reasoning.” **技术文档。**

现代神经网络简介

现代深度神经网络(DNN) - 2012年

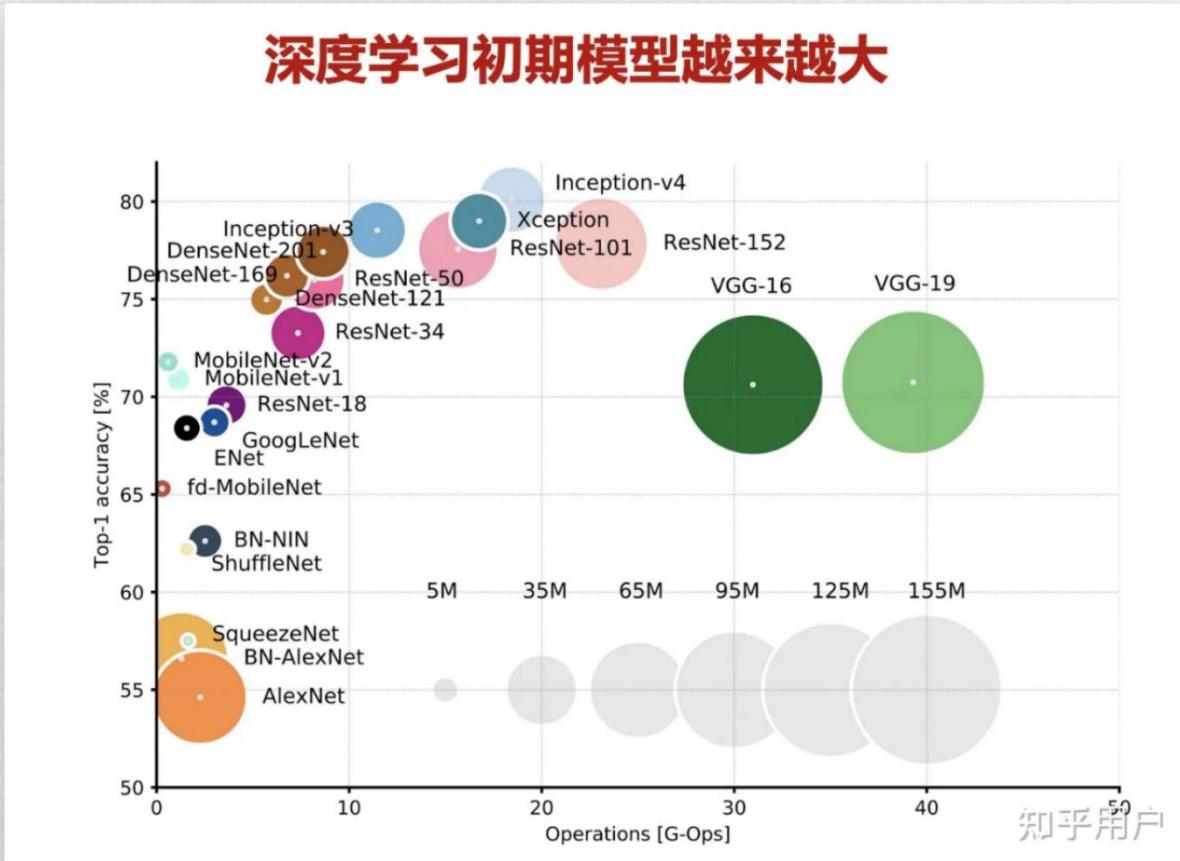
DNN是一种用于模拟人类神经系统的计算模型。

- 使用多层神经元组合来处理复杂的输入数据。
- 可以学习到高维输入特征的复杂关系，如图像视频、语音、自然语言等。
- 可扩展，依赖算力和数据。



ImageNet Classification with Deep Convolutional Neural Networks, 2012.
作者: Alex Krizhevsky, Ilya Sutskever ([OpenAI](#)), Geoffrey E. Hinton

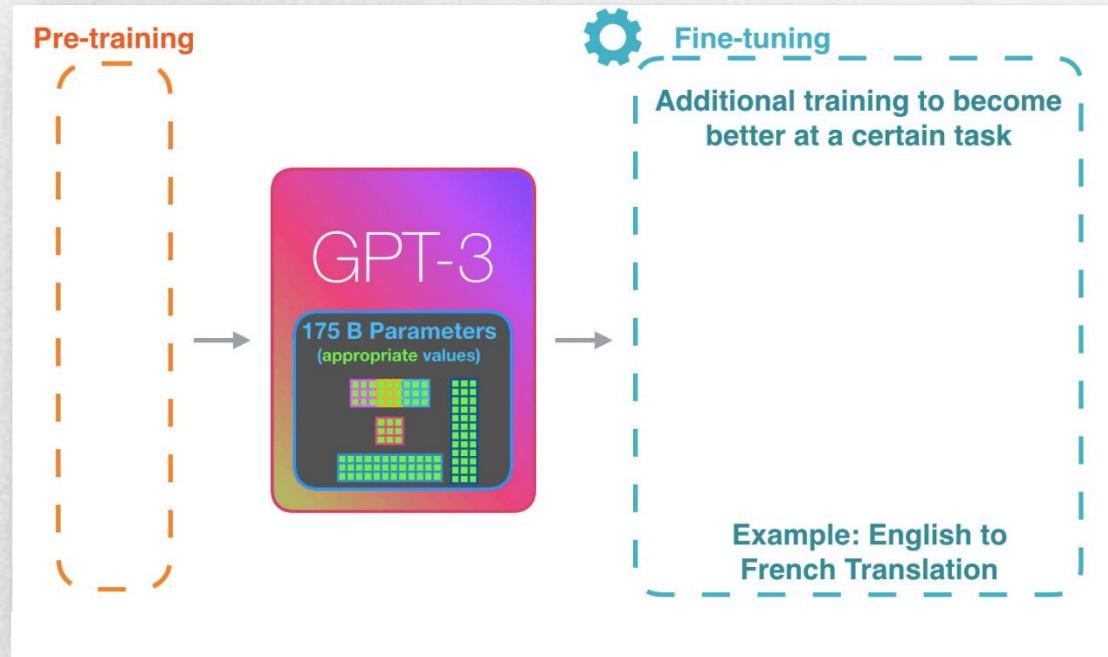
2012-2017 DNN模型规模稳步增长



图片来源于知乎

大语言模型 (Large Language Model)

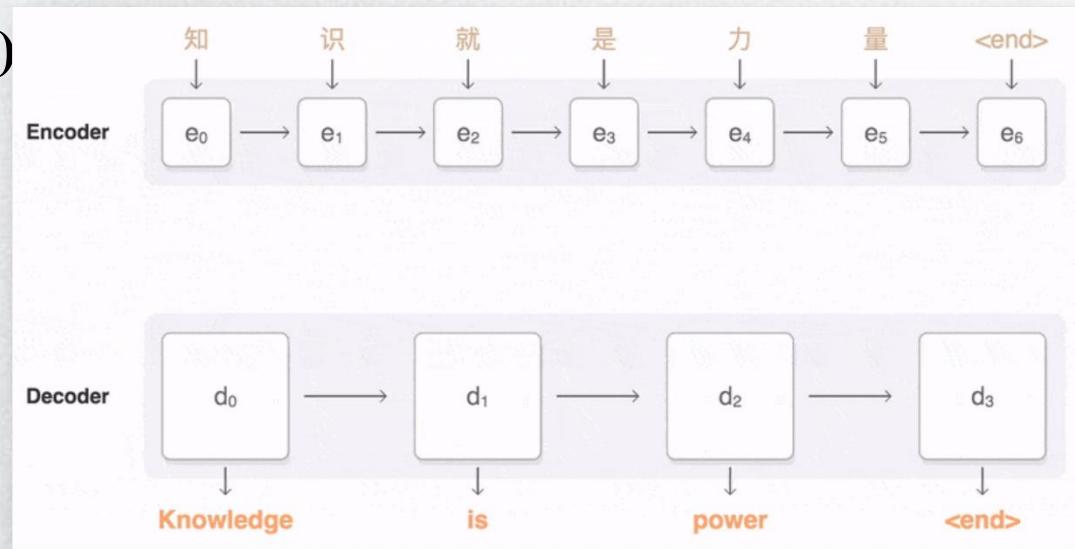
- 参数个数1千亿-2万亿
- 采用预训练加任务微调
- 超级数据中心进行训练



语言模型 Transformer - 2017年

新一代序列数据模型 (参数过亿)

- 基于编码-解码模型
- 计算所有输入单词之间的相关程度(编码)
- 计算所有输入-输出单词之间的相关程度(解码)
- 模型计算量大大增加



ChatGLMModel(
 (word_embeddings): Embedding(130528, 4096)
 (layers): ModuleList(
 (0-27): 28 x GLMBlock(
 (input_layernorm): LayerNorm((4096,), eps=1e-05, elementwise_affine=True)
 (attention): SelfAttention(
 (rotary_emb): RotaryEmbedding()
 (query_key_value): Linear(in_features=4096, out_features=12288, bias=True)
 (dense): Linear(in_features=4096, out_features=4096, bias=True)
)
 (post_attention_layernorm): LayerNorm((4096,), eps=1e-05, elementwise_affine=True)
 (mlp): GLU(
 (dense_h_to_4h): Linear(in_features=4096, out_features=16384, bias=True)
 (dense_4h_to_h): Linear(in_features=16384, out_features=4096, bias=True)
)
)
)
 (final_layernorm): LayerNorm((4096,), eps=1e-05, elementwise_affine=True)
)

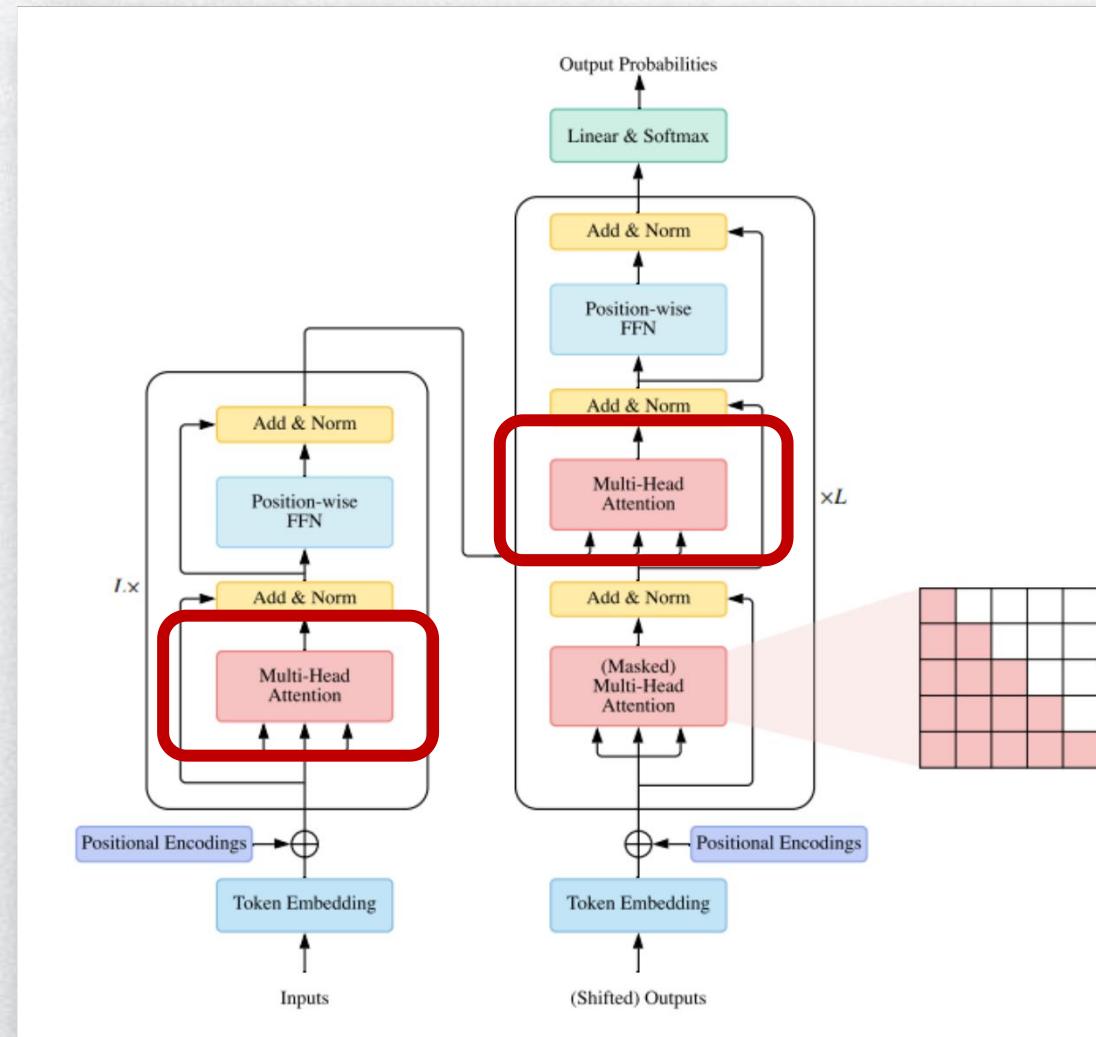
28 layer- 60亿参数模型

70 layer- 1300 亿参数模型

基本单元: 注意力模块

使用基于键值对的注意力机制，以实现：

- 自注意力(self-attention) 对输入进行编码
- 交叉注意力(cross-attention) 对输出进行解码
- 更好理解上下文
- 可以输出更长的文字, 理解历史对话

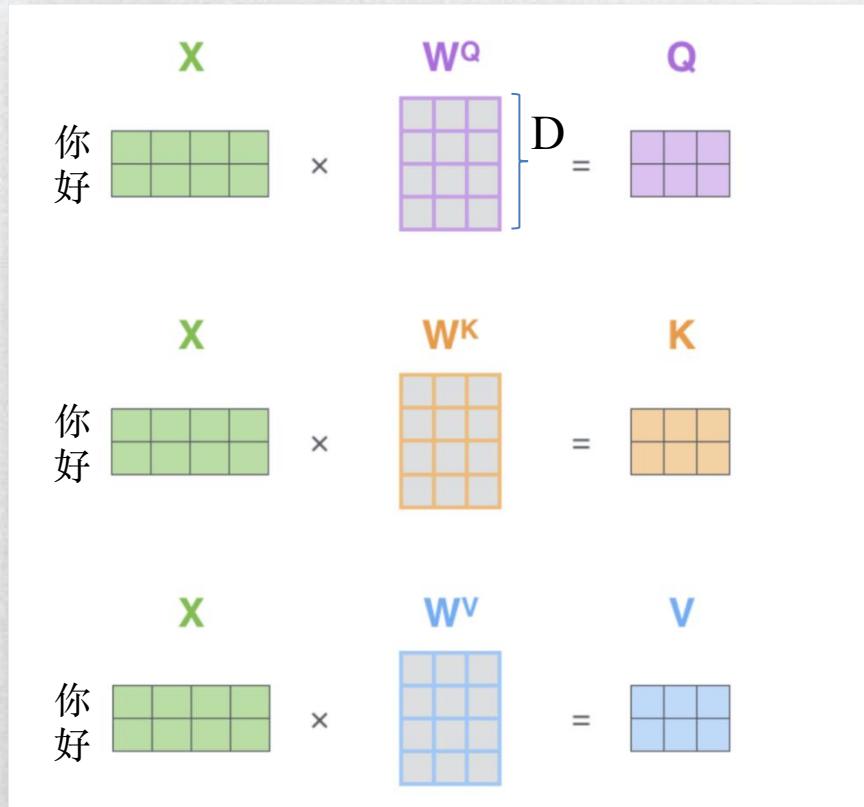


基于键值对的注意力机制(主要的计算消耗单元)

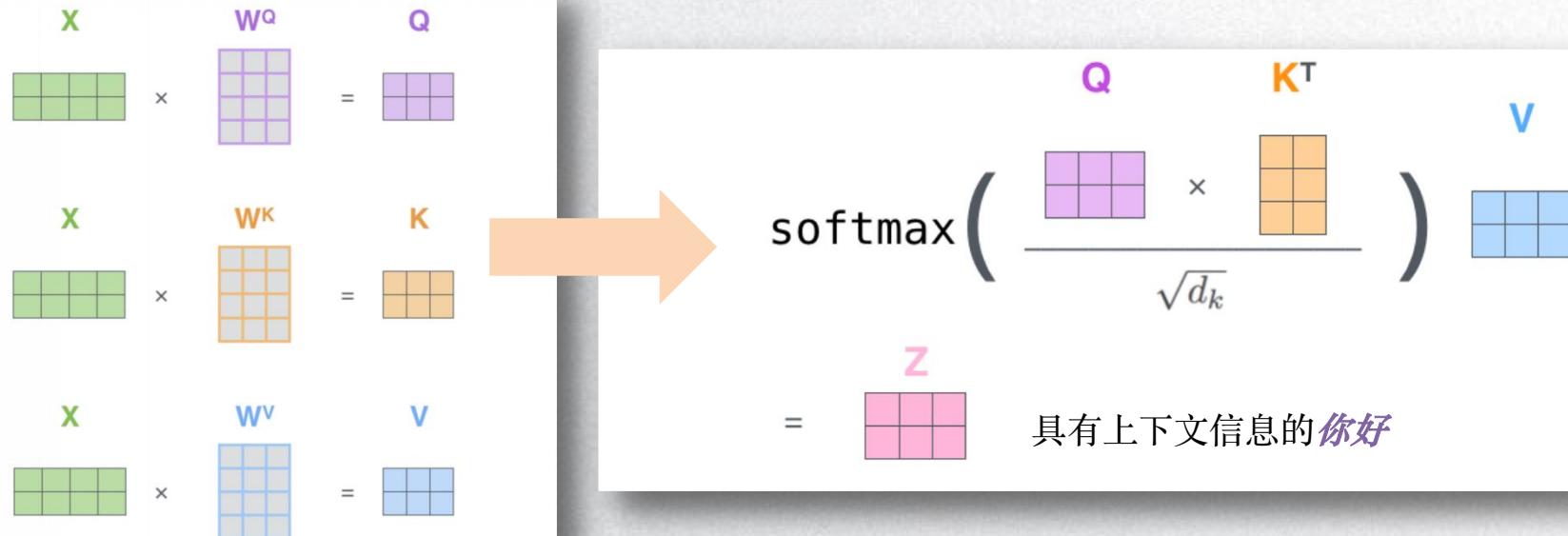
对一句话每一个单词

- 计算Query (查询)
- 计算Key (键)
- 计算Value (值)

参数量为 $O(D^2)$



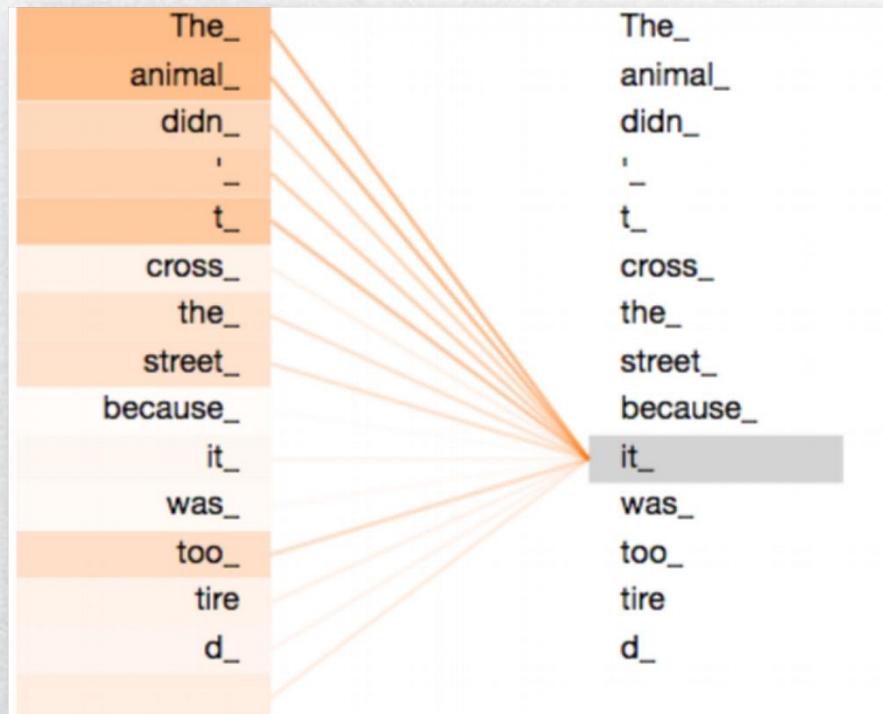
基于键值对的注意力机制(主要的计算消耗单元)



计算量为 $O(N*D^2)$, N 为单词个数

注意力机制的目的 – 更好的编码和解码!

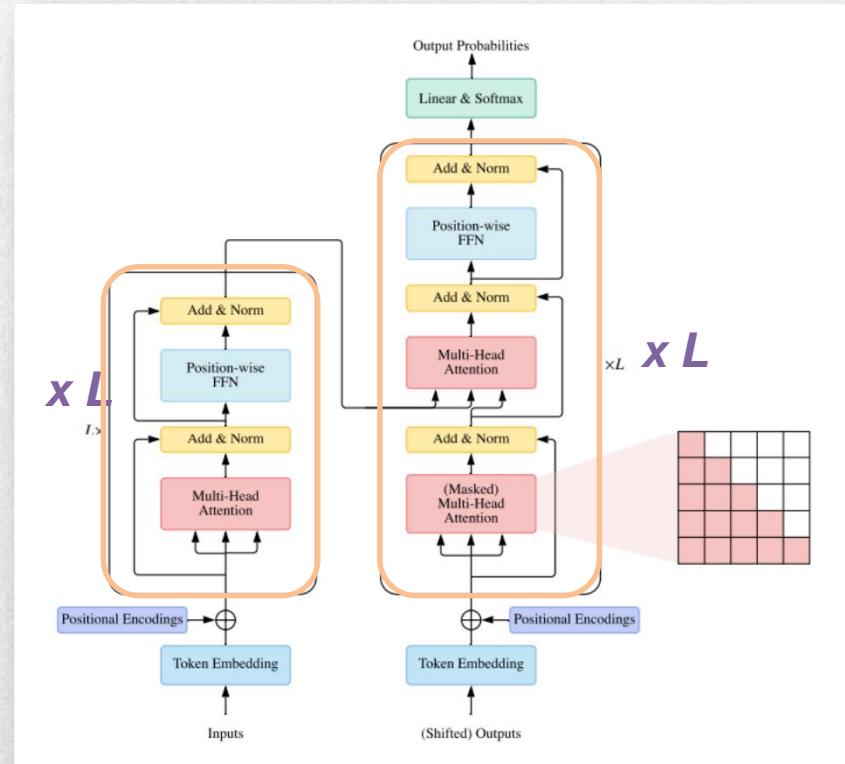
- 通过注意力机制编码
- 每一个单词均能找到合适的上下文信息
- 例如, it 关联到
 - animal,
 - cross-the-street,
 - too tired



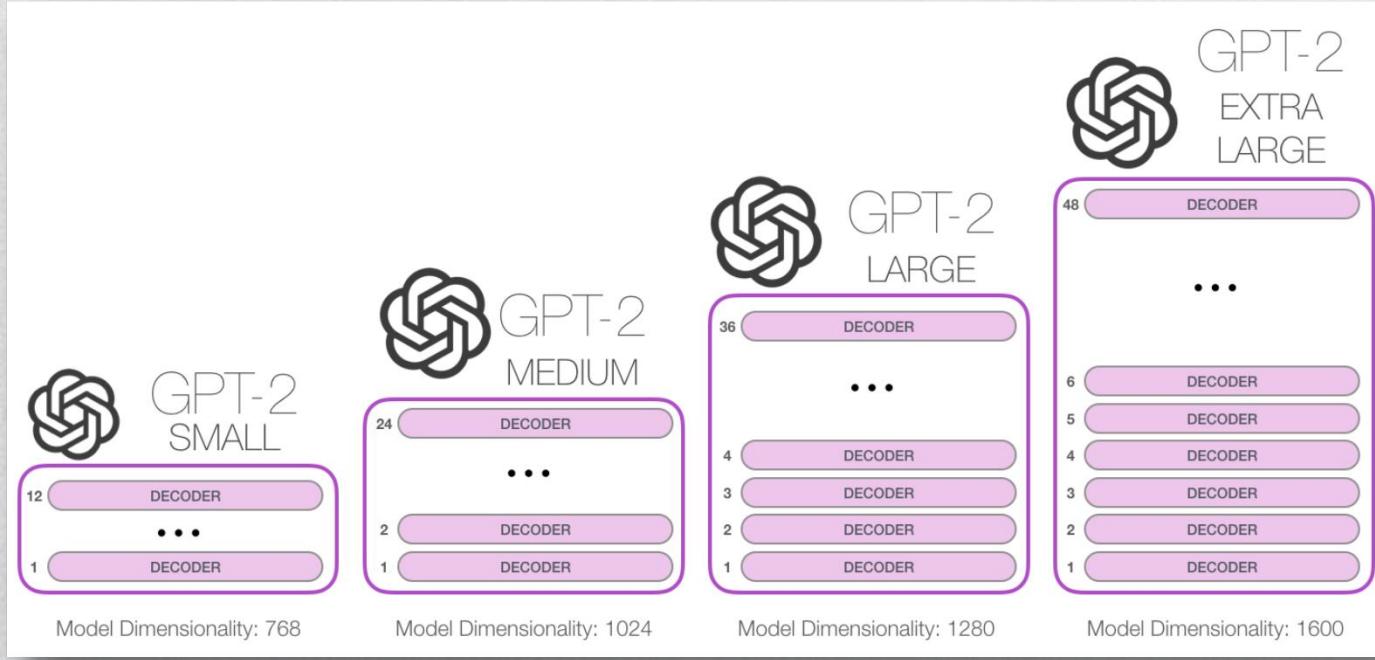
The animal didn't cross the street because it was too tired.

模型参数膨胀的秘密

- 大量堆叠注意力模块层数 L
e.g., $L=96$ (ChatGPT)
- 隐藏层维度 D 增长为 2048
- 注意力特征参数 $O(D^2 \cdot L)$
- 其他参数, 如词嵌入, 输出层
 - 达到上亿级别



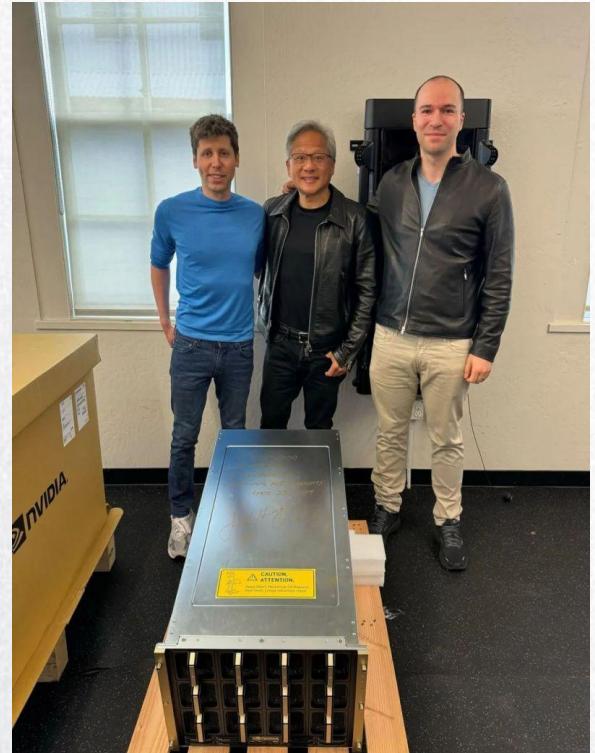
层数(纵向)叠加



维度(横向)增加

Bigger is better ?

Yes, said by OpenAI and Nvidia.



缩放定律 Scaling Law (2020年提出)

- 中文大多翻译为“尺度定律”，但是“缩放定律”更直观，是对模型能力的“缩放”定量研究。
- 当增加AI大模型的**参数量、训练数据量、计算量**时，模型的性能会以某种可预测的方式提高。
- 这种现象最早由OpenAI在其GPT语言模型的研究中，系统地提出并验证。

缩放定律 Scaling Law

- Scaling law的公式研究表明，模型误差 (E) 与模型规模 (N) 、数据量 (D) 以及计算量 (C) 之间存在某种幂律关系。

$$E \propto \left(\frac{1}{N}\right)^\alpha + \left(\frac{1}{D}\right)^\beta + \left(\frac{1}{C}\right)^\gamma$$

- 通过增加模型规模、数据量和计算量，可以显著降低模型误差，从而提升模型性能。

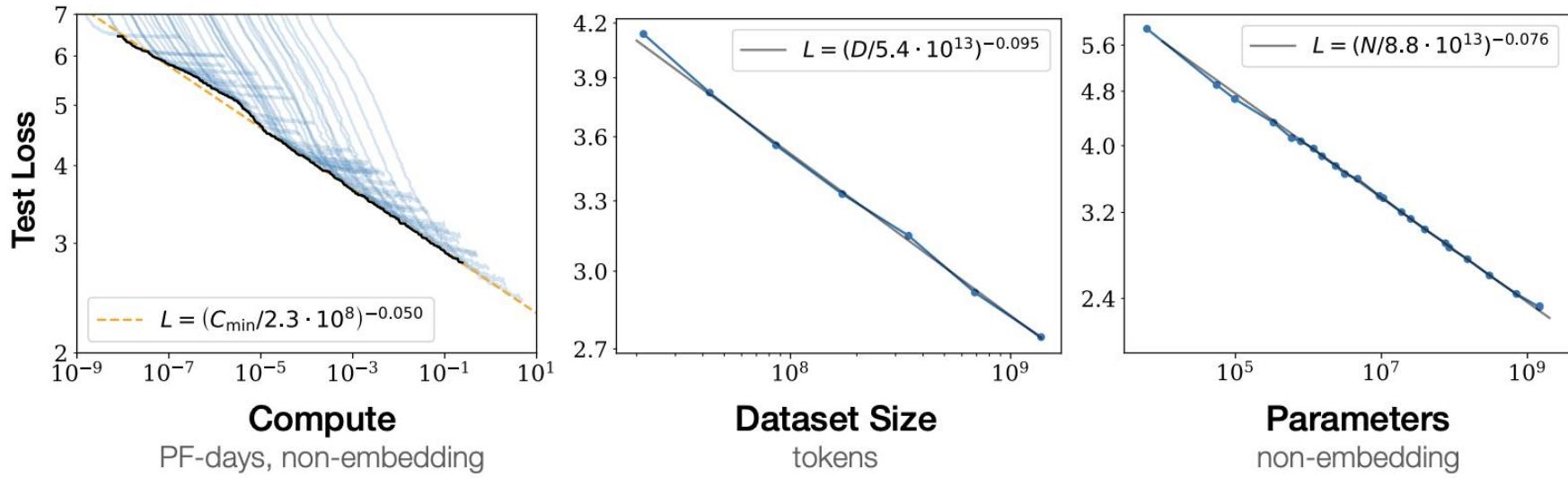


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

缩放定律的启示

- 模型性能受三个因素共同影响，其中**计算量对性能的提升最为显著，其次是模型参数。数据集大小的影响相对较小。**
- 若假设计算量不受限制，那么模型越大越好

Hoffmann et al.'s model

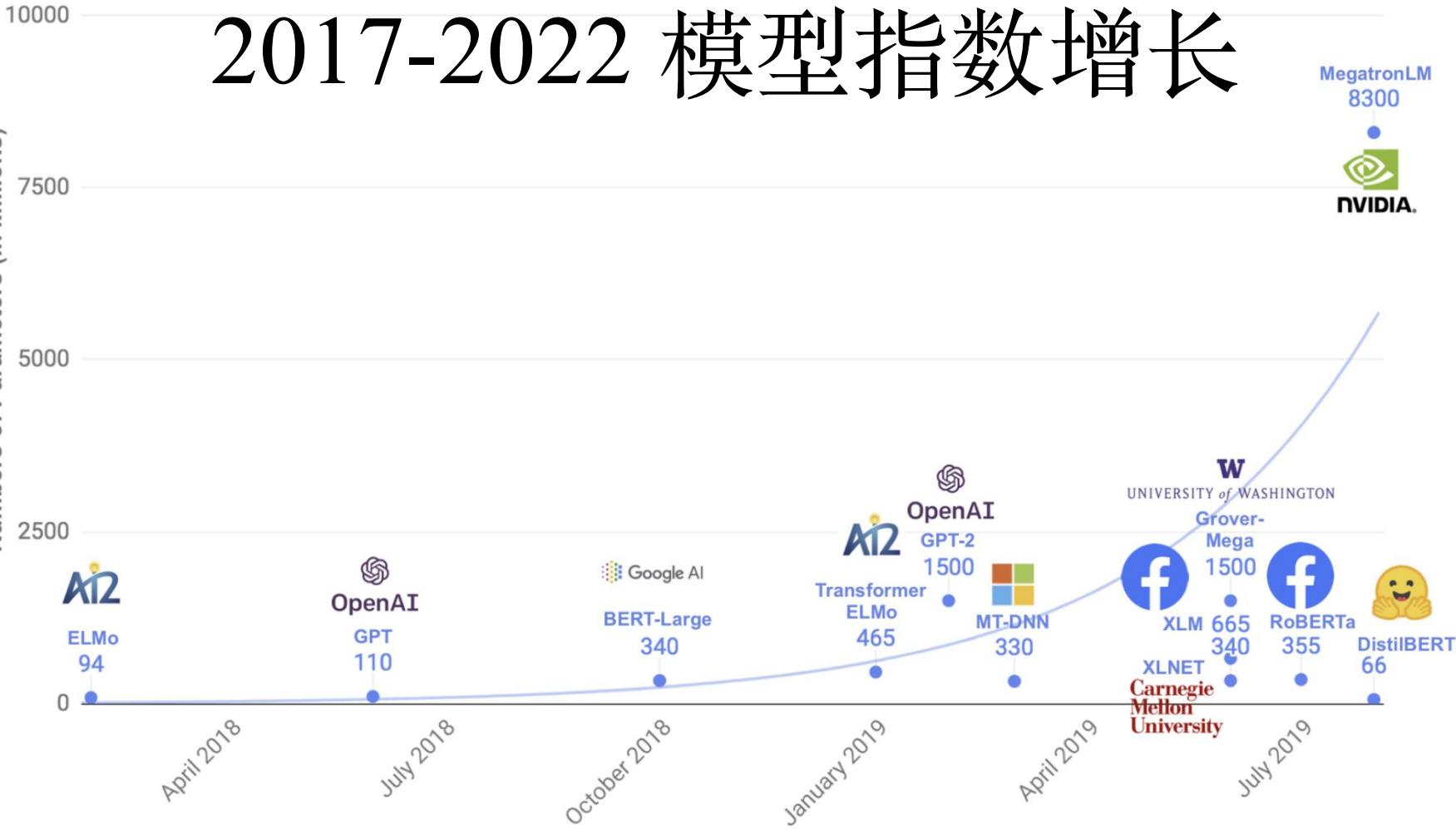
$$L(N, D) = 1.69 + \frac{406.4}{N^{0.34}} + \frac{410.7}{D^{0.28}} \quad (4)$$

缩放定律的启示

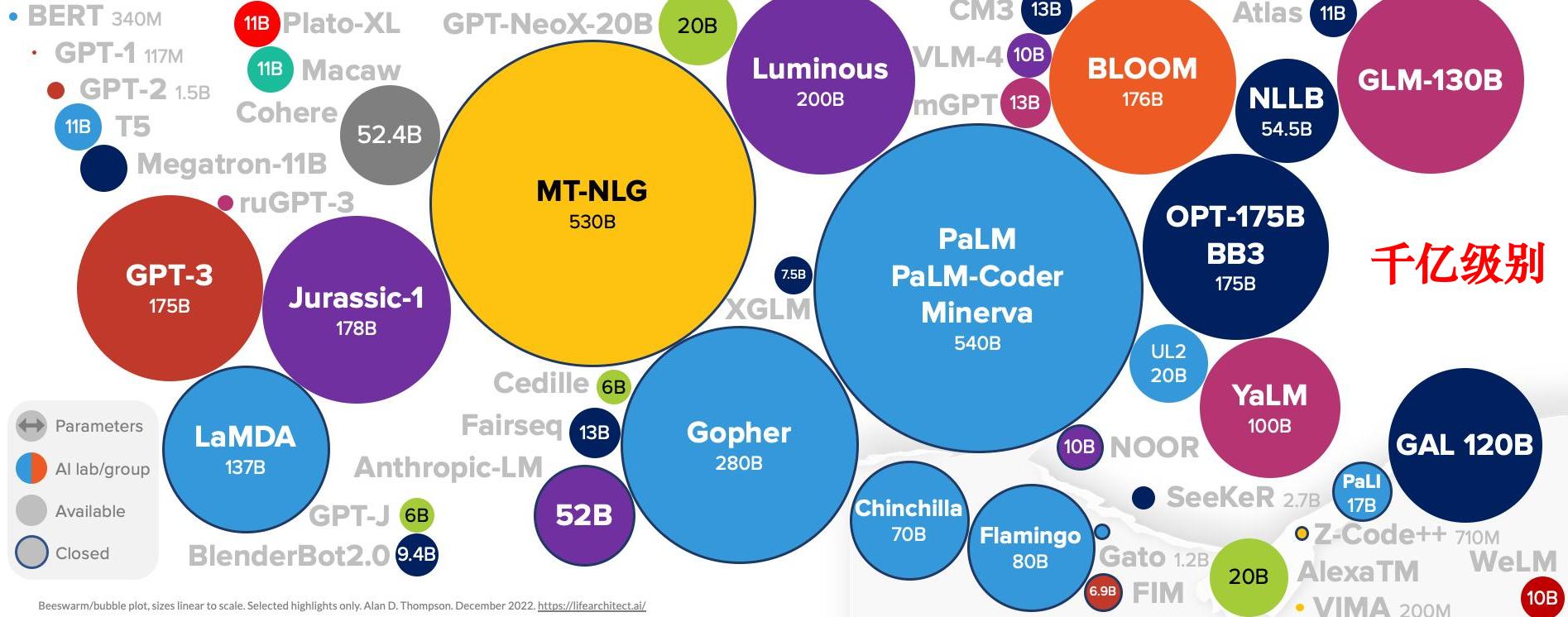
- 在有限的计算资源下，为了获得最佳性能，可以通过训练大型模型并在其达到最佳状态前终止训练。
- 为了实现计算成本的最优，模型的规模和训练数据量（Token 数）应保持成比例的增长，即模型规模翻倍时，训练数据量也应翻倍。
- 在训练大型语言模型时，应重视高质量数据集的扩充，但只有在数据质量高的情况下，扩大数据集的优势才能得到充分发挥。

2017-2022 模型指数增长

Numbers of Parameters (in Millions)



LANGUAGE MODEL SIZES TO DEC/2022

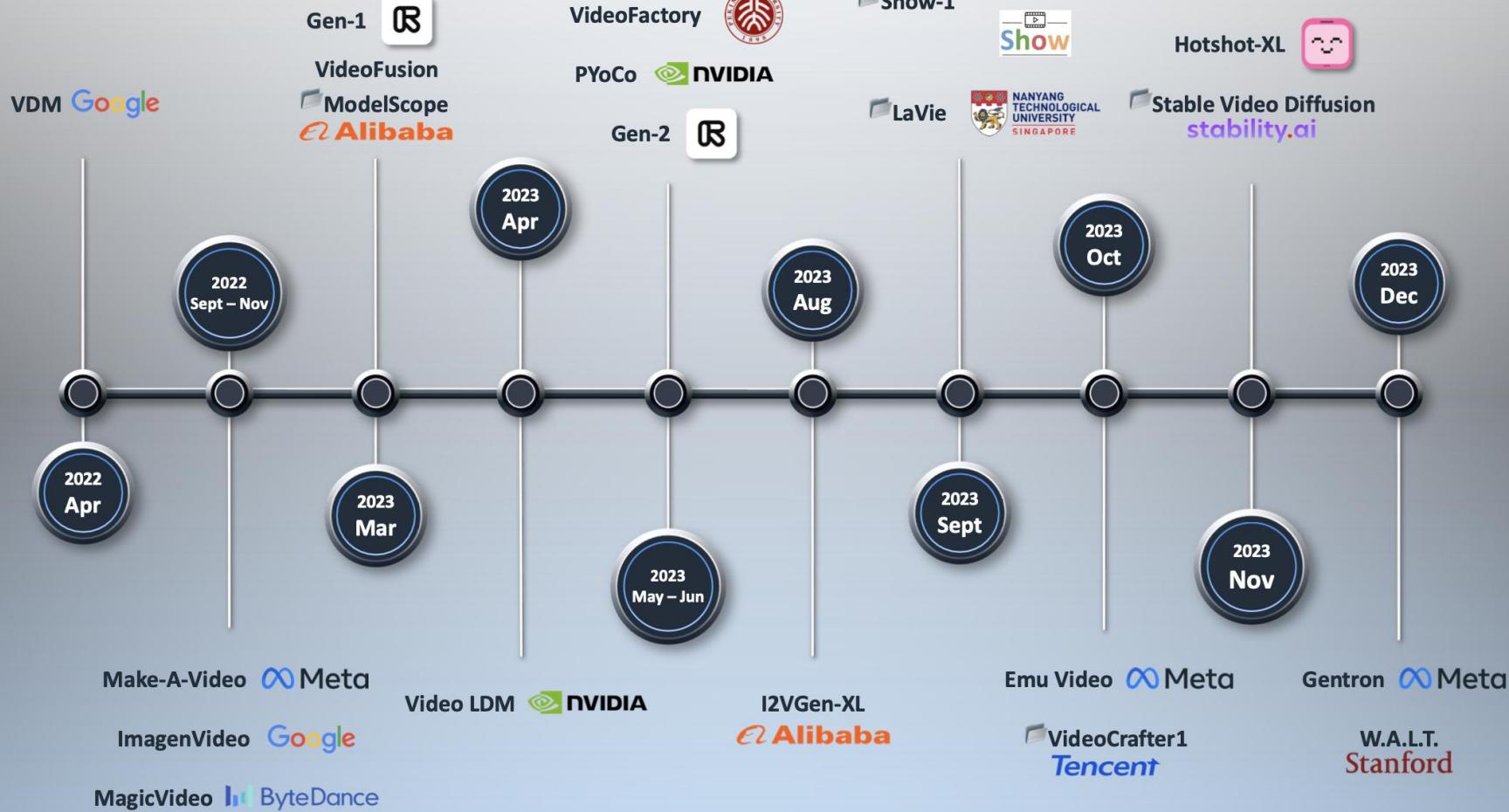


现代视觉大模型展示

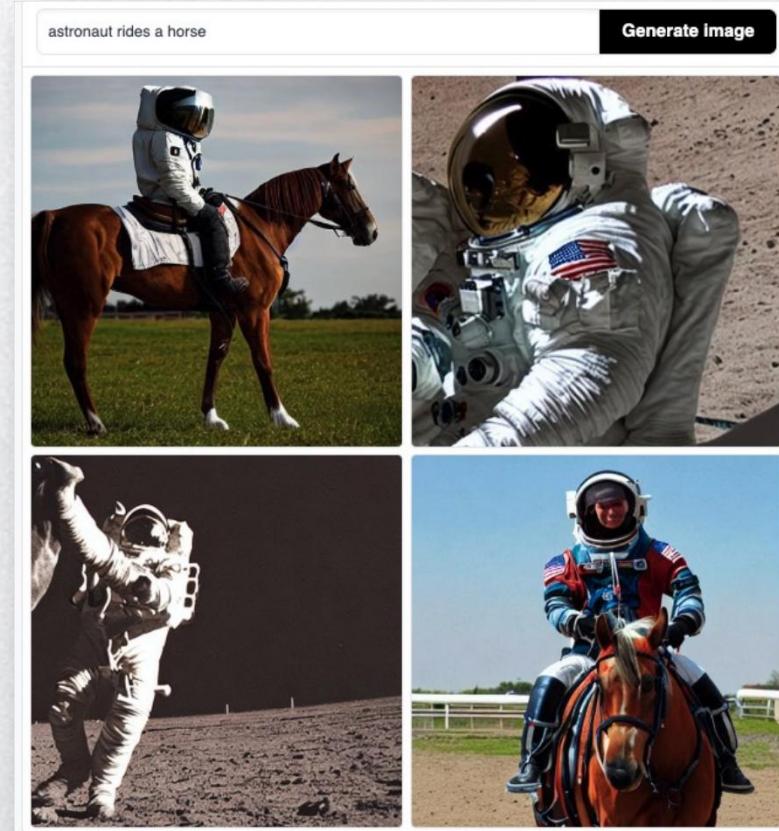
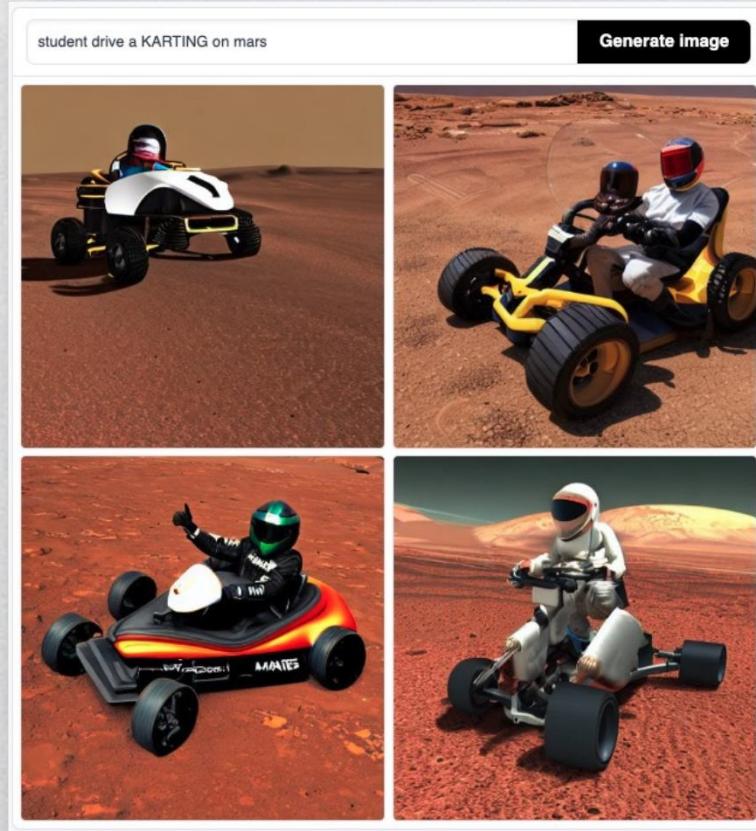
- 在2024年2月，OpenAI 提出Sora, 一种强大的文生视频 (Text-to-Video) 技术。
- “Sora 展现出的能力表明，视频模型是一条通往强大模拟器的希望之路，可以模拟物理世界、数字世界以及生活在这些世界中的对象、动物和人。”
- 值得注意的是，Sora并不是原创或唯一的文生视频模型，在这之前已有多款模型。

* <https://openai.com/index/video-generation-models-as-world-simulators/>

Video Foundation Model



图像生成 - Diffusion扩散模型



图片动态化 Cinemagraph

原始图片+掩码



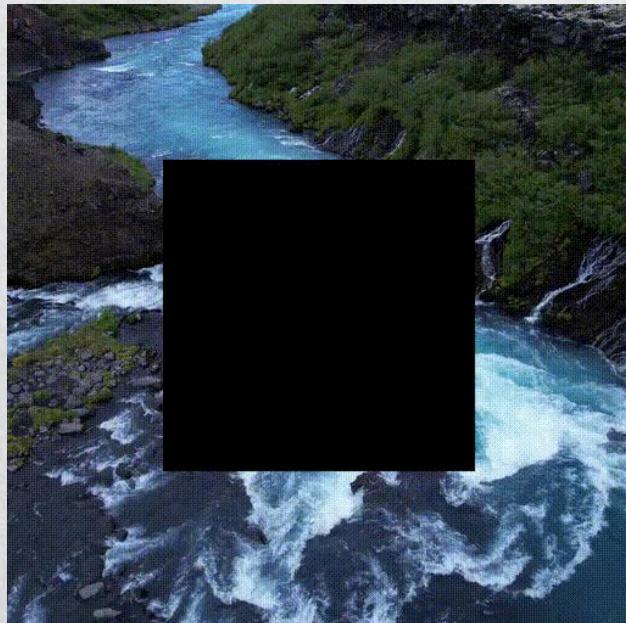
动态视频



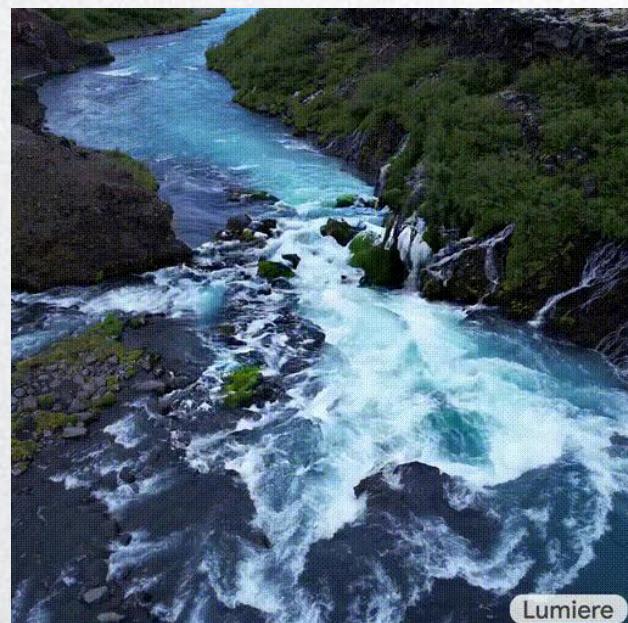
视频补全

Video Inpainting

原始视频



补全视频



Lumiere

风格迁移 Video Stylization

原始视频



"Made of colorful toy bricks"



"Made of flowers"



人物部分风格迁移

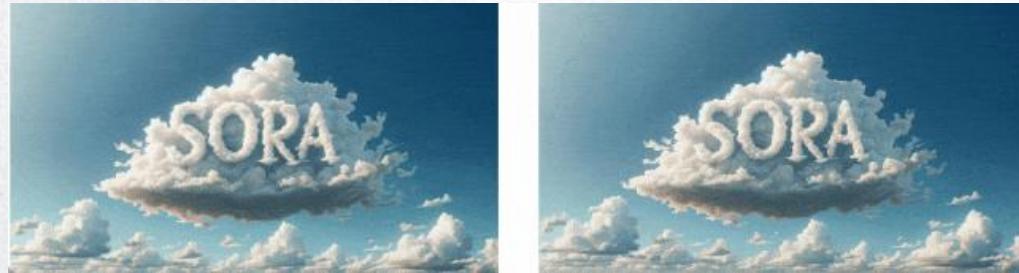
原始视频



"wearing a gold strapless gown"



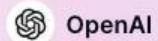
什么是SORA?



- 2024年2月美国硅谷OpenAI 公司发布**文本生成视频(Text-to-Video, T2V)模型** Sora 。
- Sora 不仅仅是一个视频生成模型，它旨在探索 AI 如何在理解真实世界运动和交互，所以被认为是一个“世界模拟器”。
- Sora 可以生成长达1分钟的高清视频，在生成时长和生成质量上较其他现有模型和产品实现了明显突破。



prompt: Exploring Space / Time with Sora.
This isn't going to replace the filmmaking process, rather it's offering an entirely new way of thinking about it. Not restricted by time, money or other people's permission, I can ideas and experiment in bold and exciting ways



OpenAI

03/27/2024, 04:55 PM



prompt: The Glenfinnan Viaduct is a historic railway bridge in Scotland, UK, that crosses over the west highland line between the towns of Mallaig and Fort William. It is a stunning sight as a steam train leaves the bridge, traveling over the arch-covered viaduct. The landscape is dotted with lush greenery and rocky mountains, creating a picturesque backdrop for the train journey. The sky is blue and the sun is shining, making for a beautiful day to explore this majestic spot.

 OpenAI

02/15/2024, 06:14 PM



66

prompt: realistic video of people relaxing at beach, then a shark jumps out of the water halfway through and surprises everyone



Tim Brooks

02/17/2024, 12:51 AM

Animating static images



In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave.

英伟达科学家 Jim Fan 评价称：

- Sora 是一个端到端可以训练的 Transformer 模型。
- Sora 通过大量视频数据，学习成为一个物理引擎，通过梯度下降来调整神经网络参数，
- Sora 可能使用虚幻引擎 5 对大量合成数据进行训练



Jim Fan 
@DrJimFan

Apparently some folks don't get "data-driven physics engine", so let me clarify. Sora is an end-to-end, diffusion transformer model. It inputs text/image and outputs video pixels directly. Sora **learns** a physics engine **implicitly** in the neural parameters by gradient descent through massive amounts of videos.

Sora is a learnable simulator, or "world model". Of course it does not call UE5 explicitly in the loop, but it's possible that UE5-generated (text, video) pairs are added as synthetic data to the training set.

视频生成的挑战及Sora的解决方法

- 你正在翻看一本世界名胜相册，不同国家、不同风格的景色照片，有宽阔的海景，有狭窄的巷道，还有夜晚灯火辉煌的城市风光。
- 但你能轻松地辨识每一张照片代表的地点和情感，因为你的大脑能够将这些不同的视觉信息统一理解。
- 算法能够处理和理解来自世界各地、不同设备拍摄的数以百万计的图片和视频吗？
- 算法能够生成以假乱真的图片和视频吗？
- OpenAI 推出的 Sora 能够理解和生成这么丰富的视觉内容。

<https://openai.com/index/video-generation-models-as-world-simulators/>

- 传统方法将所有训练视频裁剪成正方形。
- 导致主题只是部分可见，剪裁会损失主要内容信息。
- Sora 使用原始长宽比训练，生成的视频（右侧）保留了主题。





- 传统方法通常需要调整大小、进行裁剪或者是将视频剪切到标准尺寸。例如固定时长4秒，视频分辨率为 256×256 。
- Sora **采样的灵活性**：可以采样宽屏、竖屏 1920×1080 以及两者之间的分辨率。这使Sora可以直接以其天然纵横比为不同设备创建内容。
- Sora还允许以较小的尺寸快速创建内容原型。

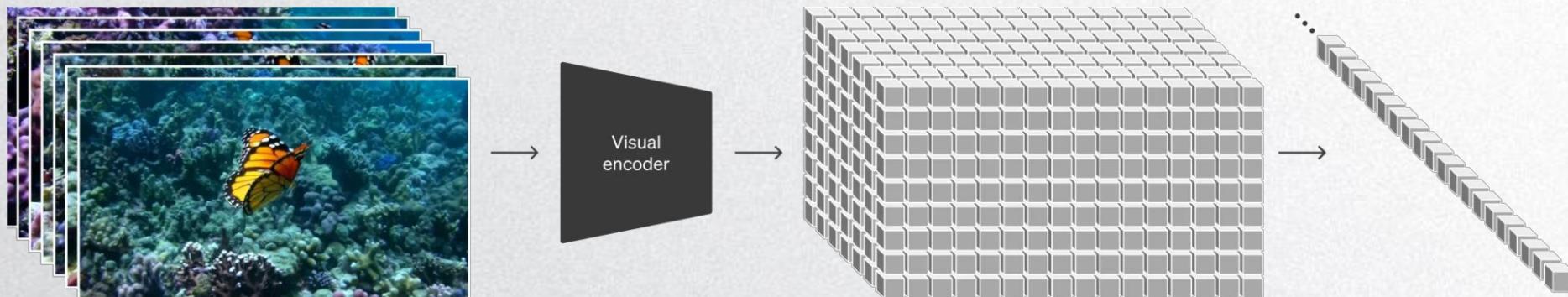
Sora 原理简介



Sora 的基础--压缩时空图像块

- 时空图像块 (Spacetime visual patches)
- 启发于大语言模型LLM，LLM 使用了 token 作为最小的词单元。
- 而Sora则是利用了图像块visual patches作为最小的视觉单元。
- 图像块是一种高度可扩展且有效的表示，可以用于在不同类型的视频和图像上训练生成模型。

Sora 的基础--压缩时空图像块



数据维度 $T \times H \times W \times C$

T : 多少帧视频

H, W : 每一帧的长宽

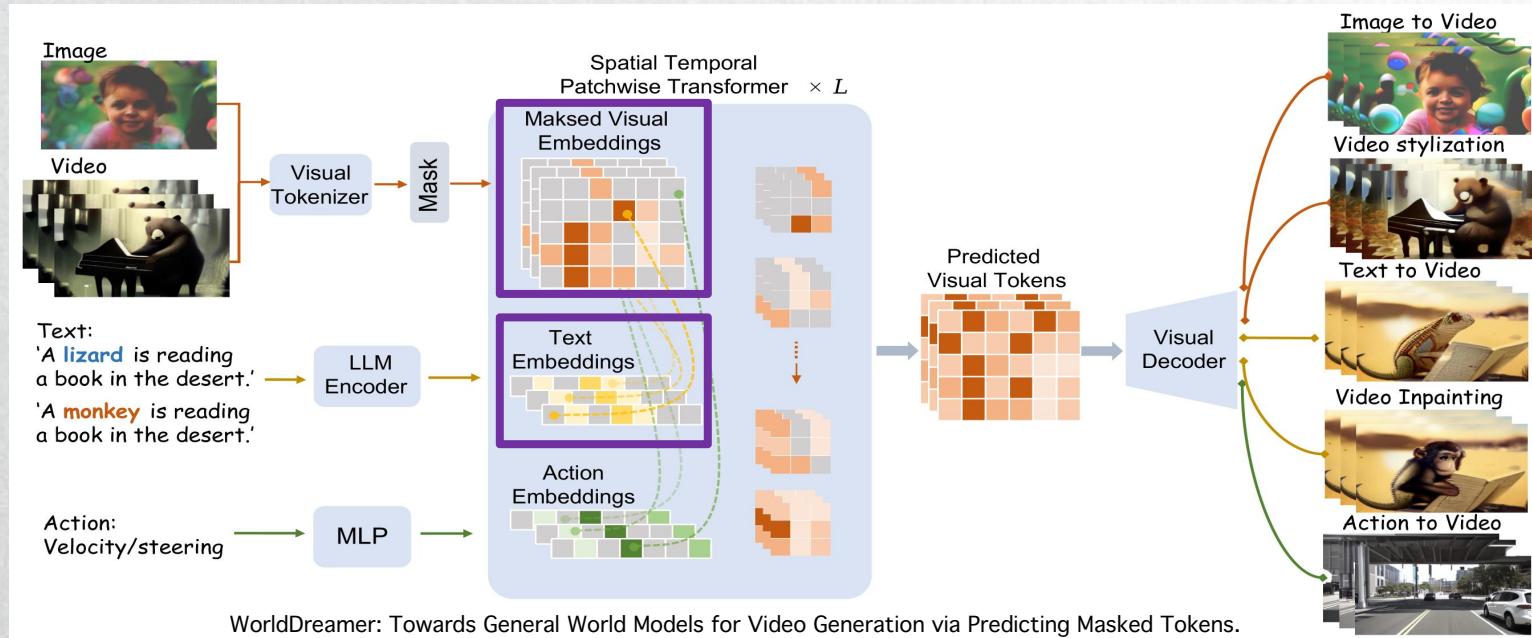
C : RGB图像的通道=3

压缩 Compress
编码 Encode

时空分解

展开成为令牌

- 这种图像块，有着更广泛的称呼：视觉令牌 token。
- 文本内容通过词嵌入，也组建成令牌形式。
- 因此Sora等视频大模型，可以将多模态数据，统一成token的向量格式。
- 因此，可以实现风格迁移、融合、文生图、文生视频等功能。



- Sora 类视觉大模型，学习如何对多模态 token 的变换和组合。
- 因此可以理解文本提示，还能利用不同类型的视觉信息，生成与文本提示相匹配的视频或图片。
- 就好比是从全世界的视觉数据中找到能够拼凑出你想象中的“海底翩翩起舞的蝴蝶”场景的片段，创造出一个全新的视觉作品。
- 这就使得大模型有了“创造性”！



海底翩翩起舞的蝴蝶！

使用时空图像块的优点

- *OpenAI的技术综述没有提供技术细节。学者推断其模型架构应当遵循了当前研究的思想。
- Sora基于扩散模型与Transformer的结合DiT，用于从高维的时空碎片张成的空间中，观察并提取丰富的时空碎片之间的关联与演化的动态过程。
- 视频其实是记录了时空信息的载体：时空碎片patch可以看作是三维空间的点集(x,y,z)的运动(t)或者说其实是个四维时空模型(x,y,z,t)。

Sora 技术细节篇

“Je pense, donc je suis” -- René Descartes

Sora 等文生视频的技术创新总结：

- 1) 使用潜在扩散模型LDM自编码器实现时间维度压缩，使得长视频生成成为可能；
- 2) 直接对LDM中潜视频进行图块化处理并直接使用Visual Transformer建模，解除输入格式限制的同时，能够创新性地实现任何像素和长宽比视频的生成；
- 3) 其训练数据集中可能包含带有物理信息的合成数据，从而使模型展现出对物理信息的初步理解能力；
- 4) 复用重标注技术，对视频数据生成高质量文字标注，借助GPT对提示词进行扩展，提升生成效果。

文生视频中的核心模型-ViT

- 视觉模型ViT能够将图像、视频等输入，转换成令牌
- 通过大量的数据，ViT 能学到高质量的中间特征
- 可以处理不同分辨率大小的图像和视频

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy^{*,†}, Lucas Beyer^{*}, Alexander Kolesnikov^{*}, Dirk Weissenborn^{*},
Xiaohua Zhai^{*}, Thomas Unterthiner^{*}, Mostafa Dehghani^{*}, Matthias Minderer^{*},
Georg Heigold^{*}, Sylvain Gelly^{*}, Jakob Uszkoreit^{*}, Neil Houlsby^{*,†}

*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

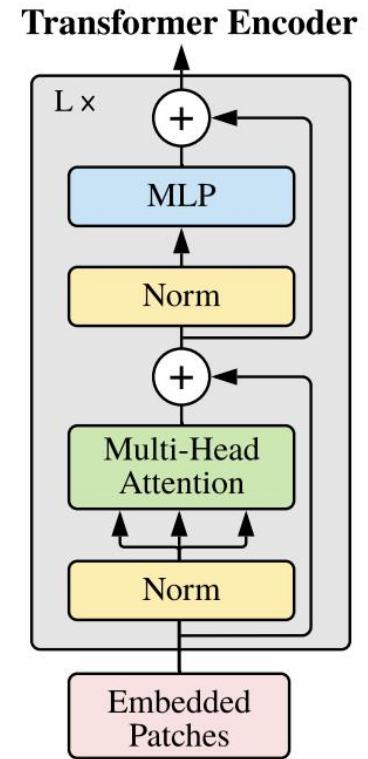
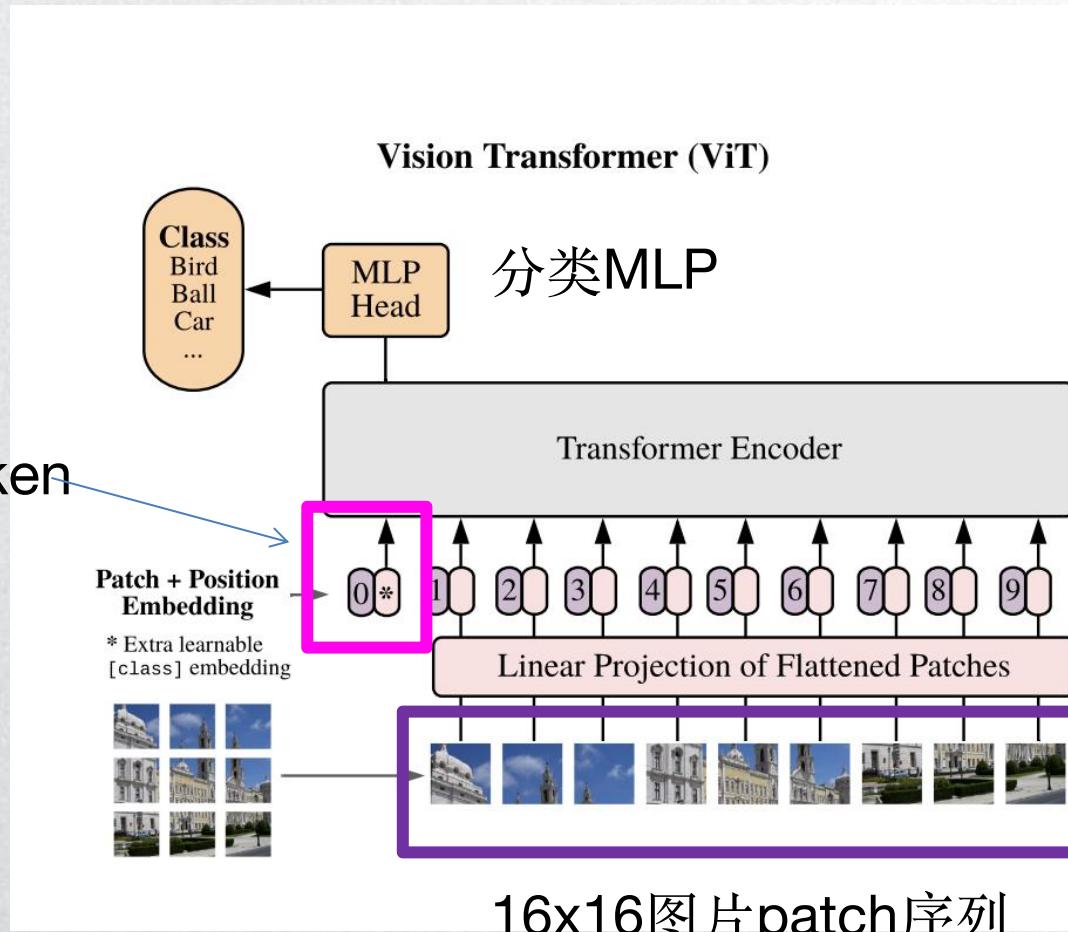
ABSTRACT

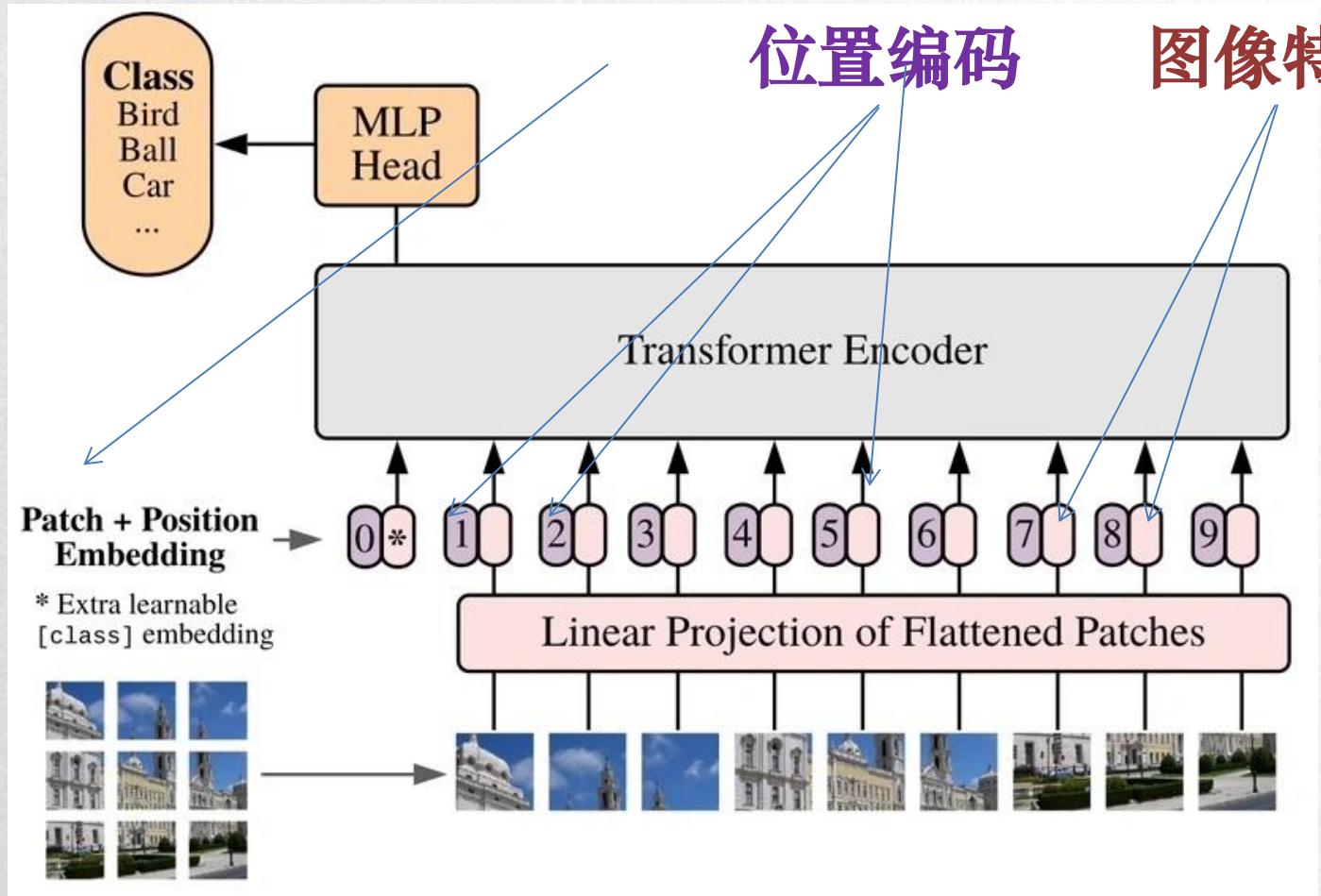
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.¹

ViT模型架构简述

- ViT将输入图片分为多个自定义大小 (如 16×16) 的图像块 (patch)
- 再将每个 patch 投影为固定长度的向量送入Transformer Encoder
- Encoder中的多头注意力机制，对图像patch序列进行编码，生成token
- 这样，2D图像或者3D视频，通过ViT生成了令牌序列，可以进行流式处理，可以自由地关注到任何其他的图像块和文本信息

[CLS] token





Huggingface / Google 官方预训练模型

```
from transformers import ViTImageProcessor, ViTForImageClassification
from PIL import Image
import requests

url = 'http://images.cocodataset.org/val2017/000000039769.jpg'
image = Image.open(requests.get(url, stream=True).raw)

processor = ViTImageProcessor.from_pretrained('google/vit-base-patch16-224')
model = ViTForImageClassification.from_pretrained('google/vit-base-patch16-224')

inputs = processor(images=image, return_tensors="pt")
outputs = model(**inputs)
logits = outputs.logits
# model predicts one of the 1000 ImageNet classes
predicted_class_idx = logits.argmax(-1).item()
print("Predicted class:", model.config.id2label[predicted_class_idx])
```

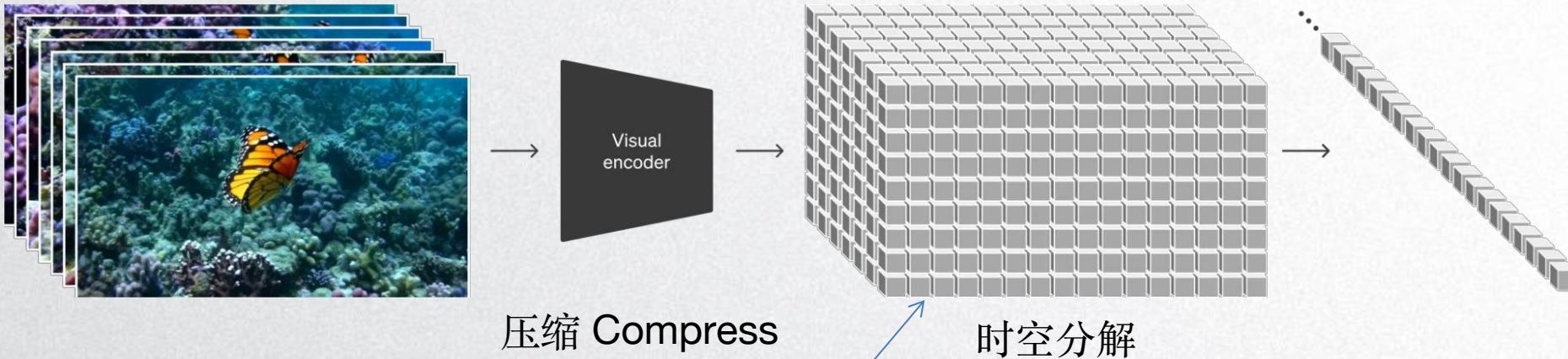
The screenshot shows the Hugging Face Model Hub interface for the 'vit-base-patch16-224' model. At the top, there's a navigation bar with icons for Google, Image Classification, Transformers, and PyTorch. Below that, it says 'License: apache-2.0'. There are tabs for 'Model card' and 'Files and versions', with 'Files and versions' being active. A dropdown menu shows 'main' selected. The main area lists the following files:

- .gitattributes
- README.md
- config.json
- flax_model.msgpack
- model.safetensors
- preprocessor_config.json
- pytorch_model.bin
- tf_model.h5

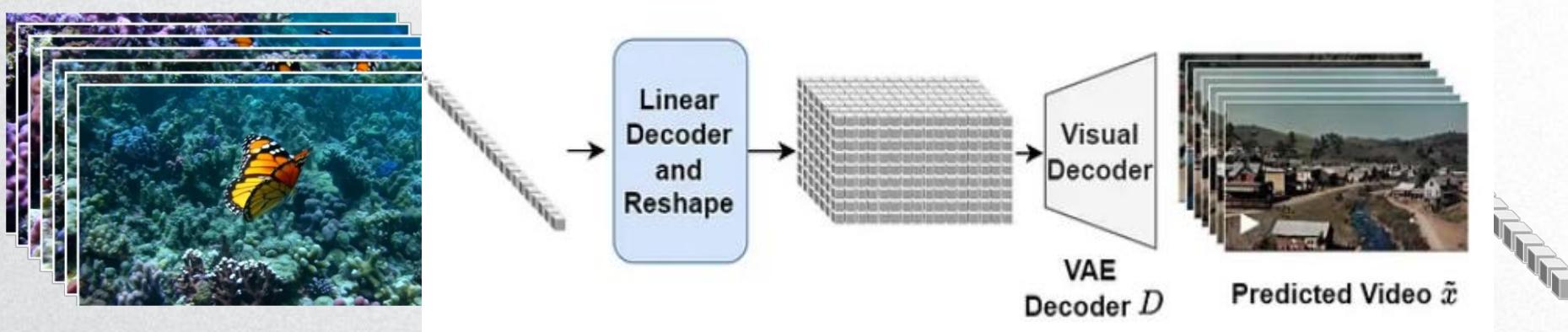
At the bottom right, there are buttons for 'Edit', 'View', and 'Add file'.

ViT模型处理token（第一层）

- 3x224x224 大小的输入图片分为 3x16x16 的patch，一共有 $196=14\times 14$ 个图片patch，组成序列
- 每个patch维度 $768=3\times 16\times 16$, 一层变换之后仍然保留 768 维度
- 输入的token序列大小为 $(196+1) \times 768$, 其中加入了 [CLS] 特殊token
- 多头 Attention 模块将输入映射到同样大小输出 197x768
- 使用 GeLU 激活函数 + LN归一化



- Sora首先将视频压缩到较低维的特征空间，然后分解为时空图像块。
- 该输出可以看做在时间和空间都被压缩的潜在特征表示。(latent feature, 由stable diffusion提出)



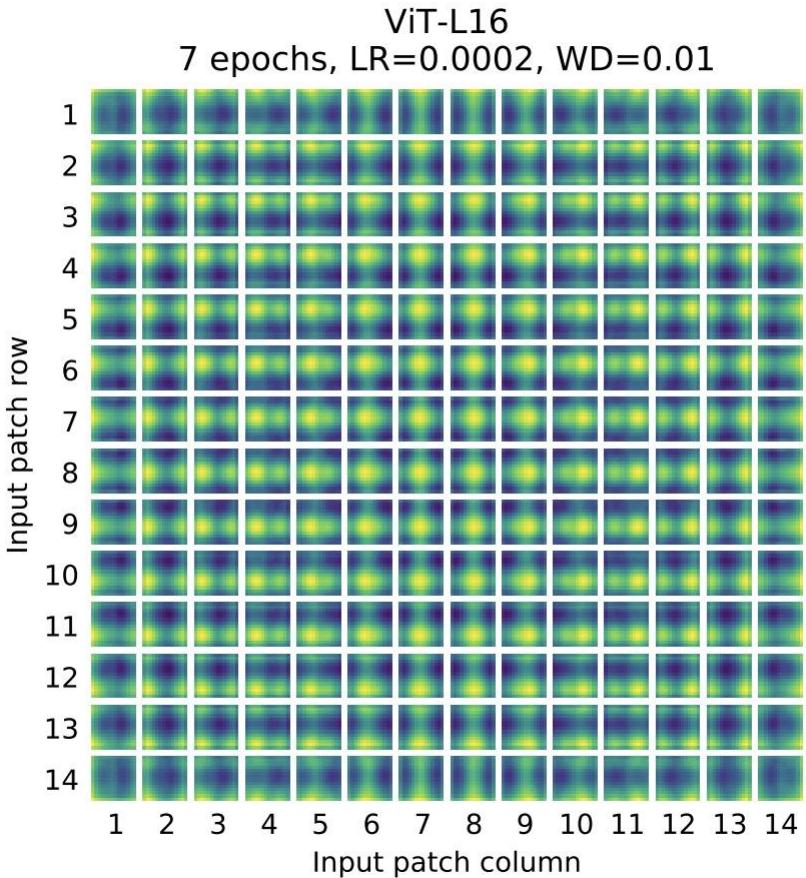
- 在这个压缩的潜在特征表示上，进行训练来生成视频。
- 利用图像解码器模型，将潜在特征再映射回像素空间。
- 最终生成高精度的视频。

ViT模型 -> DiT 扩散模型

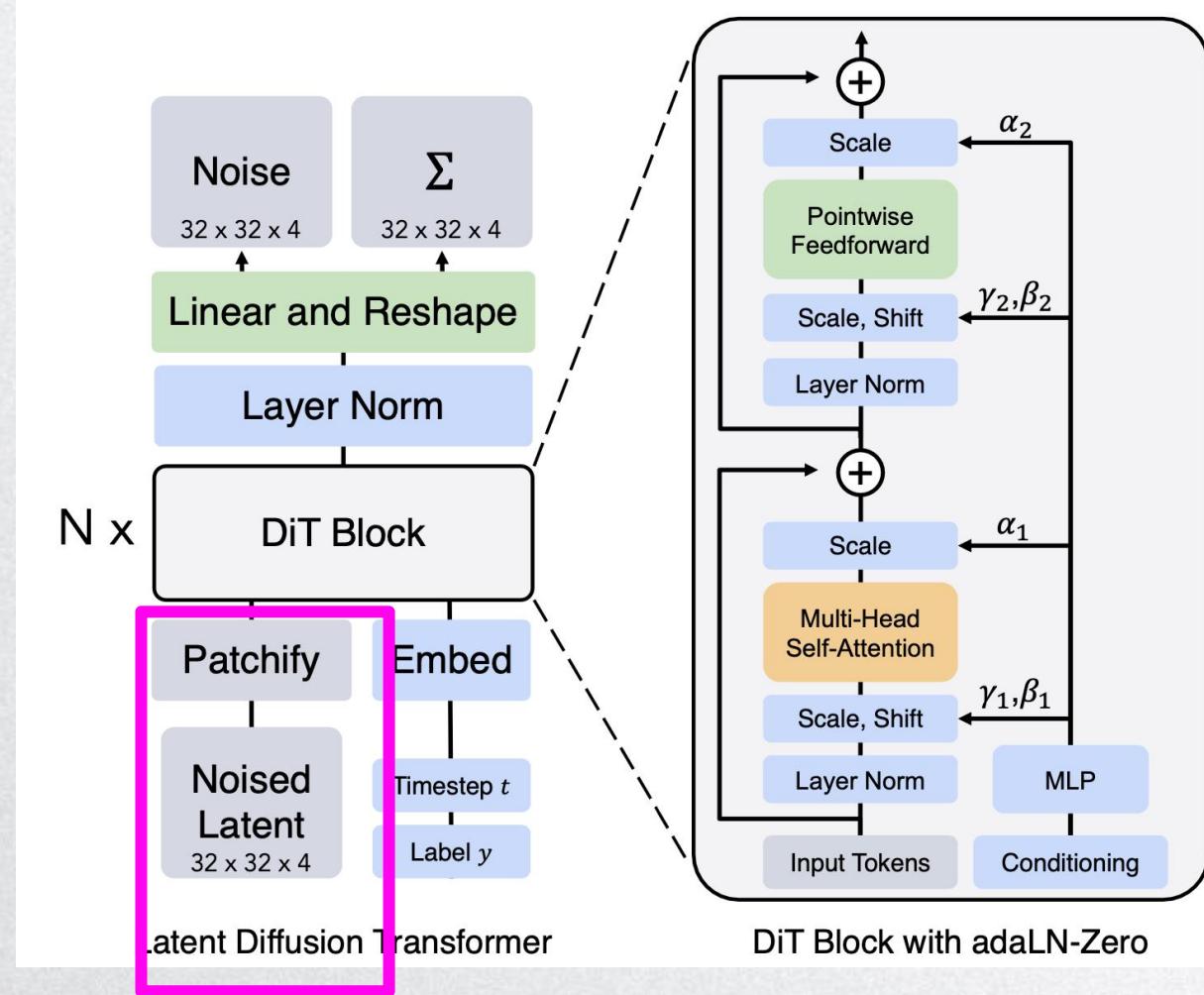
- DiT - [Scalable Diffusion Models with Transformers](#)
- 主干网络采用 ViT 结构
- 解码部分试用了扩散模型，即学习将token展开成像素

位置编码

- 1-D 位置编码: 例如 14×14 个 patch, patch 编码为 1 到 14×14
- 2-D 位置编码: 例如 14×14 patch, 分别编码 (X, Y) 再拼接起来
- 两者实际效果差别很小, 但比不用 pe 显著要好

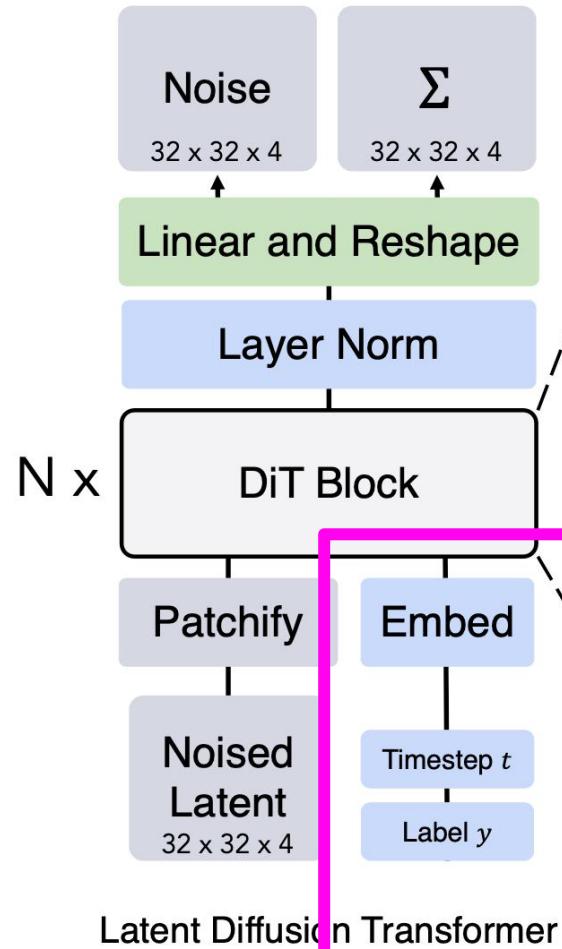


- 整个模型是参数化的去噪音模型
- 基于 Transformer 的 DiT Block 对输入的各 patch 进行自注意力编码
- 最终输出噪音
- 详情见 Diffusion 模型，最终预测结果期望为高斯噪声

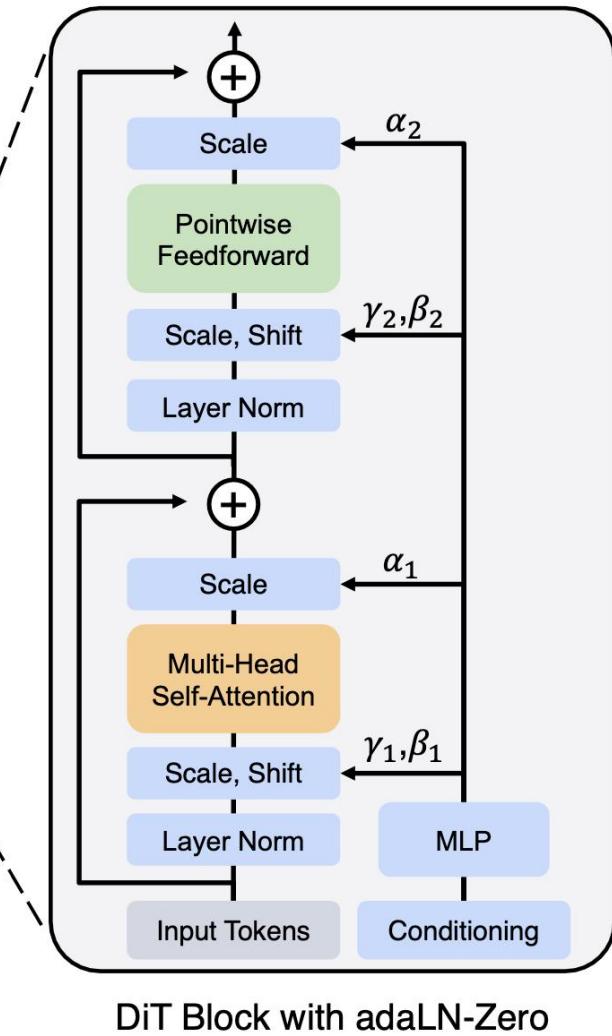


Conditional DiT 模块

- 类似 Conditional-GAN
- 对标签 y 进行嵌入
- 可以控制生成的图像标签



Latent Diffusion Transformer



DiT Block with adaLN-Zero



Figure 16. Uncurated 512×512 DiT-XL/2 samples.
Classifier-free guidance scale = 4.0
Class label = "husky" (250)

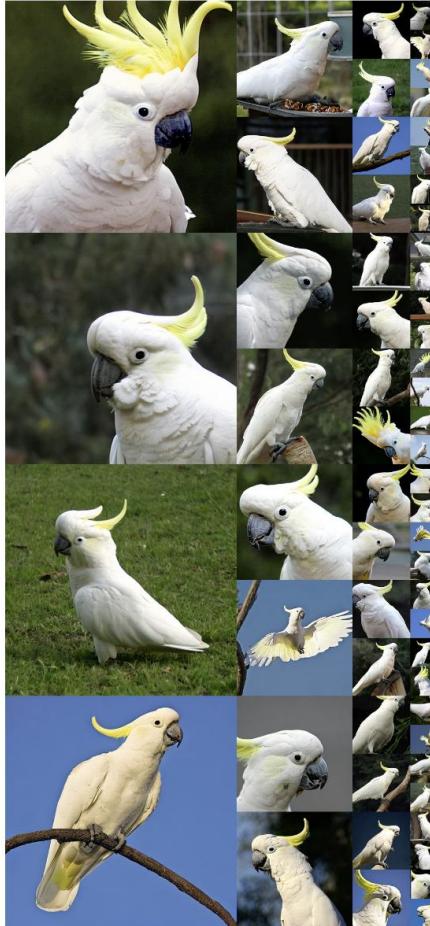


Figure 17. Uncurated 512×512 DiT-XL/2 samples.
Classifier-free guidance scale = 4.0
Class label = "sulphur-crested cockatoo" (89)

DiT forward() 函数

```
def forward(self, x, t, y):
    """
    Forward pass of DiT.

    x: (N, C, H, W) tensor of spatial inputs (images or latent representations of images)
    t: (N,) tensor of diffusion timesteps
    y: (N,) tensor of class labels
    """

    x = self.x_embedder(x) + self.pos_embed # (N, T, D), where T = H * W / patch_size
    t = self.t_embedder(t) # (N, D)
    y = self.y_embedder(y, self.training) # (N, D)
    c = t + y # (N, D)

    for block in self.blocks:
        x = block(x, c) # (N, T, D)

    x = self.final_layer(x, c) # (N, T, patch_size ** 2 * out_channels)
    x = self.unpatchify(x) # (N, out_channels, H, W)

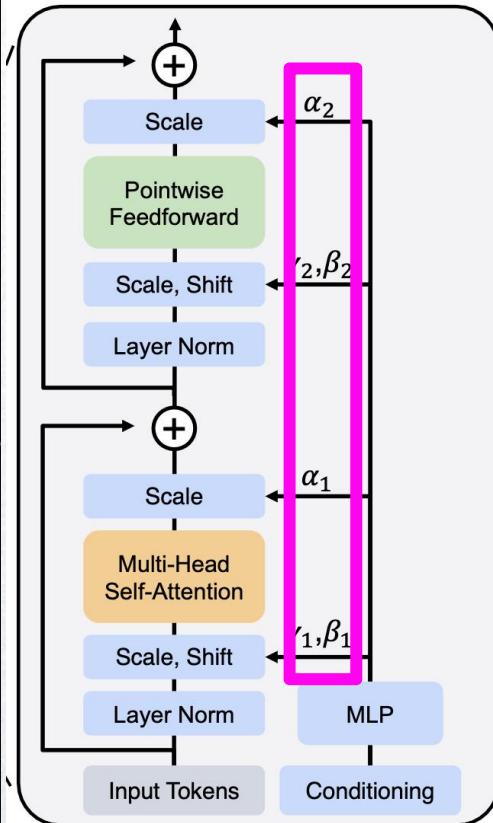
    return x
```

DiT Block 定义

```
class DiTBlock(nn.Module):
    """
    A DiT block with adaptive layer norm zero (adaLN-Zero) conditioning.
    """

    def __init__(self, hidden_size, num_heads, mlp_ratio=4.0, **block_kwargs):
        super().__init__()
        self.norm1 = nn.LayerNorm(hidden_size, elementwise_affine=False, eps=1e-6)
        self.attn = Attention(hidden_size, num_heads=num_heads, qkv_bias=True, **block_kwargs)
        self.norm2 = nn.LayerNorm(hidden_size, elementwise_affine=False, eps=1e-6)
        mlp_hidden_dim = int(hidden_size * mlp_ratio)
        approx_gelu = lambda: nn.GELU(approximate="tanh")
        self.mlp = Mlp(in_features=hidden_size, hidden_features=mlp_hidden_dim, act_layer=approx_gelu)
        self.adalN_modulation = nn.Sequential(
            nn.SiLU(),
            nn.Linear(hidden_size, 6 * hidden_size, bias=True)
        )

    def forward(self, x, c):
        shift_msa, scale_msa, gate_msa, shift_mlp, scale_mlp, gate_mlp = self.adalN_modulation(c)
        x = x + gate_msa.unsqueeze(1) * self.attn(modulate(self.norm1(x), shift_msa, scale_msa))
        x = x + gate_mlp.unsqueeze(1) * self.mlp(modulate(self.norm2(x), shift_mlp, scale_mlp))
        return x
```

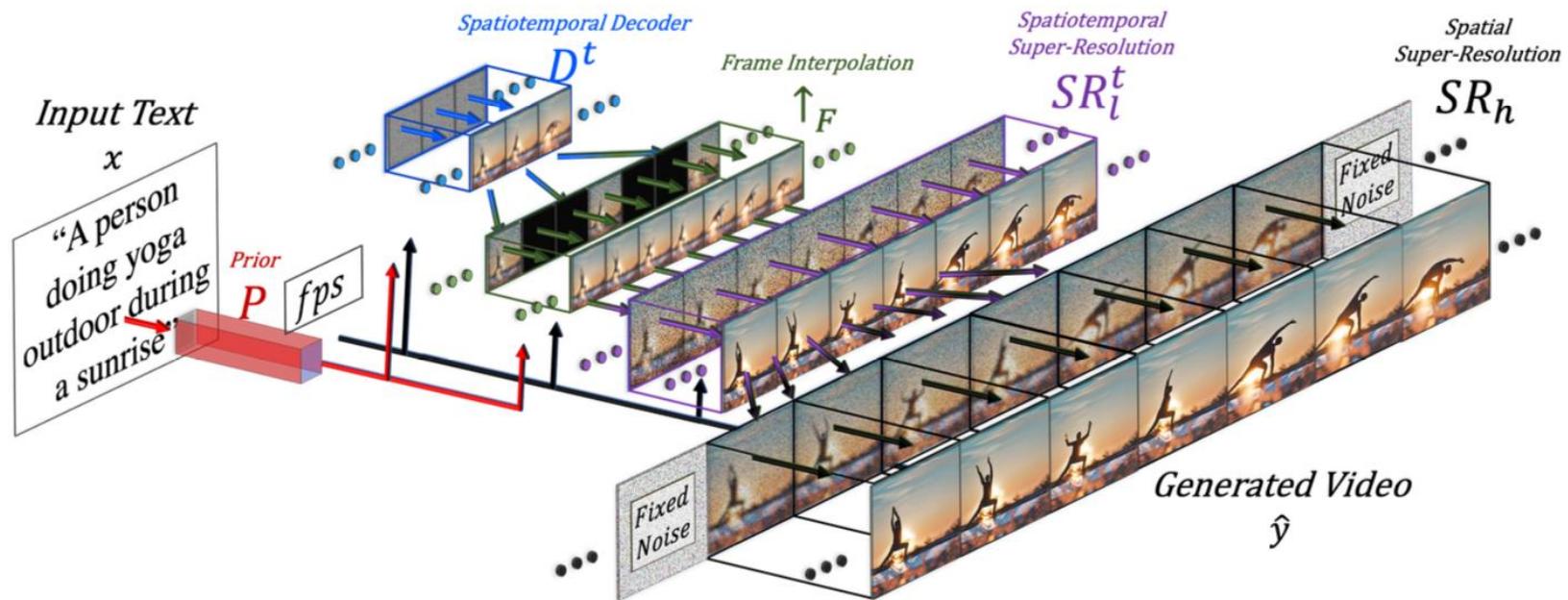


DiT Block with adaLN-Zero

Sora 参数估计和成本分析

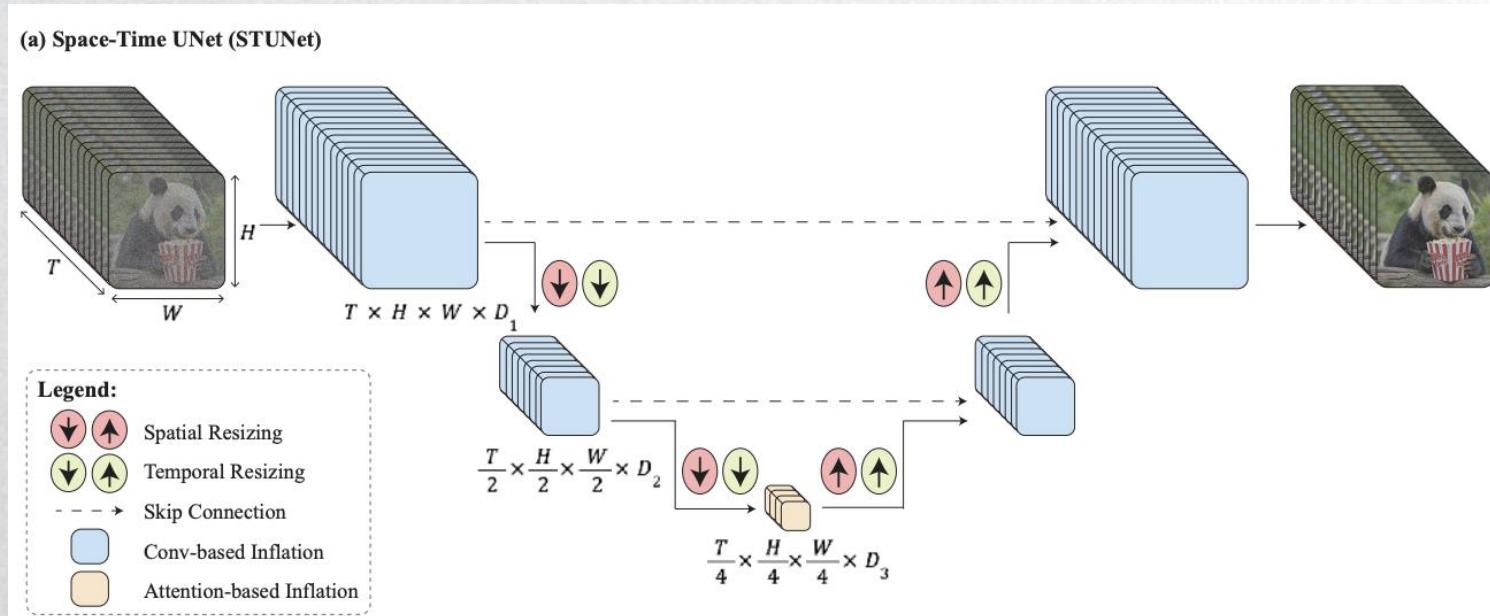
“It's far better to buy a wonderful company at a fair price than a fair company at a wonderful price.” -- Warren Buffett

传统方法 - 生成视频帧再进行插值采样



T2V模型(Google, Meta)

- 使用传统卷积网络对视频（图像流）先降维，再升维度
- 缺点：降维过程损失了较多的空间信息。升维的过程不能保证时空一致性。



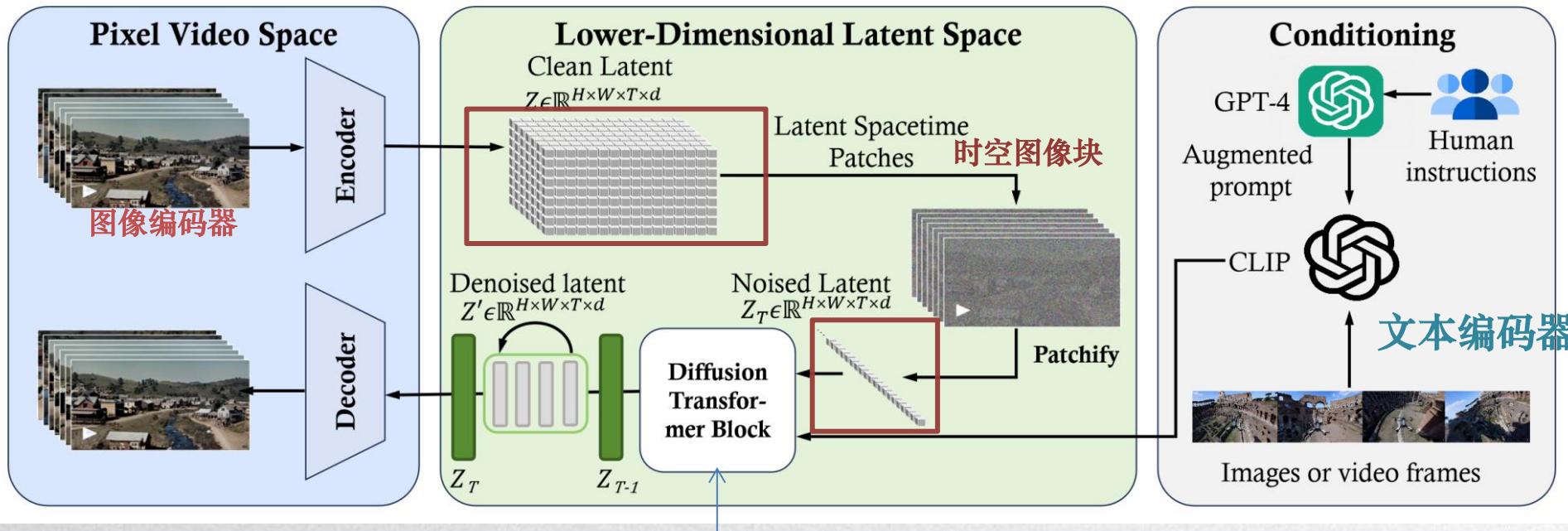
基于 T2V模型的参数

T2V模型 [Emu Video](#): 1.7B 参数模型

- 2.7B 预训练参数，**冻结不参与训练**，来自预训练的 [Text-to-Image](#) 模型，用于图像生成
- 3.0B 预训练参数，**冻结不参与训练**，用于 [Flan-T5-XL](#) 理解文本提示
- **1.7B 可训练参数**，用于增强时间维度的一致性

OpenSora - 压缩到低维隐空间再生成

Sora 是一个带有Transformer骨架的扩散模型：DiT = [VAE编码器 + ViT + DDPM + VAE解码器]。



DiT 模型，从图像块序列生成时空图像（视频）

Scalable Diffusion Models with Transformers <https://hpc-ai.com/blog/open-sora>

基于 时空图像块模型的参数

- 开源的Opensora 模型: 0.7B 参数 (700M参数)
- OpenAI 的Sora模型据推测: 3B 参数
- 视频生成的模型，去除预训练的语言模型之后，可训练参数并没有LLM（7B）那么大！
- 但是，它的计算量非常大！这是因为，视频被分解为大量的时空图像块。

大模型算力成本(软硬件)

估算：ChatGPT 输出量为3亿单词/分钟

硬件：1张H100单次运算，每秒钟能产生大约2万个单词

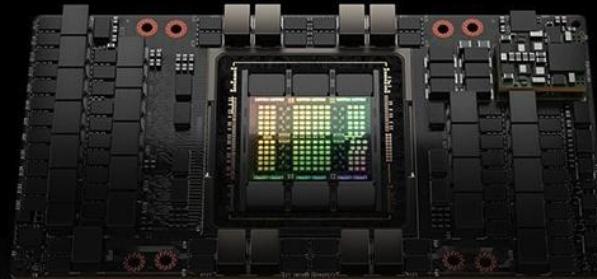
部署：3亿单词/分钟=> 2500块H100满足访问=> 7500万美元

\$30,000

NVIDIA H100 Tensor Core GPU

Extraordinary performance, scalability, and security for every data center.

[View Datasheet](#)



大模型目前的收费

Model	Pricing
gpt-4o	\$5.00 / 1M input tokens
	\$15.00 / 1M output tokens
gpt-4o-2024-08-06	\$2.50 / 1M input tokens
	\$10.00 / 1M output tokens
gpt-4o-2024-05-13	\$5.00 / 1M input tokens
	\$15.00 / 1M output tokens

生成一段1分钟的视频需要多少钱？

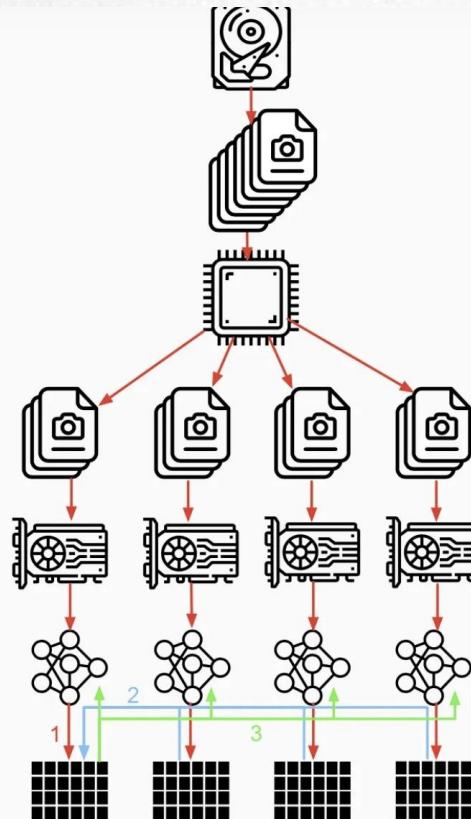
- 视频token数量 921,600 (0.9M)
 - 1分钟视频，总共有 $30\text{fps} * 60 = 1800$ 帧
 - 单帧图片 $1920 * 1280$ 分辨率，假设每个Patch尺寸为 $60 * 80$ ，一帧的画面有512个Patch
 - 文字提示 token 数量：1024（可忽略不计）
- 按照GPT4o的价格 1M token / 15美元
- **生成1分钟视频一共需要14美元！**

如果抖音全部是AI视频

- 按照抖音8亿日活，平均每人每天使用时长2h，对应每天16亿小时视频播放时长。
- 假设每个1分钟视频14美元，每天的成本为1.3万亿美元

优化策略：DP - 数据并行

- DataParallel
- 单进程多线程
- 模型权重在GPU:0上计算，再分发到其他GPU
- 显存和速度瓶颈



优化策略

- ZeRO (Zero Redundancy Optimizer) 大幅减少内存占用。每一张GPU仅保存部分的模型参数、计算部分的梯度，维护部分的优化器参数。
- 混合精度训练：训练过程中同时使用FP16（半精度浮点数）和FP32（单精度浮点数）两种精度的技术，大大减少内存占用。

优化策略 - Low Rank Adaptation (LoRA)

预训练的矩阵为 $W_0 \in \mathbb{R}^{d \times k}$, 它的更新可表示为:

$$W_0 + \Delta W = W_0 + BA, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

其中秩 $r \ll \min(d, k)$.

- 70% Params. Pre-training
- 30% Params. Fine-tuning
- 更容易在少样本上微调 !

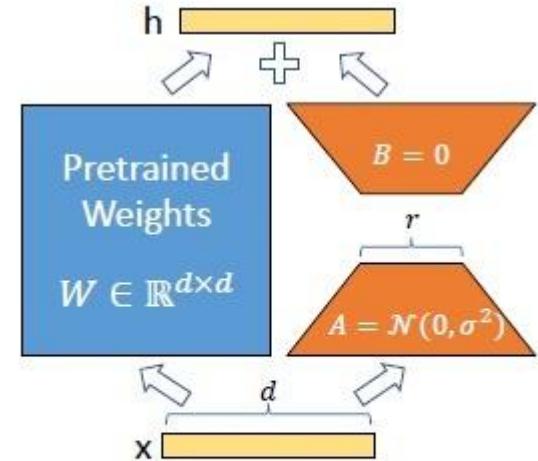
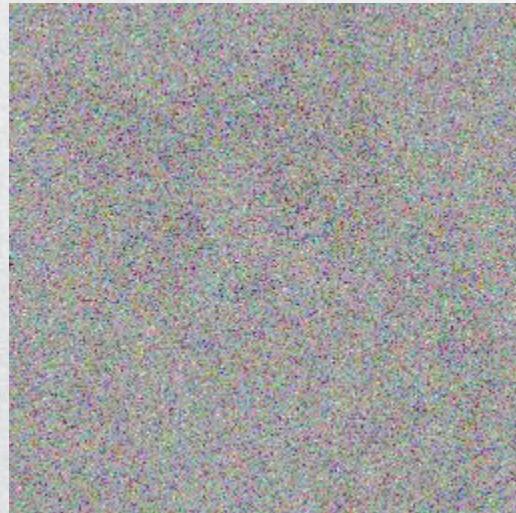


Figure 1: Our reparametrization. We only train A and B .

Sora 的语言理解部分

- 训练**文本到视频生成系统**需要大量带有相应文本字幕的视频。研究团队将 DALLE 3 中的重字幕（re-captioning）技术应用于视频。
- 具体来说，研究团队首先训练一个视频帧字幕生成器模型，然后使用它为训练集中所有视频生成文本字幕。
- 使用字幕-视频进行训练，可以提高文本保真度以及视频的整体质量。
- 研究团队还利用 GPT 将简短的用户 prompt 转换为较长的详细字幕，这使得 Sora 能够生成准确遵循用户 prompt 的高质量视频。

扩散模型 - 多模态生成大模型的基石



基础3：扩散生成(diffusion)模型

1. 什么是生成式人工智能 (Generative AI) ?

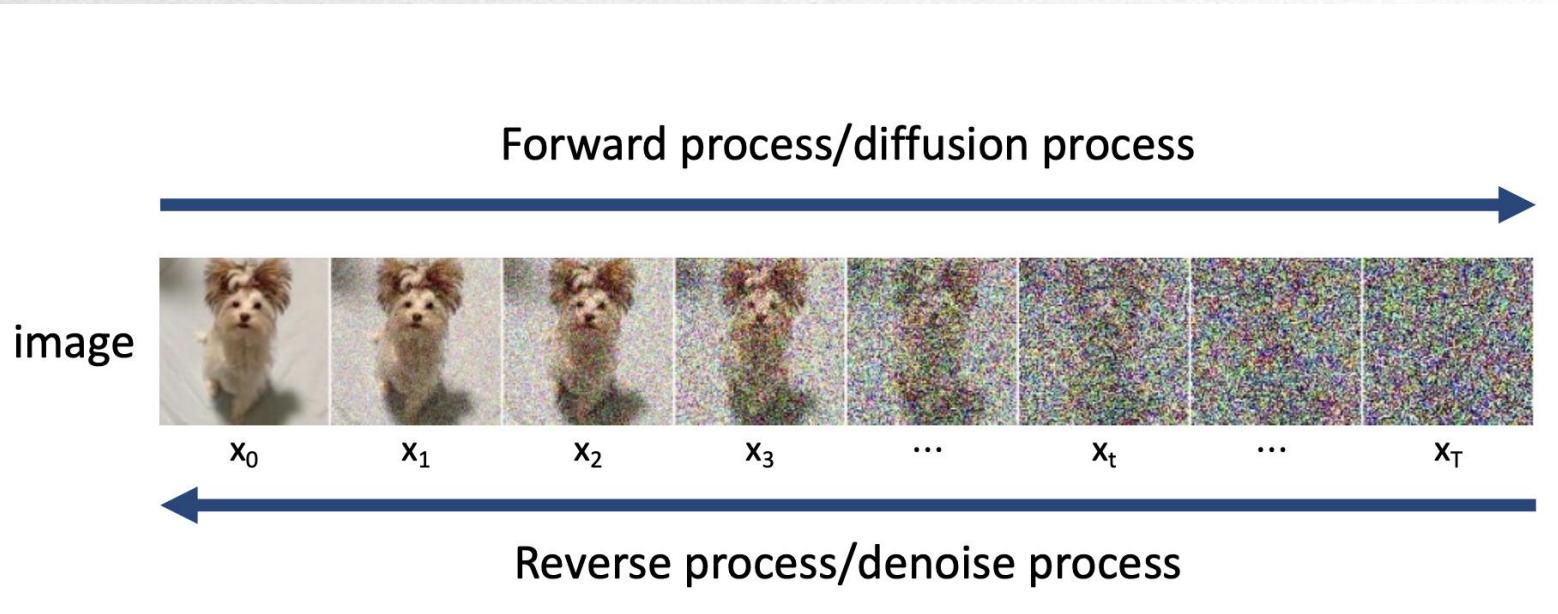
- 生成式人工智能是一种可用于创建新的内容和想法（包括对话、故事、图像、视频和音乐）的人工智能技术方法

2. 为什么压缩的时空图像块可以生成高精度的图片和视频?

- 依赖于扩散模型，图像块可以通过扩散过程逐渐生成像素空间的点，并且具有全局感知的能力，使得内容有意义

3. 什么是扩散过程/扩散模型?

DDPM (Denoising Diffusion Probabilistic Models)



Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS 2020.

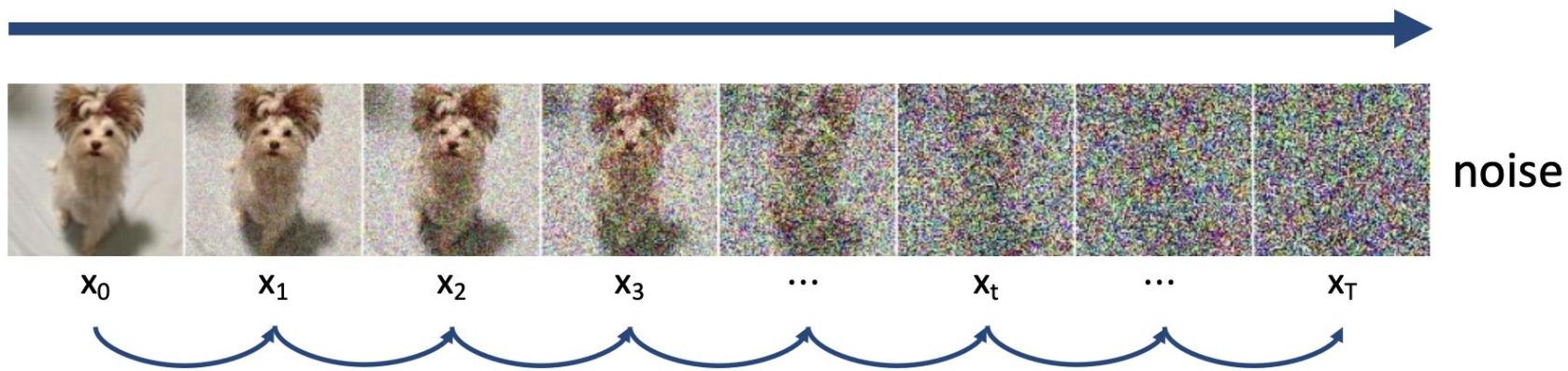
Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” ICML 2015.

Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” ICLR 2021.

“Tutorial: Video Diffusion Models.” CVPR 2024 [Tutorial](#).

正向过程：顺序地加入噪声，从 x_{t-1} 到 x_t

Forward process/diffusion process: add noise

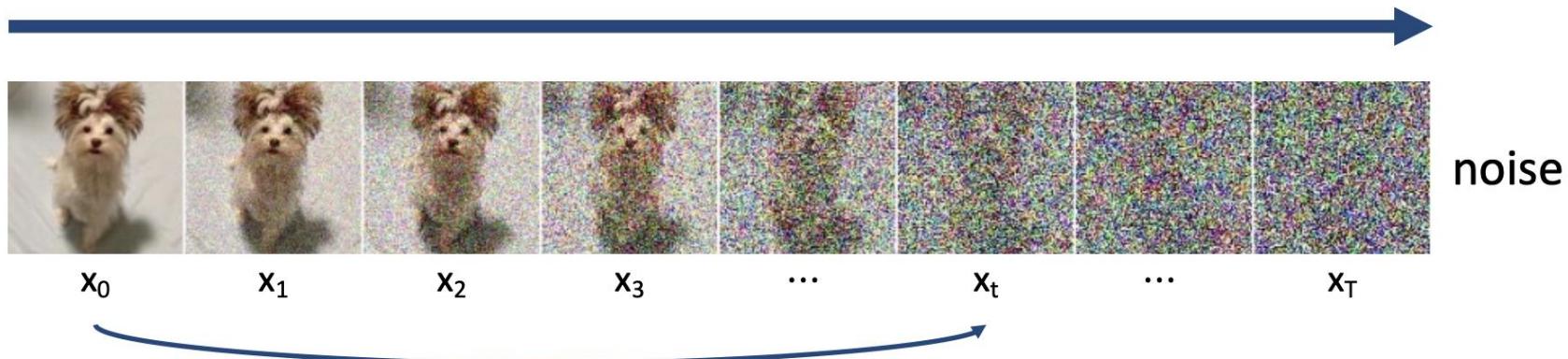


$$\Pr(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

β_t is the variance (strength) of noise

等价的，也可以直接从 x_0 到 x_t

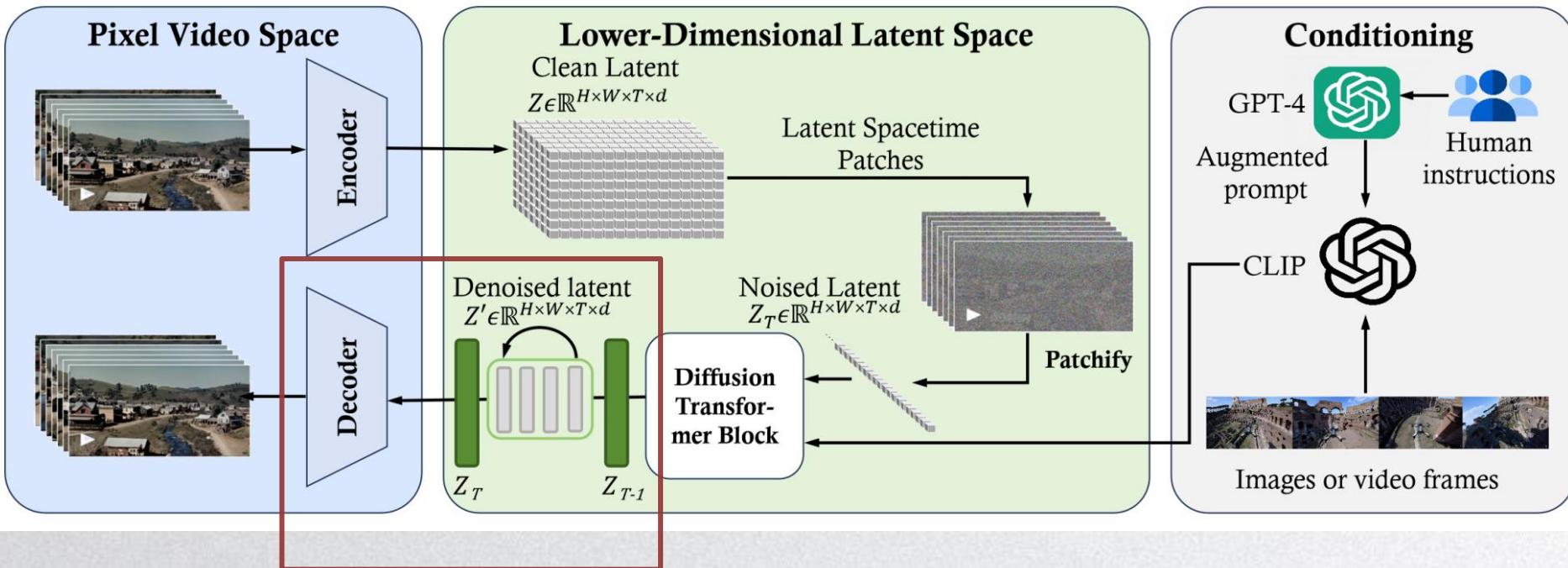
Forward process/diffusion process: add noise



$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$$

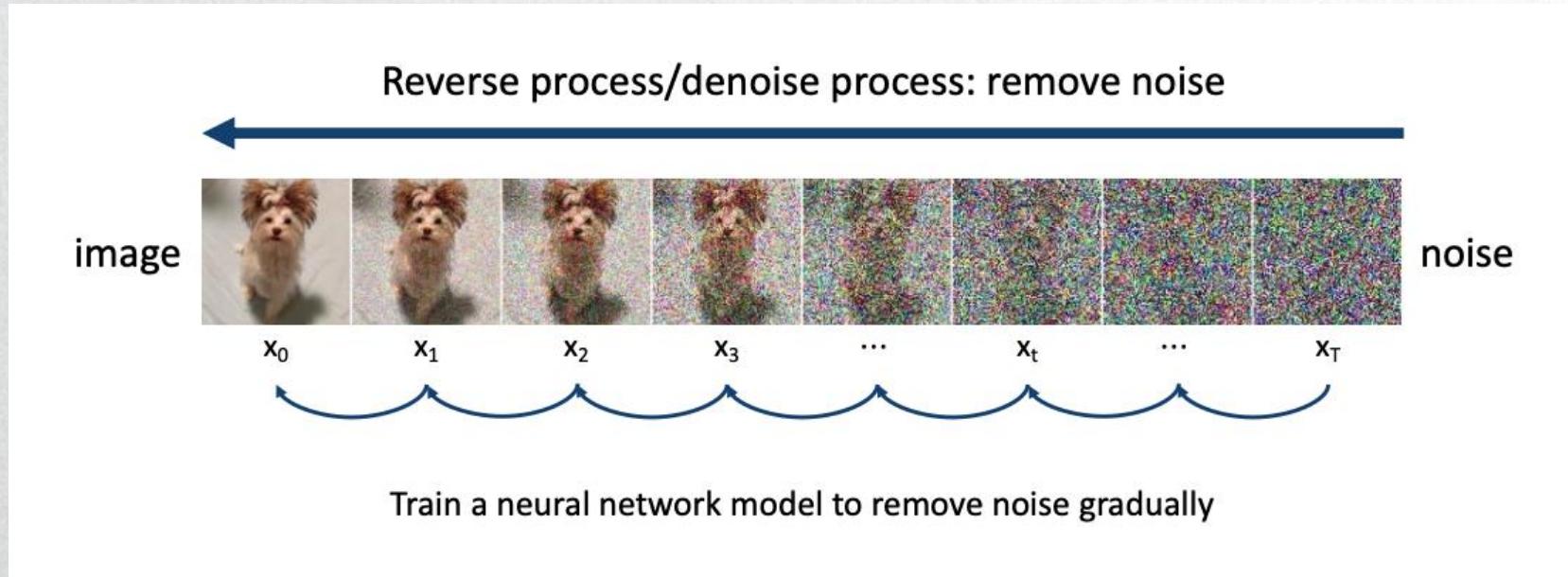
OpenSora - revisit again



扩散模型为什么要加噪声呢？

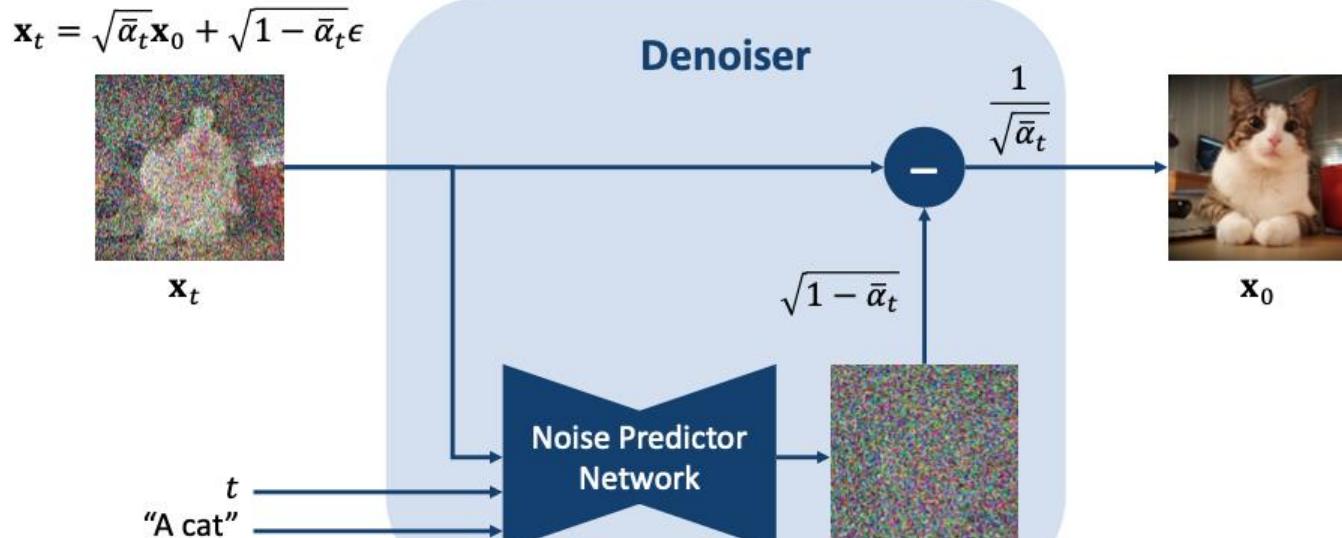
- 因为我们想要训练模型，去噪 (denoise) 的能力
- 训练的目标：从有噪音的图片里面，训练逐步去噪的能力
- 也就是说，我们需要预测噪音！
- Training objective: one-step predict the noise w.r.t. the original image
- 这样，就可以从含有噪音图片（甚至是完全的噪音）

反向过程：顺序地反向消除噪声，从 x_t 到 x_{t-1}



如何去噪：学习一个神经网络来预测噪音

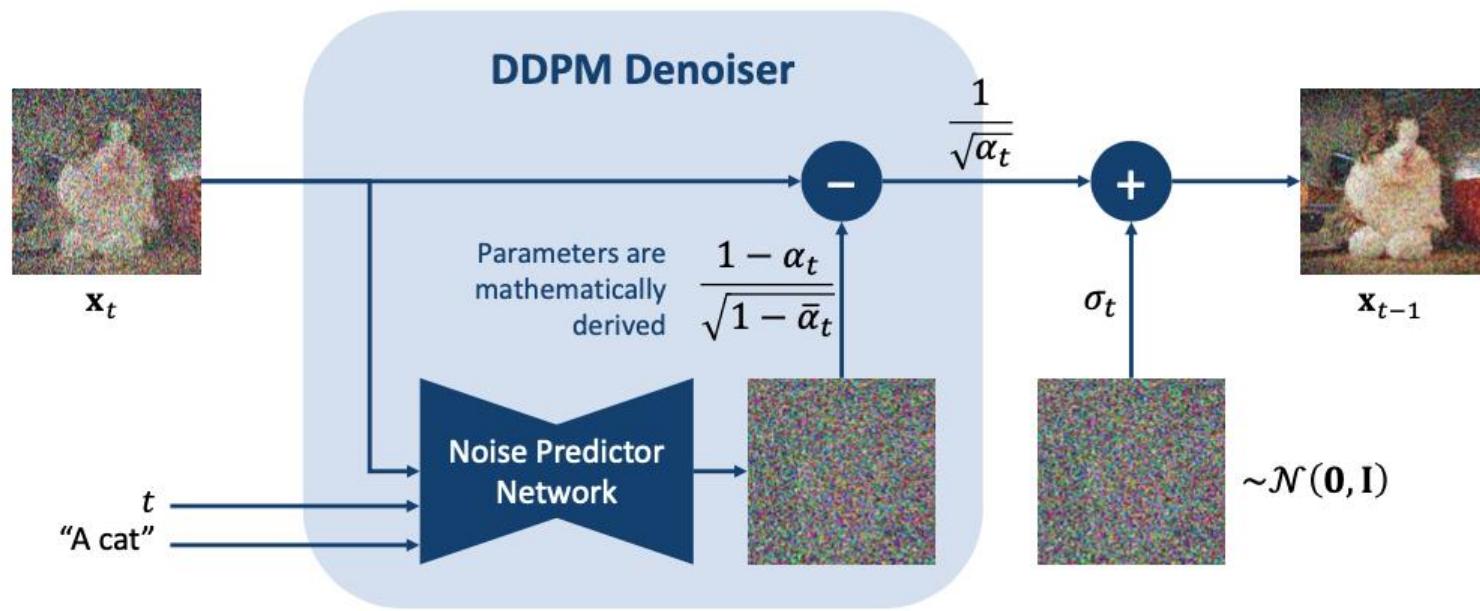
Training objective: one-step predict the noise w.r.t. the original image



训练时，利用原始高清图片 \mathbf{x}_0 ，可以训练这个网络预测噪音

如何去噪：学习一个神经网络来预测噪音

During generation: denoise step-by-step, in each step, add noise after noise removal



在推理时，用训练好的网络，可以对噪声去噪，生成高清图片

基于以上技术 Sora涌现出物理世界模拟能力

- Sora 模型表现出模拟物理世界中的人、动物和环境的某些方面的能力，也就是理解物理引擎中的世界模型。
- 例如：自动驾驶的很多 corner case 都是在类似游戏的仿真器里面模拟的，这样可以生成大量的现实中少见的数据。

- **三维一致性。** Sora 可以生成动态摄像机运动的视频。随着摄像机的移动和旋转，人物和场景元素在三维空间中的移动是一致的。



长序列连贯性和目标持久性。

- 长视频内容保持时间一致性。
- 即使人、动物和物体被遮挡或离开画面，Sora 模型也能保持它们的存在。
- 同一角色的多个镜头，在整个视频中保持相同外观。



与物体进行互动

- Sora 有时可以模拟以简单方式影响世界状态的动作。
- 例如，画家可以在画布上留下新的笔触，这些笔触会随着时间的推移而持续
- 一个人可以吃汉堡并留下咬痕。



艺术创作



应用举例

- 作为AI/CS专业的学生,如何提高自己的代码和写作水平。
- 作为商科管理类的学生，如何使用大模型进行文本分析，以及市场趋势、情感的分析。
- 作为艺术类的学生，如何将自己的灵感编辑成视频语音。

“编写一个乒乓小游戏”

作为一名代码新手，制作一个乒乓球小游戏是一个学习编程和游戏开发的绝佳项目。以下是你可能遵循的步骤：

1. 选择编程语言和开发工具

- Python + Pygame: Python是一种易于学习的语言，Pygame是一个用于制作简单2D游戏的Python库。
- Scratch: 这是一个面向初学者的图形化编程工具，适合快速上手游戏开发。
- JavaScript + HTML5 Canvas: 如果你对网页开发感兴趣，可以使用JavaScript在浏览器中制作游戏。

2. 设置开发环境

Python + Pygame

1. 安装Python: 从Python官网下载并安装最新版本的Python。
2. 安装Pygame: 打开命令提示符或终端，运行`pip install pygame`。

Scratch

- 直接访问Scratch官网，无需安装。可以在浏览器中编程。

JavaScript + HTML5 Canvas

- 你需要一个文本编辑器（如Visual Studio Code）和一个现代浏览器。

3. 学习基础知识

- 编程基础：变量、数据类型、条件语句、循环、函数等。
- 游戏开发概念：游戏循环、事件处理、渲染、碰撞检测等。

4. 设计游戏

- 游戏规则：两个玩家控制挡板，球在屏幕上反弹，未能住球的一方对手得分。
- 界面元素：球、左右板、右侧板、得分显示等。

5. 编写代码

以下以Python和Pygame为例：

导入库并初始化

```
python
import pygame
pygame.init()
```

Copy code

设置游戏窗口

```
python
screen = pygame.display.set_mode((800, 600))
pygame.display.set_caption("乒乓球小游戏")
```

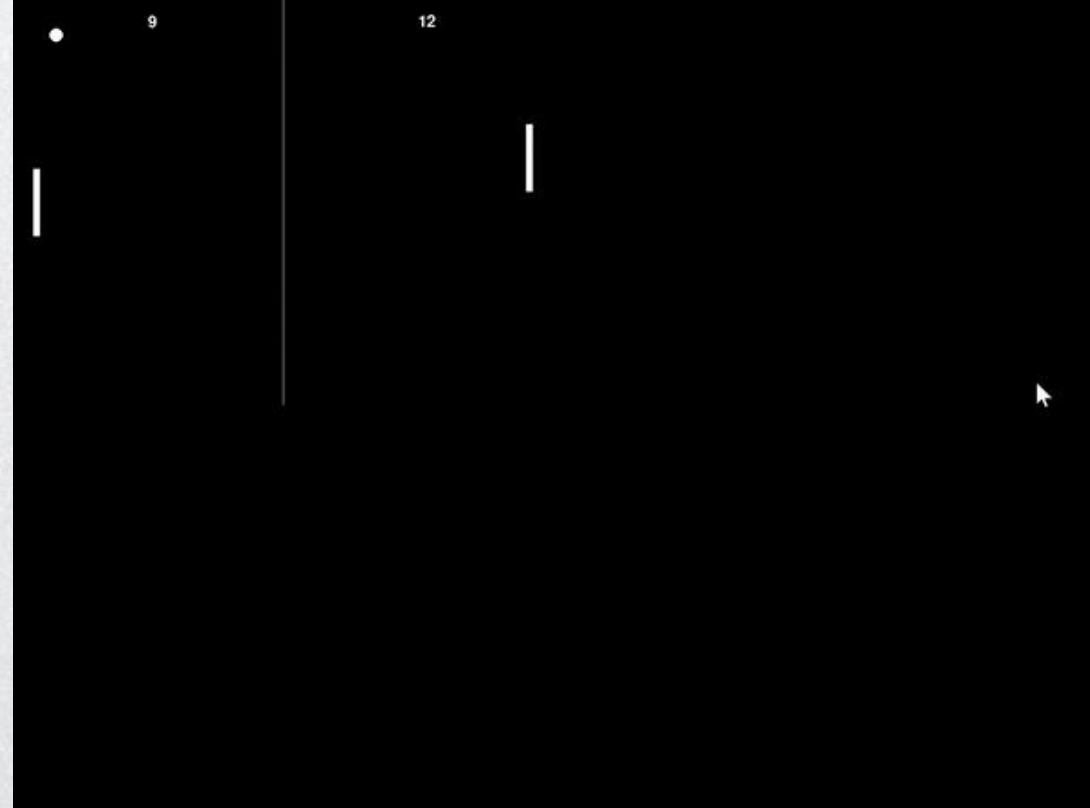
Copy code

定义游戏元素

```
python
# 白色
WHITE = (255, 255, 255)
BLACK = (0, 0, 0)

# 球
ball_pos = [400, 300]
```

Copy code



代码 debug

找出以下 python 代码的 bug:

```
a = [1,2,3,4]  
print(a[4])
```



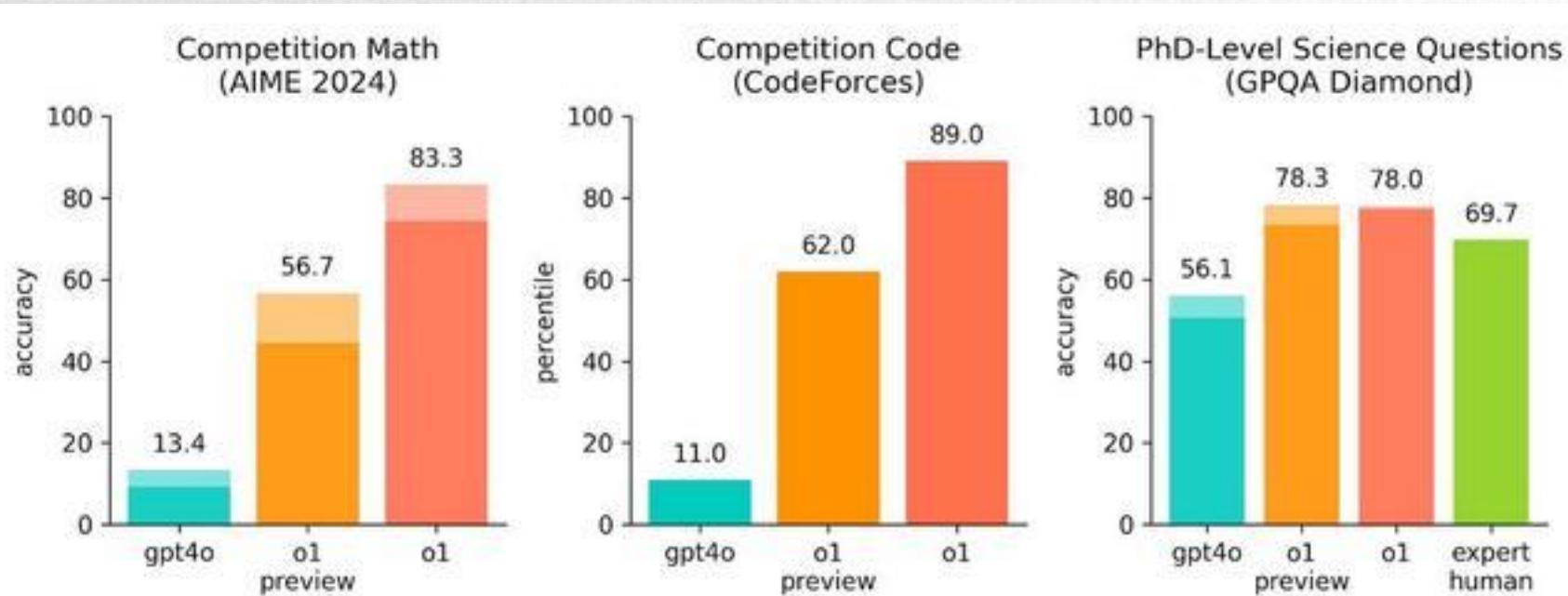
The bug in the Python code is that it tries to access an index that is out of range. The list `a` has elements at indices 0, 1, 2, and 3. When you attempt to access `a[4]`, there is no element at this index since the indexing starts from 0.

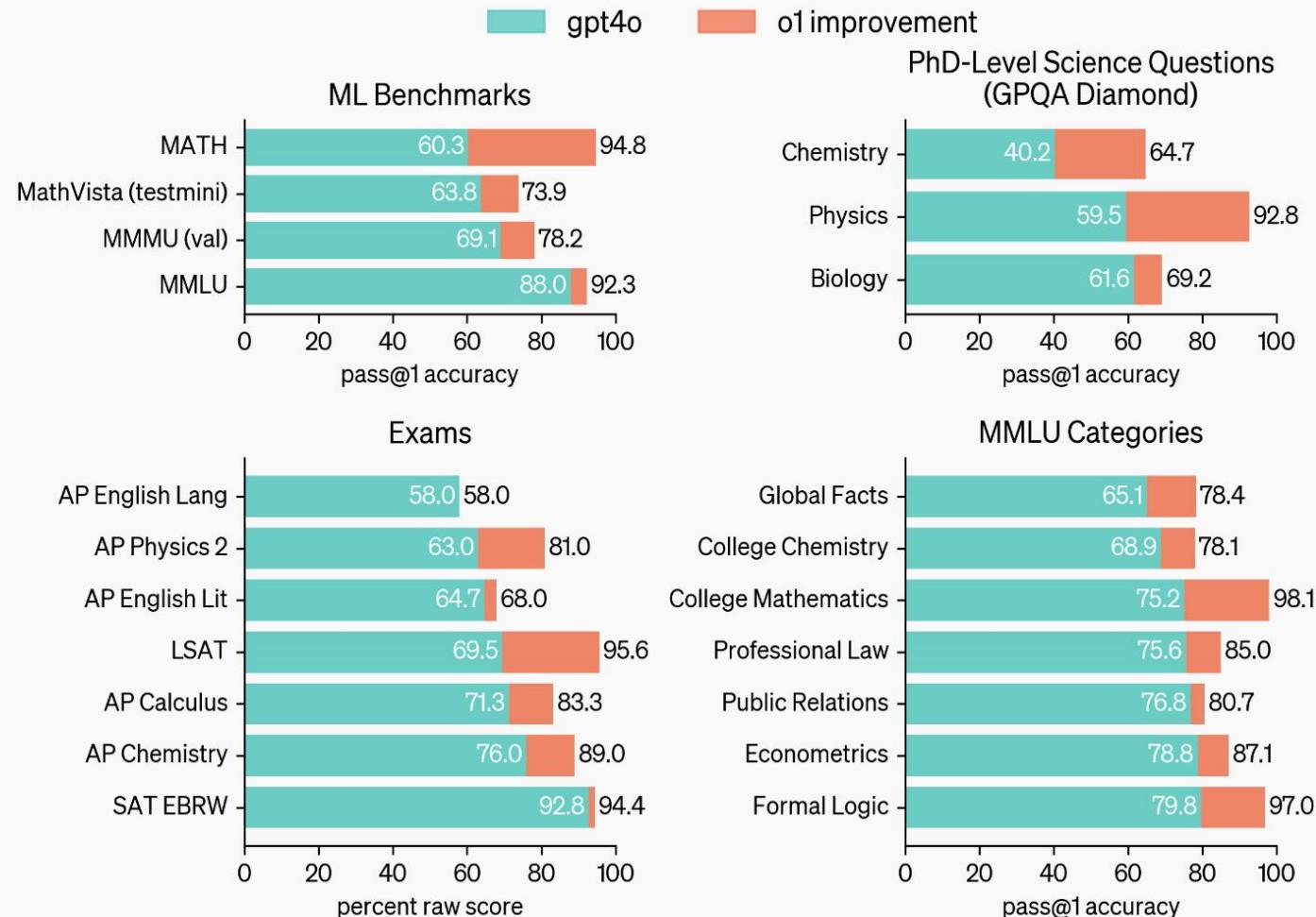
To fix the bug, you can either print the last element correctly by using the index 3, like this:

```
1 a = [1, 2, 3, 4]  
2 print(a[3])
```

STEM 数学竞赛、编程能力

- GPT-o1在编码、数学和各个科学领域的问题上，可能已经超越了普通人类的水平。







我用AI生成的图片，别人拿去卖钱，那么构成侵权吗？

法律问题解析



关于您使用AI生成的图片，如果别人未经授权将其用于商业目的（例如出售），这可能构成侵权。具体来说，以下几个方面可能需要考虑：

- 版权归属：**如果您是AI生成图片的创作者，并且您拥有该作品的版权，那么未经您的许可，他人不得擅自使用您的作品进行商业活动。
- 使用许可：**如果您在生成图片时使用了某些特定的软件或服务，需要查看该软件或服务的使用条款。有些服务可能会规定生成的作品版权归属问题。
- 侵权行为：**如果他人未经授权使用您的作品进行商业活动，这可能构成版权侵权。您可以采取法律手段维护自己的权益。

为了更好地保护自己的权益，建议您：

- 保留创作证据：**保存好创作过程中的相关记录，如创作时间、创作工具等。
- 明确版权归属：**如果使用了第三方平台或工具生成图片

修改英文语法



修改英文文法：He go to school early everyday.



He goes to school early every day.

字数：7字 | 输入tokens: 498 | 输出tokens: 8



Request ID: 757e4f44-2fbf-989b-9833-c76658390054

写一封请假信

写一封请假信，请假十天，去海南玩。限制200字。



尊敬的[主管/老师]：

您好！我拟于[具体日期]至[具体日期]期间请假，共10天，赴海南度假，以休养身心，提升状态。目前工作/学业已做妥当安排，确保交接无碍。期间保持通讯畅通，应对紧急事宜。归来后，愿以更饱满热情投入，必要时可加班补回。望批准，感荷。

敬祈理解，

[您的名字]

[当前日期]

字数：140字 | 输入tokens：375 | 输出tokens：101



Request ID: e0850fd3-662f-950c-9fb2-fec005e20285 [🔗](#)

文献检索



提供10篇，近十年来，人工智能研究最经典的论文列表。



近十年来，人工智能领域经历了快速发展，许多开创性论文不仅推动了技术进步，还深刻影响了行业应用。以下是2010年代至2020年代初部分具有代表性的经典论文：

1. "**Deep Residual Learning for Image Recognition**" (2015) - Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

引入了残差网络(ResNet)，解决了深度神经网络训练中的梯度消失问题，极大提升了深度学习模型在计算机视觉任务中的表现。

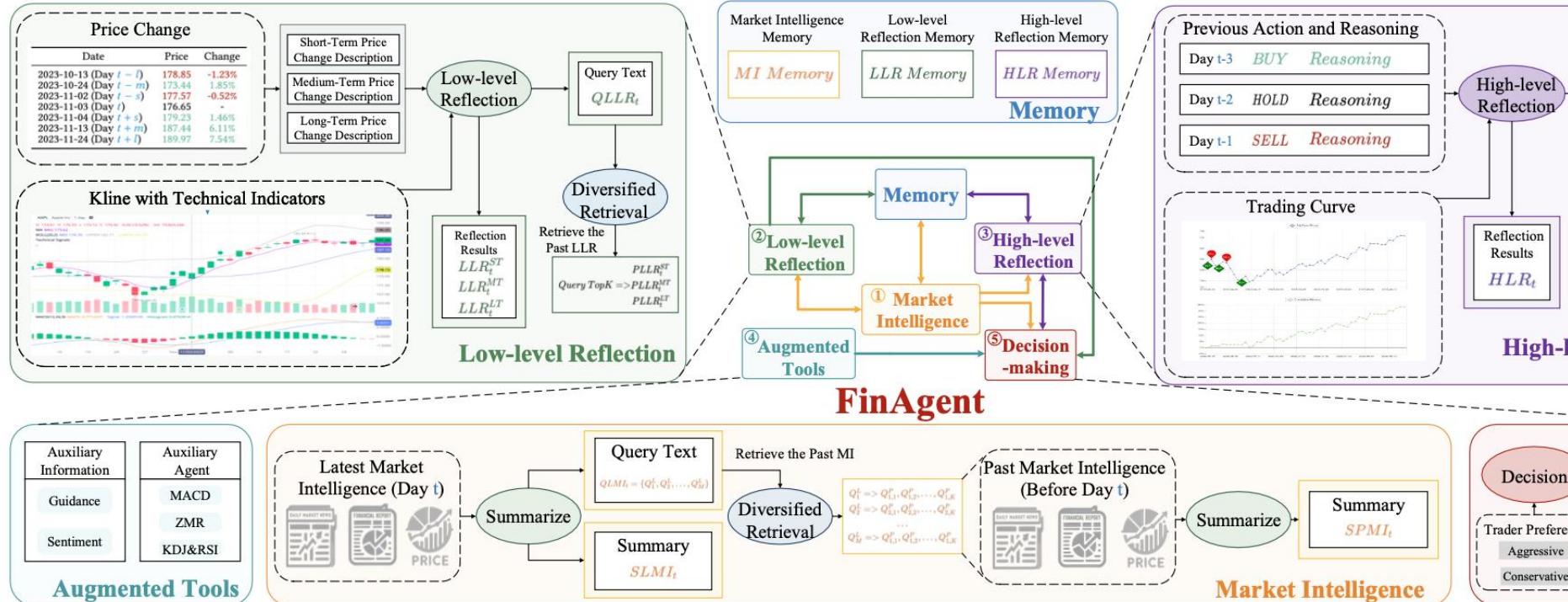
2. "**Generative Adversarial Nets**" (2014) - Ian Goodfellow, Jean Pouget-Abadie, et al.

首次提出了生成对抗网络(GANs)，为无监督学习和数据生成提供了革命性的框架，广泛应用于图像、视频、音频生成等领域。

3. "**Attention Is All You Need**" (2017) - Ashish Vaswani, Noam Shazeer, et al.

提出Transformer模型，通过自注意力机制彻底改变了自然语言处理领域，成为现代NLP模型如BERT和GPT系列的基础。

金融市场预测分析



A Multimodal Foundation Agent for Financial Trading.

公司业绩预测

证券研究报告：贵州茅台

核心观点

2023年，公司实现营收1476.94亿元，同比增长19.01%，归母净利润747.34亿元，增长19.16%。这一强劲的业绩表明，尽管市场环境充满挑战，公司依然能够通过优化销售策略和调整产品矩阵，持续增长和盈利。特别是，公司对直销渠道的重视促进了营收的显著增加，而在产品方面，系列酒的快速增长开始改变收入结构，为公司的持续发展注入新动力。这些成绩反映了公司不仅在行业中保持了领先地位，同时也拥有良好的市场竞争力。对于未来，公司通过进一步优化直销渠道和扩展产品矩阵的策略，展现出对增强市场竞争力和持续增长能力的明确信心。

业绩总览

公司发布了2023年的主要经营数据：营收达到约1476.94亿元，同比增长约19.01%；归属于母公司的净利润为747.34亿元，同比增长约19.16%。该数据显示公司在面临挑战的市场环境中继续展现出强劲的成长动力和盈利能力，稳健的经营业绩体现了其在行业中的领先地位和良好的市场竞争力。

FinRobot

<https://ai4finance.org/>

<https://github.com/AI4Finance-Foundation/FinRobo>

Report date: 2024-04-03

基本数据

评级	买入
目标价	1953.65 – 2600.0 (md. 2246.9)
近六月日平均成交量 (RMBmn)	0.03
收盘价 (RMB)	1715.11
总市值 (RMBbn)	2125499.24
52周价格范围 (RMB)	1555.55 – 1935.00

股价表现

Forecaster for Kweichow Moutai on 2024-01-28

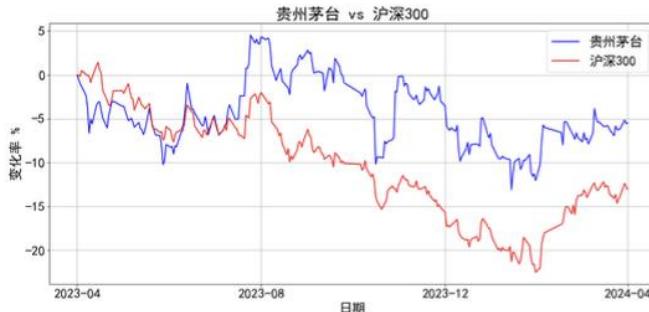
[积极发展]:

1. 贵州茅台的购买率持续上涨，表明市场对茅台的需求持续增长。**(News)**
2. 公司的新闻发布，如“美茅台美世界”2024全球品牌文化活动，表明公司继续追求国际化发展，并且将继续以“美”的国际表达，坚持追求美，实现看好未来的发展。**(News)**
3. 茅台的营业总收入同比增长率达到了19.80%，表明公司的经营状况良好，具有稳定的盈利能力。**(Basic Financials)**
4. 茅台的流动比率和速动比率都相对较高，表明公司的资金流动性较好，可以更好地应对市场的压力。**(Basic Financials)**

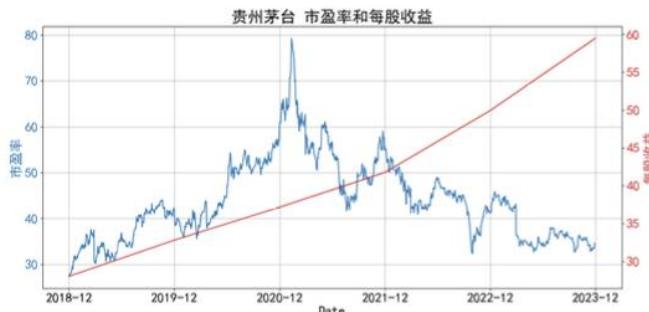
[潜在担忧]:

1. 茅台的资产负债率达到了17.98%，显示公司的负债压力较大，可能会对公司的盈利能力产生影响。**(Basic Financials)**
2. 公司未来可能面临的市场竞争压力可能会增加，如果公司无法继续提供高质量的产品和服务，可能会受到市场反应。**(News)**
3. 茅台的购买率上涨，可能会导致市场价格的偏高，可能会影响到公司的股价。**(Stock Price)**

股价表现



市盈率和每股收益



[预测和分析]:

预测涨跌幅: 上涨0-1%

总结分析: 虽然贵州茅台的资产负债率较高, 但公司的经营状况良好, 营业总收入同比增长率达到了19.80%, 显示公司在市场的需求和竞争中具有稳定的盈利能力。考虑到茅台的购买率持续上涨, 表明市场对茅台的需求持续增长, 可能会对公司的股价产生积极影响。同时, 公司的新闻发布, 如“美茅台美世界”2024全球品牌文化活动, 表明公司继续追求国际化发展, 可能会带来未来的发展机会。因此, 预计茅台的股价在下一周将上涨0-1%。

模拟数字世界

- Sora 还能模拟数字虚拟世界。
- Sora 可以通过基本策略同时控制 Minecraft 中的玩家，同时高保真地呈现世界及其动态。
- 只需在 Sora 的提示字幕中提及 Minecraft，就能激发这些功能。



Sora 局限性讨论 -- 技术层面

- 它不能准确模拟许多基本交互的物理现象，如玻璃碎裂。
- 不总能产生正确的物体状态变化。
- 官方主页列举了该模型的其他常见失效模式，例如长时间样本中出现的不一致性或物体的自发出现。
- 生成一分钟视频的成本及其昂贵，据估算单个视频就要10美元，包括GPU硬件和电费。



通过生成像素来对真实世界建模，未必是真正的通用人工之路



Yann LeCun   @ylecun · Feb 17 ...

Let me clear a *huge* misunderstanding here.
The generation of mostly realistic-looking videos from prompts *does not* indicate that a system understands the physical world.
Generation is very different from causal prediction from a world model.
The space of plausible videos is...

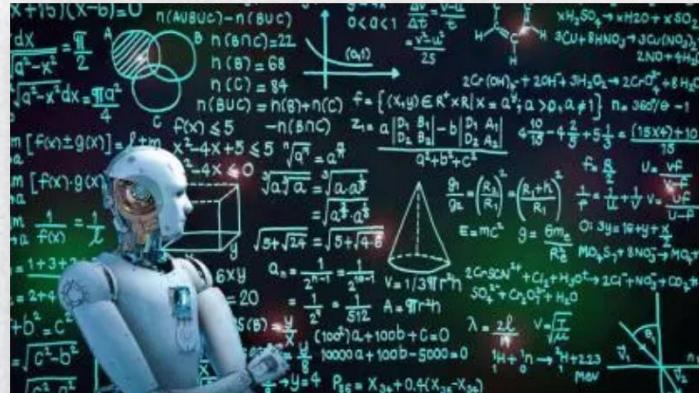
让我在这里澄清一个*巨大*的误解。
从提示中生成大部分看起来逼真的视频*并不*表明系统理解物理世界。
生成与世界模型的因果预测有很大不同。
似是而非的视频空间是.....

缺点和可能的负面影响 -- 从法律社 会

- 版权问题：艺术家的创作数据，被未经允许地拿来训练模型。
- 造假问题：目前一些所谓的“AI生成视频”，仍然需要大量人工修复和后期剪辑，让公众误导以为AI已经进展迅速。
- 内容问题：AI产生的虚假信息可能误导孩童，也可能产生严重法律问题，如虚假证据、暴力色情内容、政治问题。

模型滥用的担忧

- 输出虚假、不实的信息和知识
 - 论文引用和链接不存在
 - 专业领域的事实错误
- 数据泄露问题
 - 隐私泄露被AI爬取
 - 研究资料通过对话被套取
- 恶意内容生成
 - 病毒代码, “毁灭人类计划书”



国内首例“文生图”侵权案

- 原告李某某使用开源软件Stable Diffusion，通过提示词以及随机数种子等操作生成涉案图片，发布在小红书平台。
- 被告刘某某在百家号发布文章，文章配图使用了涉案图片。
- 北京互联网法院经审理认为，涉案图片符合作品的定义，属于作品；原告享有图片的著作权；被告应当承担侵权责任。
- 法院判决被告赔礼道歉并赔偿原告500元。

文生视频(Text-to-Video)的前景与未来

- AI生成的视频，能够在相当长的时空范围内，基本不违反物理世界的常见规律（比如重力、光电、碰撞等）。如果模型规模进一步提升，它有可能模拟生成物理世界的一切视频。
- 文生视频模型是大模型从侧重空间关联转向了加强时间关联，通过海量视频中对时空碎片的动态关联演化的学习。
- 未来或许人类可以将物理、化学、生物等等学科的范畴，可视化为图像或视频，交给视频生成大模型去学习，辅助发现其中蕴含的潜在规律。
- 集成更多的物理引擎、数学物理的定律，让模型在生成视频时能参考物理规则，提高物理互动的真实性。

人工智能的前景和未来

- 为什么AI模拟器不可以模拟巴以冲突、中美关系，模拟人类从山顶洞走向农耕文明的过程呢？
- “世界模拟器”通过模拟不同的事件和情景，预测未来的发展趋势，或可辅助决策制定，将深入到政治经济、人类社会等宏观层面。
- 也应该可以深入到病毒传播、交通规划等微观领域。这一切最终是否会改变各种学科研究的方式？
- AI这种不知疲倦的自我模拟和微调能力，是否会进一步自我进化，在模拟器中实现属于自己的闭环的世界呢？
- 还好，电源开关目前还掌握在人类手中。



AI 视频

KLING Creative Space

[文生视频](#) [图生视频](#)

💡 创意描述

一只大熊猫在湖边弹吉他

11 / 500 |

推荐尝试：油画少女 帆船 赛博城市 白猫 狮子



⚙️ 参数设置

创意想象力 创意相关性

生成模式

立即生成 (10 灵感值)

生成时长：



可灵AI



56 | 开会员，得灵感值



所有视频 ▾

视频生成中~
请 5min 后再查看

课后作业 -- 人人都是AI导演

请满足以下所有要求：

1. 在网络上搜索3-5种视频生成方法或者网站(如Runway，可灵)
2. 使用其中的2个方法或者网站，给同一段提示输入，生成视频
3. 该视频要求：输入自己的头像或照片，以自己为主角。输入一段prompt提示，用来生成一段故事或场景动作，如“我在宿舍开演唱会，我的室友都在给我加油”。时长不低于30秒。
4. 分析该视频的特点，列举该生成的视频是否达到了以下的效果，通过截图具体的视频帧，验证你的分析。

效果1：可以生成具有不同的长宽比的视频

效果2：你的照片或者头像，一直是视频的主角，没有发生明显变化。

效果3：视频中出现了多个人物，他们的动作也符合常识规律

效果4：视频中的背景前后一致，并且与你输入的prompt提示一致。



The background features a light gray textured surface with several overlapping circles of different sizes and colors: dark blue, white, and light gray. Some circles have a thin black outline, while others are solid. They are scattered across the frame, creating a sense of depth and motion.

THANKS

QA