



北京大学
PEKING UNIVERSITY

生成未必理解： 基于扩散模型能否实现视觉世界模型？

袁粒 Li Yuan

School of ECE, Peking University Shenzhen Graduate School



“What I cannot create, I do not understand”

— Richard Feynman

“What I can generate, I do understand”

“What I can understand, I do generate”

?



基于扩散的生成真的理解了吗？



跑步机上的人反着跑步，不符合逻辑

Prompt: Step-printing scene of a person running, cinematic film shot in 35mm.



杯子水先撒出来，后破碎，不符合事实

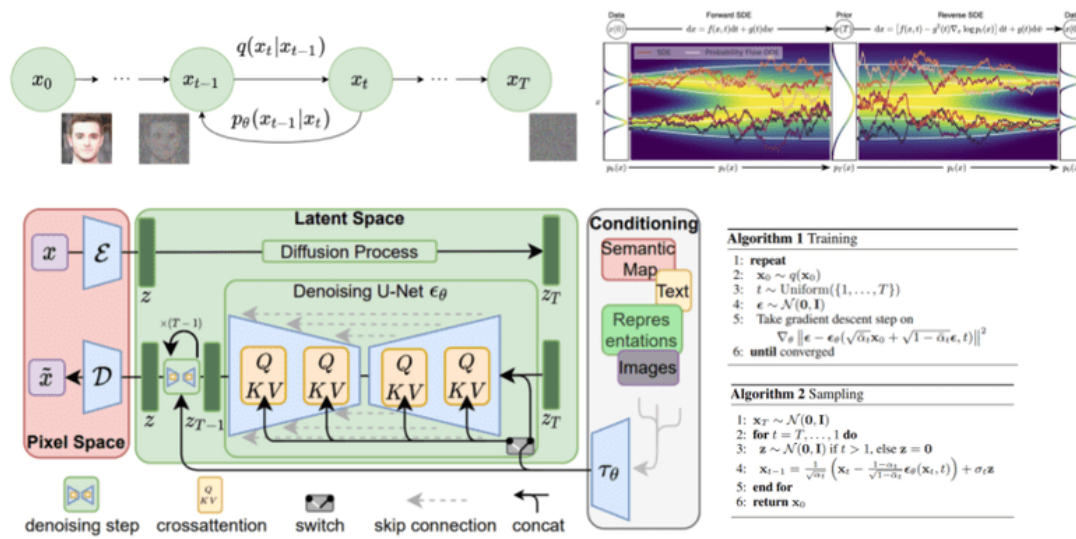
Prompt: Glass shattering with red liquid and ice cubes



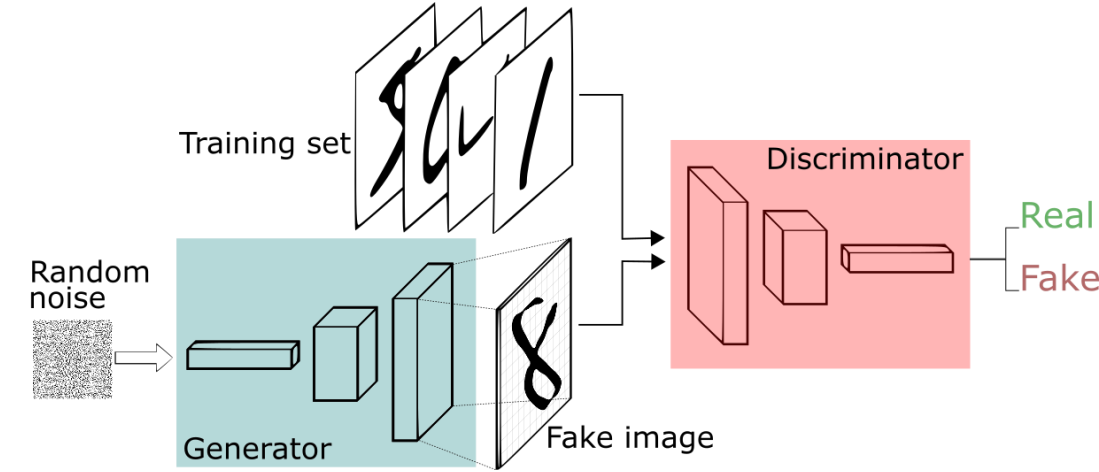
小狼的数量时而五只，时而三只或四只

Prompt: Five gray wolf pups frolicking and chasing each other around a remote gravel road

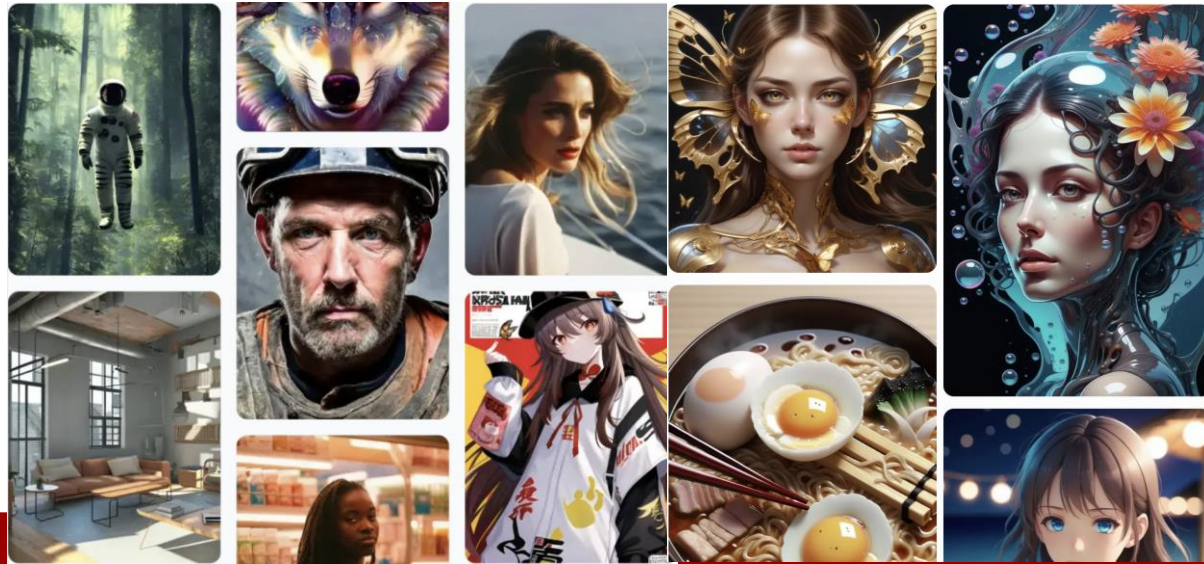
Diffusions Beat GANs



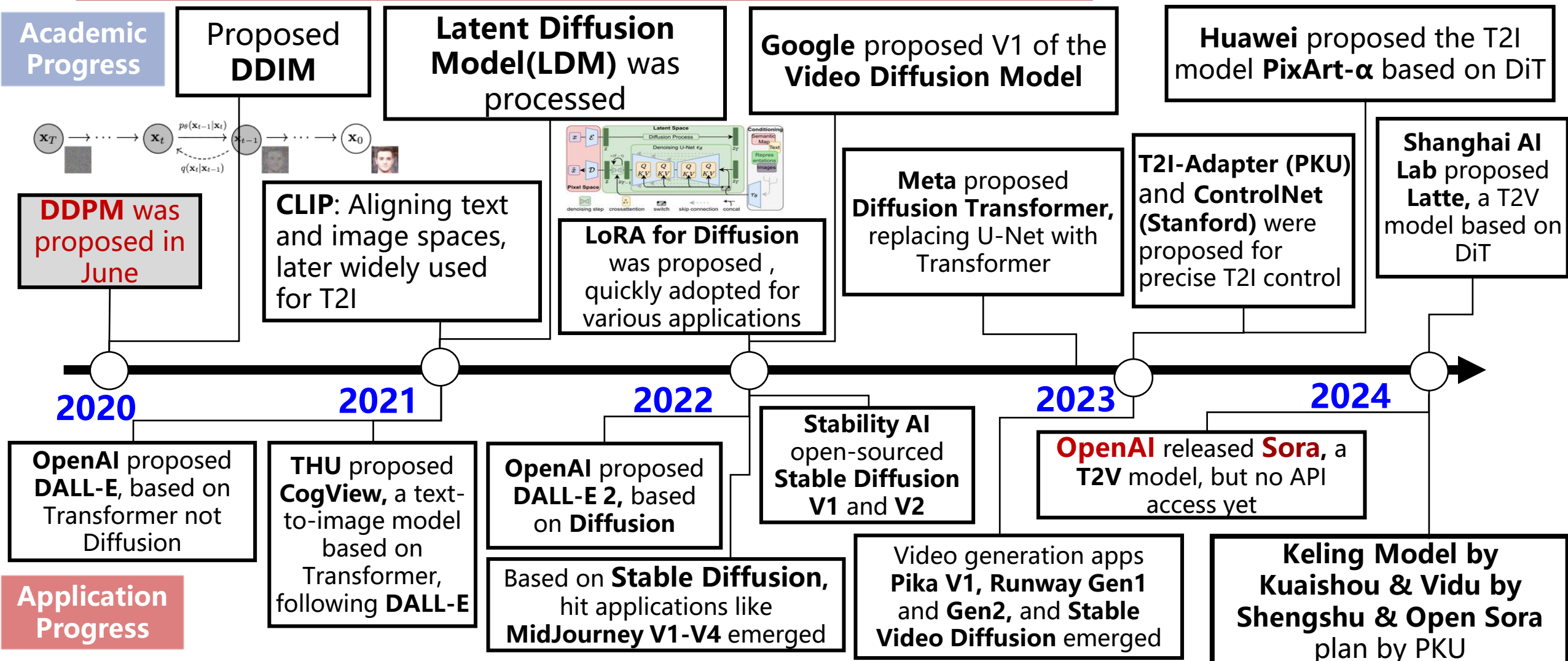
V.S.



V.S.

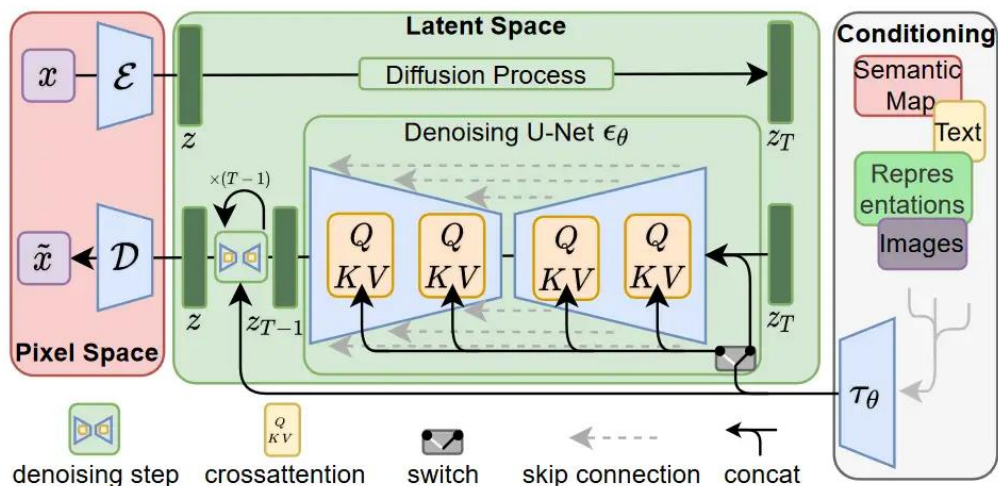


基于Diffusion的视觉生成发展

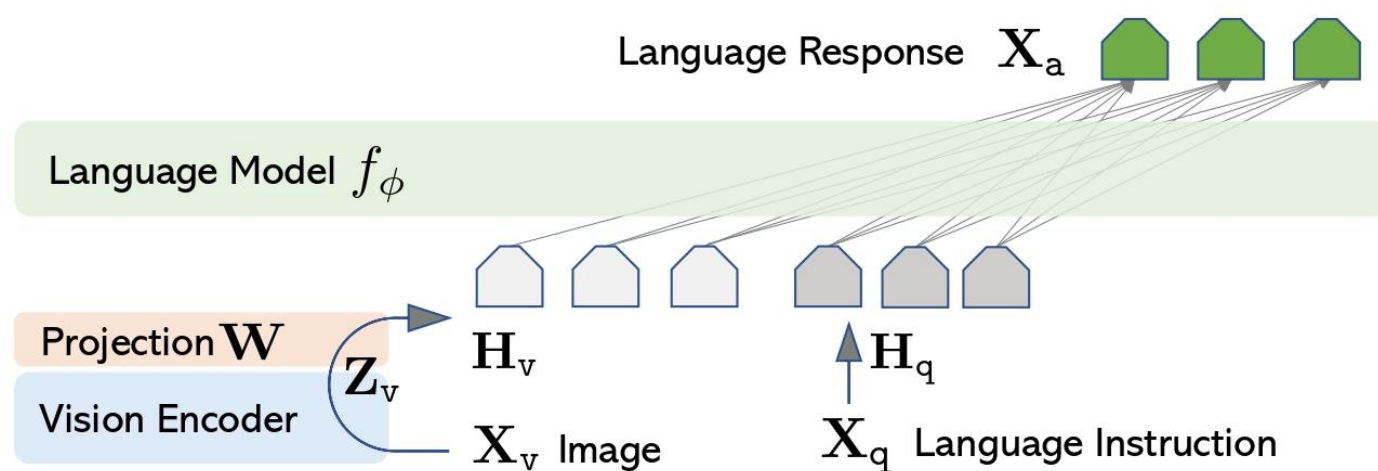


视觉生成和视觉理解两条路线完全割裂

视觉生成模型：Diffusion 系列



视觉理解模型：LLaVA系列



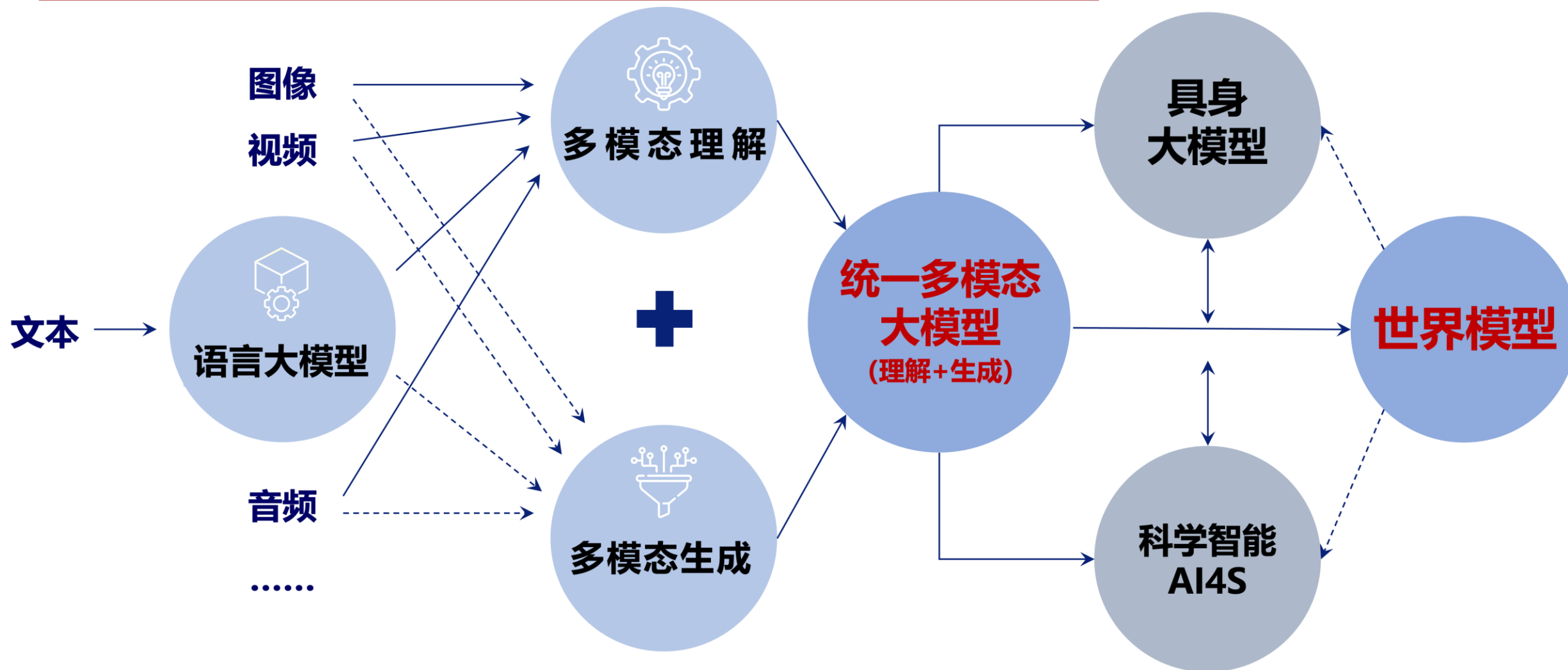
$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon} [\|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t, c)\|^2]$$

$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{y_i} [-\log P_{\theta}(y_i | y_{<i})]$$

- 生成未必理解、理解不能（视觉）生成；
- 建模方式不一样：视觉生成依靠扩散模型，视觉理解依靠LLM（自回归生成）；
- 模型参数规模不一样：视觉理解可以千亿规模，生成模型大部分数十亿到百亿；



世界模型的可能路径 — 统一是必然





世界模型的可能路径 —— 统一是必然

理解代表性工作:

- 多模态视觉理解Baseline: **Video-LLaVA**, EMNLP'24, 3.0k Star
- 多模态细粒度理解: Chat-UniVi, CVPR'24
- 多模态理解思维链: LLaVA-CoT, 1.6k star

生成代表性工作:

- 视频生成基模型: **Open-Sora Plan**, 11.7k star
- 物理知识注入的质变生成: MagicTime(1.3k star), ChronoMagic, NeurIPS'24
- 动态生成: Progressive123 (ICLR'24), Repaint123 (ECCV'24), ViewCrafter
- 多模态可控生成: ConsisID, 生成人物ID一致性

多模态

多模态



PKU-YUAN-Lab (袁粒课题组-北大信工)

Open codes from YUAN Lab at PKU

Overview

Repositories 26

Projects

Packages

People 17

Popular repositories

Open-Sora-Plan

Public

This project aim to reproduce Sora (Open AI T2V model), we wish the open source community contribute to this project.

Python 11.7k 1k

ChatLaw

Public

ChatLaw: A Powerful LLM Tailored for Chinese Legal. 中文法律大模型

7k 546

Video-LLaVA

Public

【EMNLP 2024🔥】Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Python 3.1k 220

MoE-LLaVA

Public

Mixture-of-Experts for Large Vision-Language Models

Python 2k 129

LLaVA-CoT

Public

LLaVA-CoT, a visual language model capable of spontaneous, systematic reasoning

Python 1.6k 61

MagicTime

Public

MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators

Python 1.3k 125

Open-Sora Plan 视频生成开源计划总览

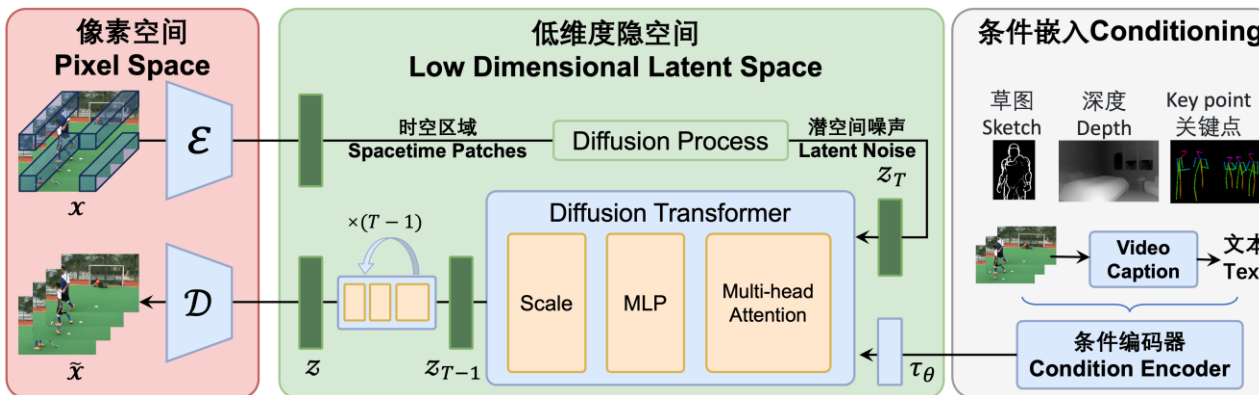


北京大学
PEKING UNIVERSITY

HOT 趋势

2月OpenAI发布了全新的文生视频模型Sora，率先开启了AI视频生成的ChatGPT时刻。AI视频生成催生了以多种创意需求，受到社会广泛关注。当前AI视频生成领域主要靠业界巨头的闭源模型推动，开源项目少。

模型结构



亮点

- 支持10s级高清视频生成
- 兼容文生图和文生视频
- 支持任意分辨率和任意时长
- 支持任意帧控制注入的图生视频及视频续写
- 借助昇腾基础软件与调优工具，充分释放硬件性能
- 高质量数据集处理及模型评测管线



发起视频生成开源计划

项目发布一周内获得14万余次访问，获得国内外上万名程序员点赞，登上GitHub Trending全球榜单第一，累计更新模型和代码数百次，融合开源社区的贡献上百次，代码被克隆上千次。



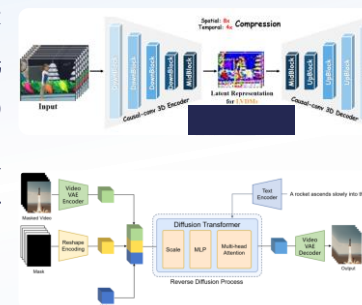
持续的项目活跃度

Open-Sora Plan项目持续更新近半年，累计新增代码8万余行，近期单周访问近3万次，收到社区和业界的持续关注，有多项后继的视频生成开源项目和闭源模型部分借鉴了我们的实现方式。



前沿的技术创新

为了实现高分辨率长时长的视频内容生成，团队先后提出了256倍压缩的3D因果VAE、任意分辨率及时长训练、稀疏注意力计算加速、任意帧控制的图生视频模型等技术。



借助昇腾基础软件与调优工具：充分释放硬件性能，快速支撑原生训练



北京大学
PEKING UNIVERSITY



TorchNPU

标准化接口，兼容开源生态

源码迁移仅需两行代码
多流内存复用优化内存
多级流水优化调度性能



MindSpeed

昇腾分布式训练加速库

兼容Megatron原生框架
内存Swap降低FA重计算开销
并行/内存/通信/计算全方位定制优化

CANN

昇腾异构计算架构

算子开发体系，图优化引擎

昇腾基础算子

DVPP/BLAS/NN/HCCl

昇腾高性能融合算子

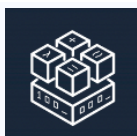
Flash Attention支持多种内存排布

去除冗余Transpose

AscendC算子自定义开发

原生开发RoPE3D高性能融合算子

基于Ascend C Debug能力快速调优



MindStudio Insight

昇腾Profiling工具套件

计算/通信/内存调优分析工具
Advisor自动化给出优化建议
实时监控训练进程，识别异常



MSProbe

昇腾精度工具分析套件

支持在线监控梯度异常
快速定位精度异常代码段
支持异构算力精度一致性对比

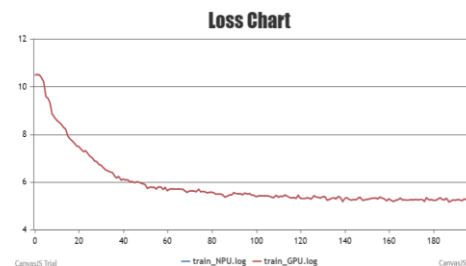


图1. Open-Sora Plan模型精度绝对值对比

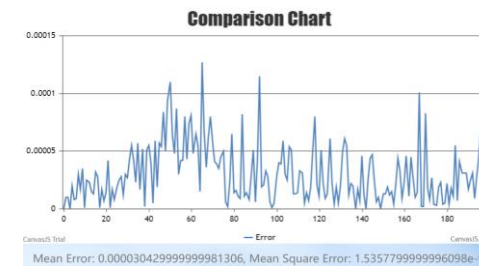


图2. Open-Sora Plan模型精度相对误差对比

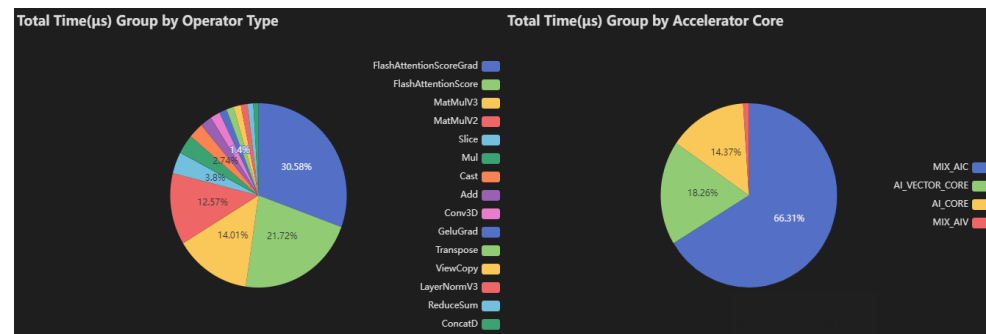
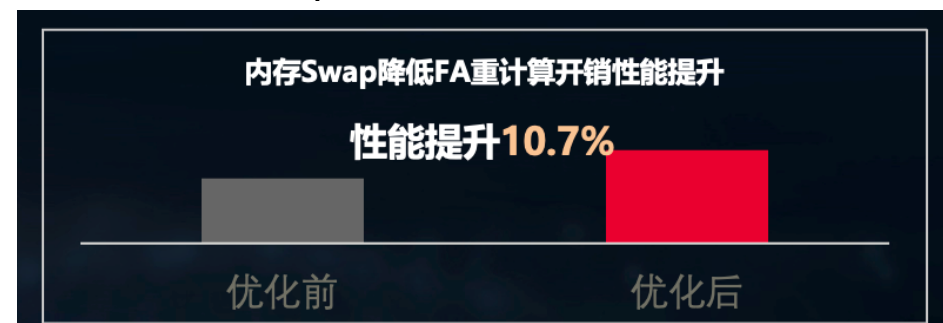


图3. Open-Sora Plan模型各算子耗时占比分析



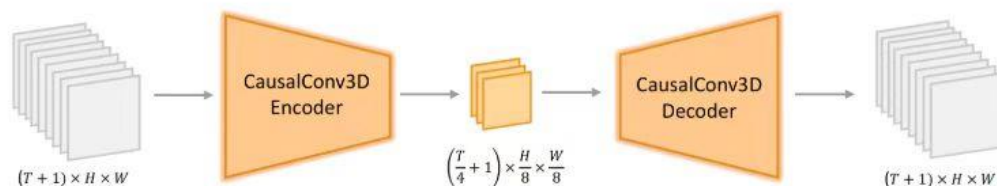
Open-Sora Plan 视频生成开源计划总览



北京大学
PEKING UNIVERSITY

3D VAE

HunyuanVideo trains a 3D VAE with CausalConv3D (Partially referred to [Open-Sora-Plan](#)) to compress pixel-space videos and images into a compact latent space. We set the compression ratios of video length, space, and channel to 4, 8, and 16 respectively. This can significantly reduce the number of tokens for the subsequent diffusion transformer model, allowing us to train videos at the original resolution and frame rate.



on, integrating more than 30 pull requests from models and codes more than 200 times within a

more than 140k visits in a single week

content
Views
Unique visitors

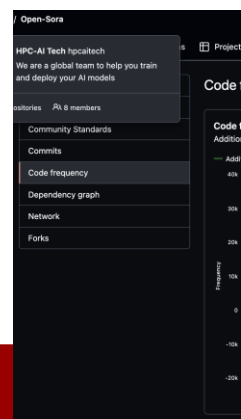
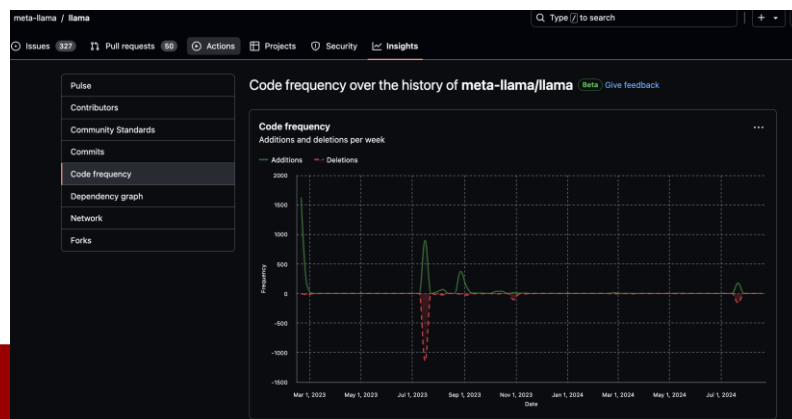


Source	Views
github.com	3,149
Bing	767
huggingface.co	298
cn.bing.com	280
jira.enflame.cn	191
reddit.com	72
paperswithcode.com	60
gitee.com	49
heehel.com	47

Acknowledgment

We extend our heartfelt appreciation for the great contribution to the open-source community, especially Open-Sora-Plan, as we build our diffusion transformer (DiT) based on Open-Sora-Plan v1.2.

- [Open-Sora-Plan](#): A project aims to create a simple and scalable repo, to reproduce Sora.
- [Open-Sora](#): An initiative dedicated to efficiently producing high-quality video.
- [ColossalAI](#): A powerful large model parallel acceleration and optimization system.
- [VideoSys](#): An open-source project that provides a user-friendly and high-performance infrastructure for video generation.
- [DiT](#): Scalable Diffusion Models with Transformers.
- [PixArt](#): An open-source DiT-based text-to-image model.
- [StabilityAI VAE](#): A powerful image VAE model.
- [CLIP](#): A powerful text-image embedding model.
- [T5](#): A powerful text encoder.
- [Playground](#): A state-of-the-art open-source model in text-to-image generation.
- [EMA-VFI](#): A video frame interpolation model.



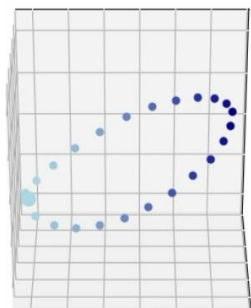
视觉世界模型探索1：相机位姿先验4D生成



北京大学
PEKING UNIVERSITY

物理先验知识作为模型输入，以相机位姿作为物理先验建模，输入视频生成模型中进行训练，构建不同视角下的4D世界 (已开源)

相机位姿发生变化后视角变化的4D生成



视觉世界模型探索2：全景3D生成



北京大学
PEKING UNIVERSITY

全景二维信息作为先验知识进行建模，输入视频生成模型中进行训练，构建不同视角下的**3D全景世界**，可用于游戏世界模拟和探索

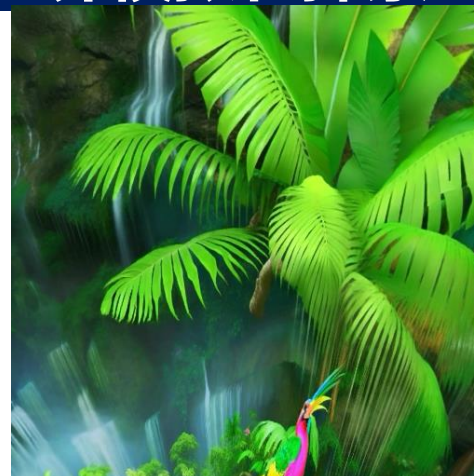
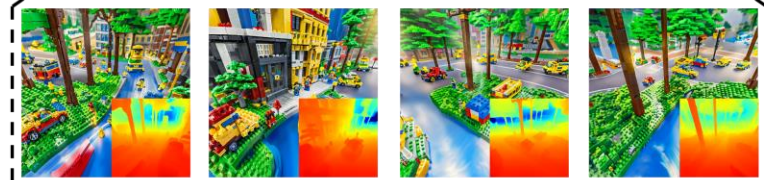
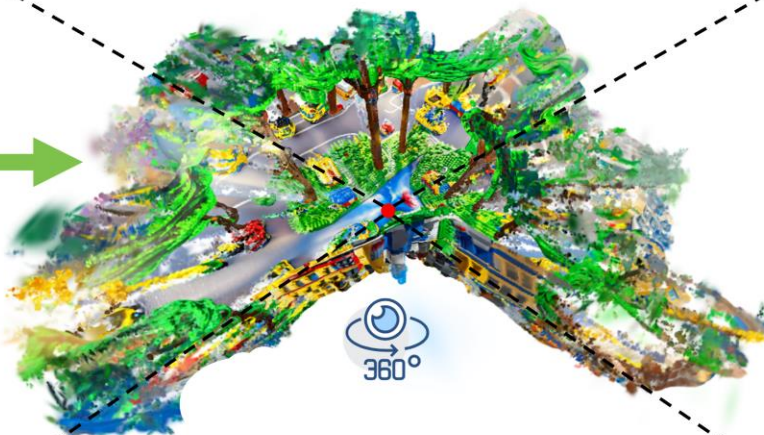
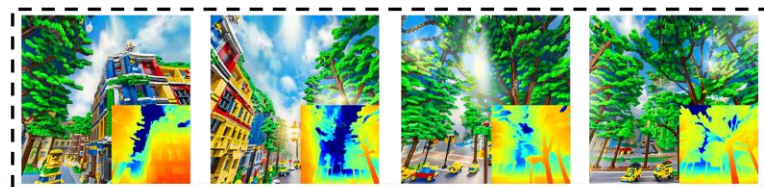


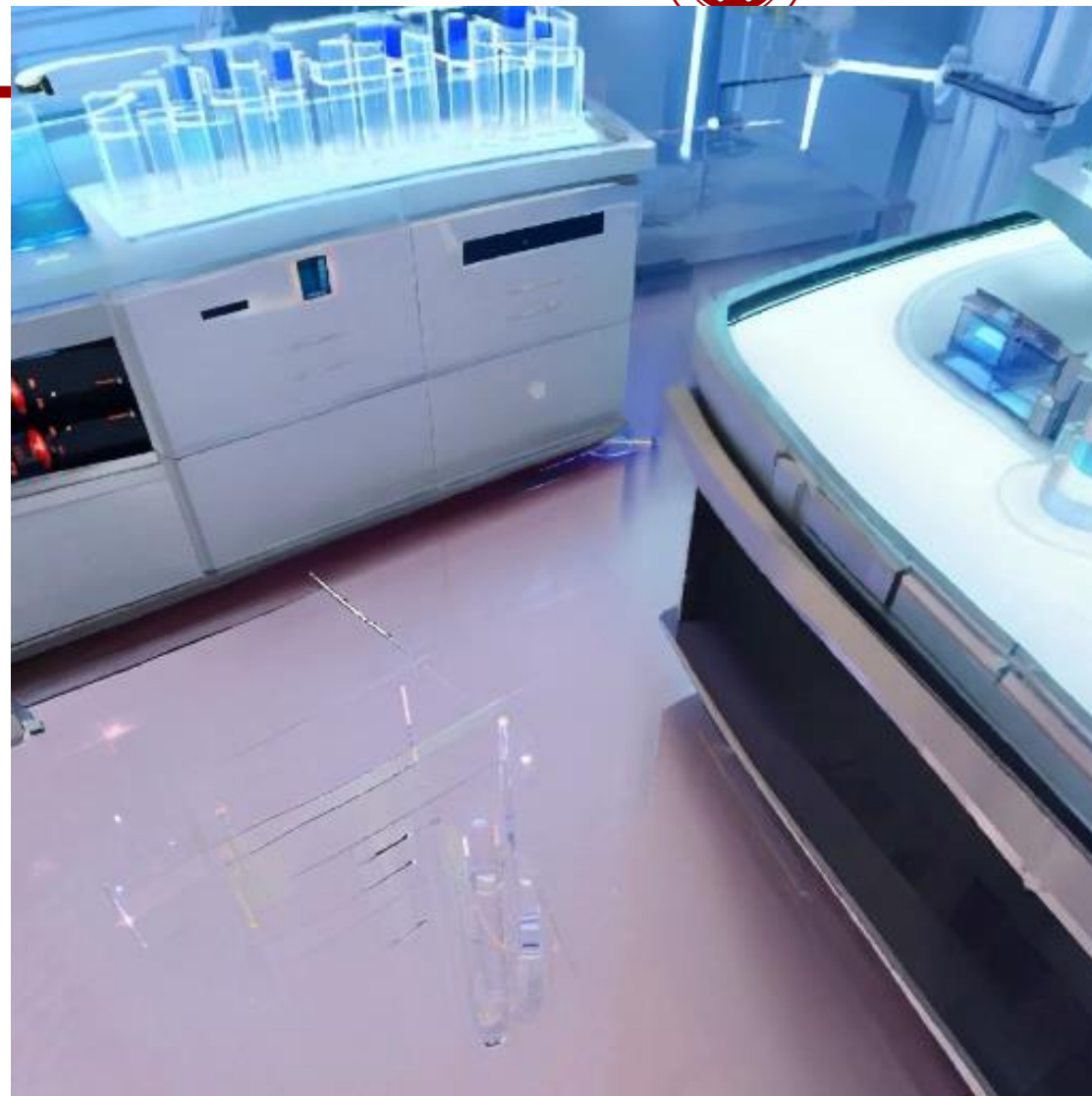
城市里面有乐高商店、乐高路、路灯、街上的汽车和乐高人、公园里的乐高树和湖泊。

统一架构构建全景图



全景先验知识输入到
Open-Sora视频生成模型

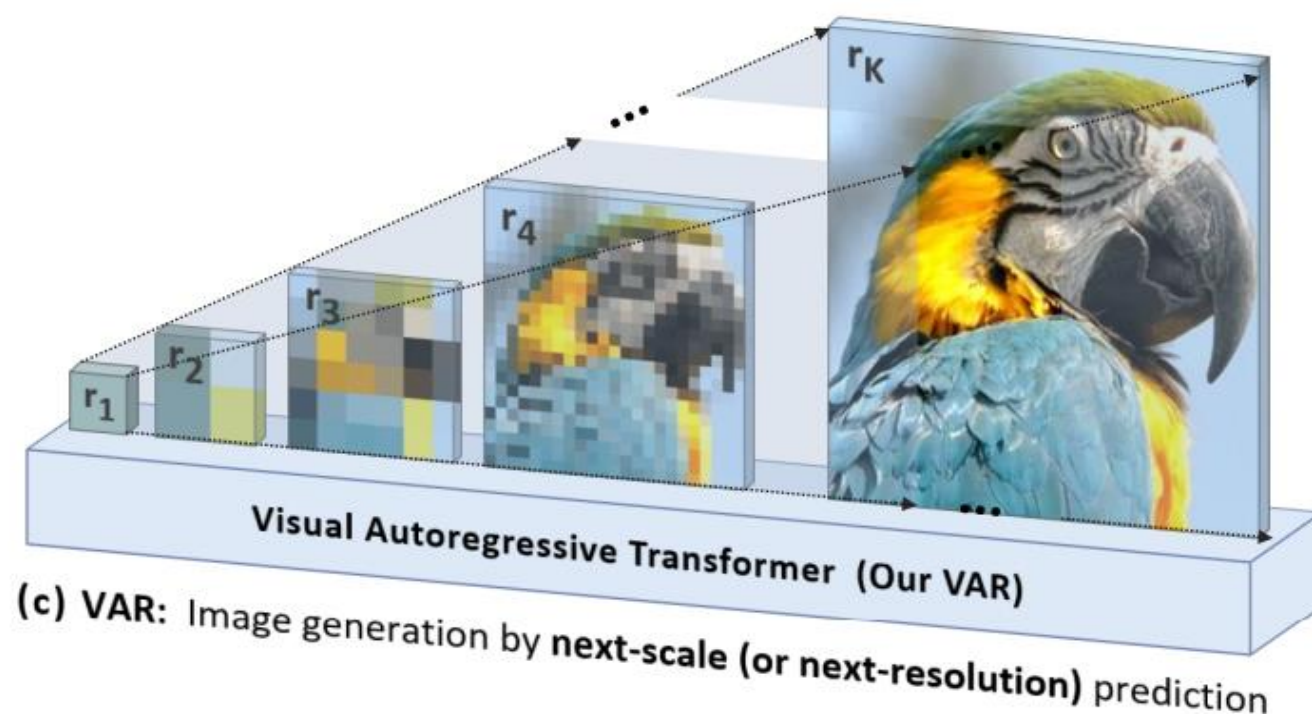
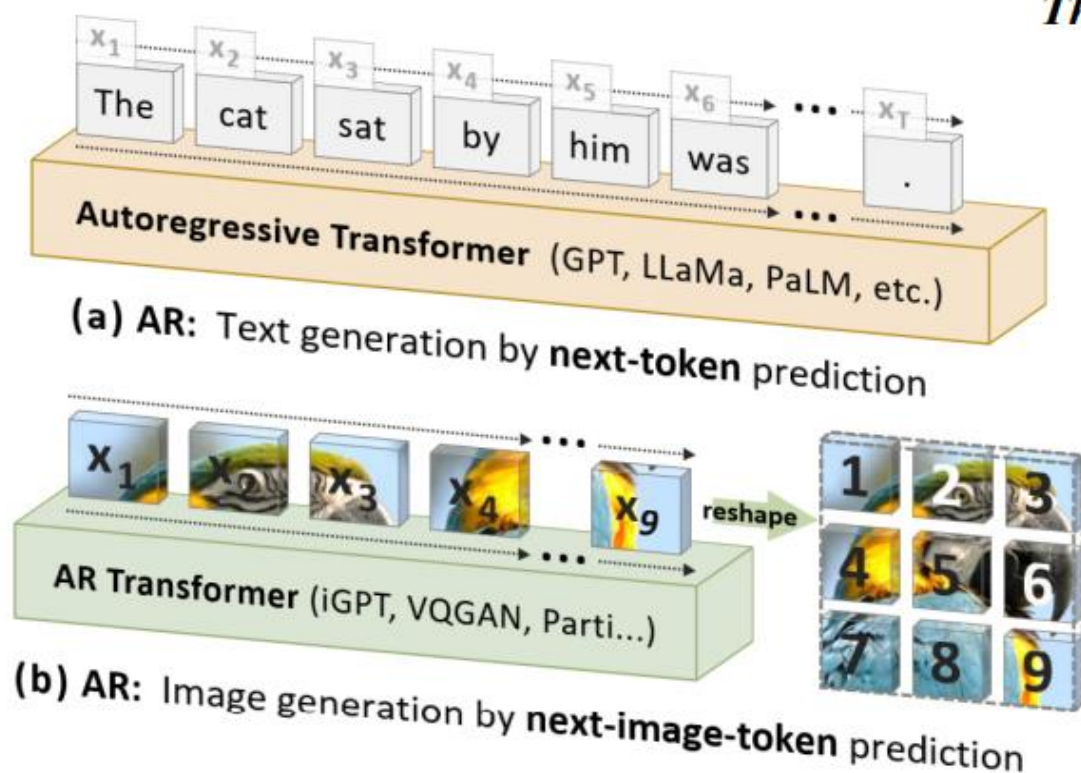




视觉生成架构是否已经收敛到Diffusion?

Autoregression model by next-resolution (next-scale) prediction

*Three Different Autoregressive
Generative Models*



视觉生成架构是否已经收敛到Diffusion?

Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation

Peize Sun¹ Yi Jiang^{2†} Shoufa Chen¹ Shilong Zhang¹ Bingyue Peng²

Ping Luo^{1*} Zehuan Yuan^{2*}

¹The University of Hong Kong ²ByteDance

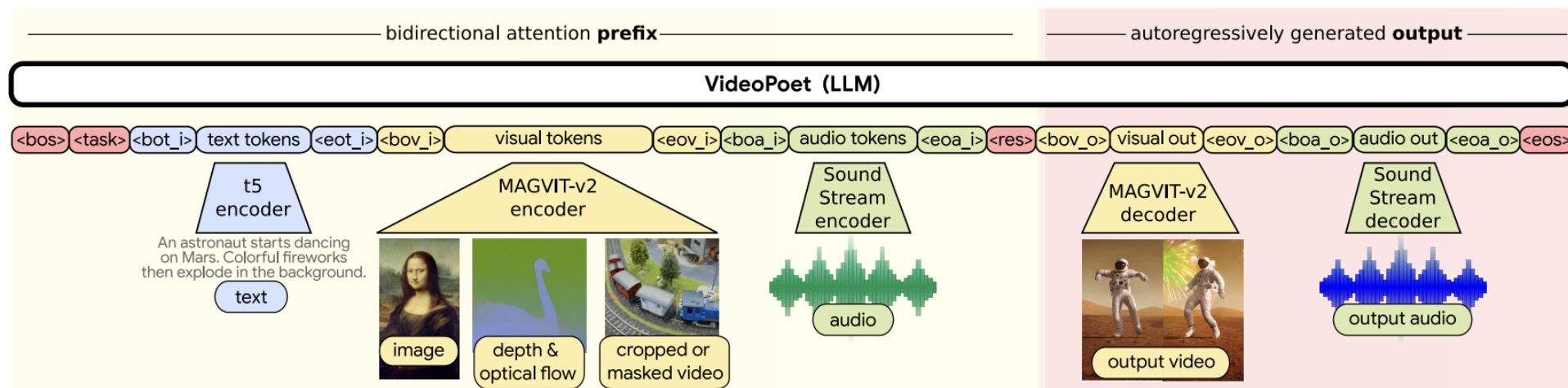
Codes and models: <https://github.com/FoundationVision/LlamaGen>



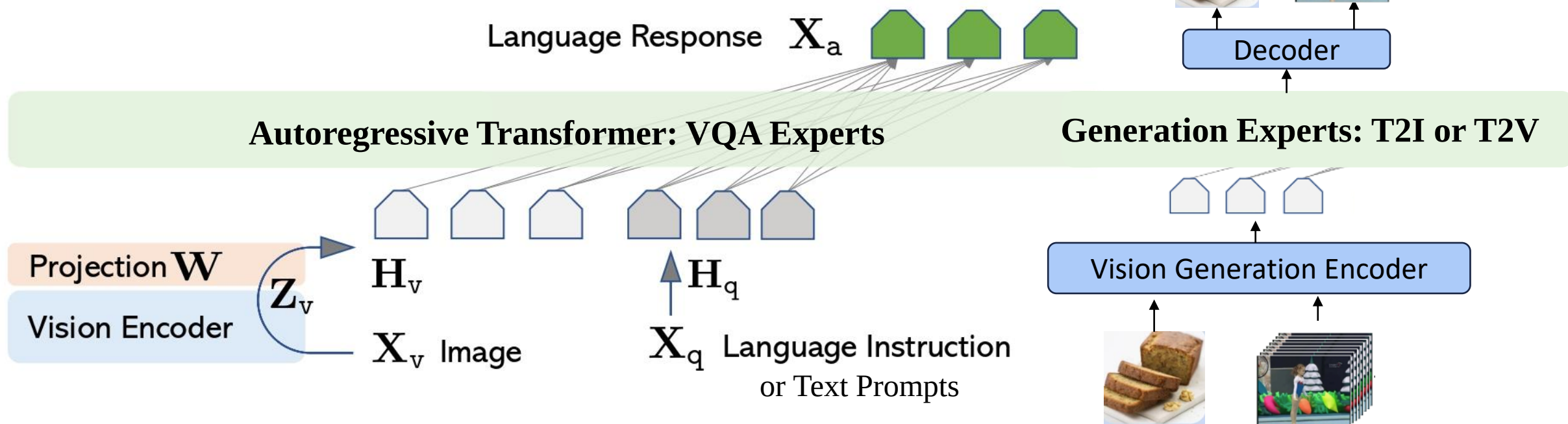
Figure 1: **Image generation with vanilla autoregressive models.** We show samples from our class-conditional image (top row) and text-conditional image (bottom row) generation models.

Type	Model	#Para.	FID↓	IS↑	Precision↑	Recall↑
GAN	BigGAN [Brock et al. 2018]	112M	6.95	224.5	0.89	0.38
	GigaGAN [Kang et al. 2023]	569M	3.45	225.5	0.84	0.61
	StyleGAN-XL [Sauer et al. 2022]	166M	2.30	265.1	0.78	0.53
Diffusion	ADM [Dhariwal & Nichol 2021]	554M	10.94	101.0	0.69	0.63
	CDM [Ho et al. 2022b]	—	4.88	158.7	—	—
	LDM-4 [Rombach et al. 2022]	400M	3.60	247.7	—	—
	DiT-XL/2 [Peebles & Xie 2023]	675M	2.27	278.2	0.83	0.57
Mask.	MaskGIT [Chang et al. 2022]	227M	6.18	182.1	0.80	0.51
	MaskGIT-re [Chang et al. 2022]	227M	4.02	355.6	—	—
AR	VQGAN [Esser et al. 2021]	227M	18.65	80.4	0.78	0.26
	VQGAN [Esser et al. 2021]	1.4B	15.78	74.3	—	—
	VQGAN-re [Esser et al. 2021]	1.4B	5.20	280.3	—	—
	ViT-VQGAN [Yu et al. 2021]	1.7B	4.17	175.1	—	—
	ViT-VQGAN-re [Yu et al. 2021]	1.7B	3.04	227.4	—	—
	RQTran. [Lee et al. 2022]	3.8B	7.55	134.0	—	—
	RQTran-re [Lee et al. 2022]	3.8B	3.80	323.7	—	—
AR	LlamaGen-B (cfg=2.00)	111M	5.46	193.61	0.83	0.45
	LlamaGen-L (cfg=2.00)	343M	3.07	256.06	0.83	0.52
	LlamaGen-XL (cfg=1.75)	775M	2.62	244.08	0.80	0.57
	LlamaGen-XXL (cfg=1.75)	1.4B	2.34	253.90	0.80	0.59
	LlamaGen-3B (cfg=1.65)	3.1B	2.18	263.33	0.81	0.58
	LlamaGen-3B (cfg=1.75)	3.1B	2.32	280.10	0.82	0.56
	LlamaGen-3B (cfg=2.00)	3.1B	2.81	311.59	0.84	0.54

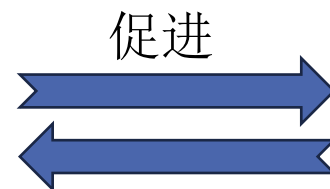
视觉生成架构是否已经收敛到Diffusion?



如何实现生成和理解统一的原生框架？



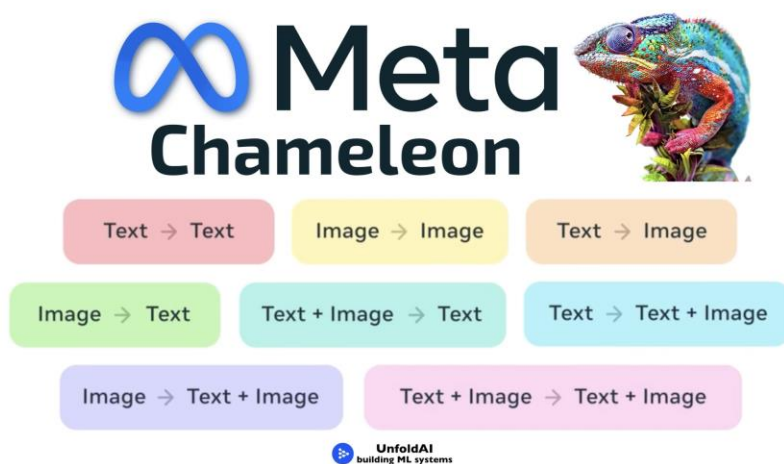
视觉理解部分



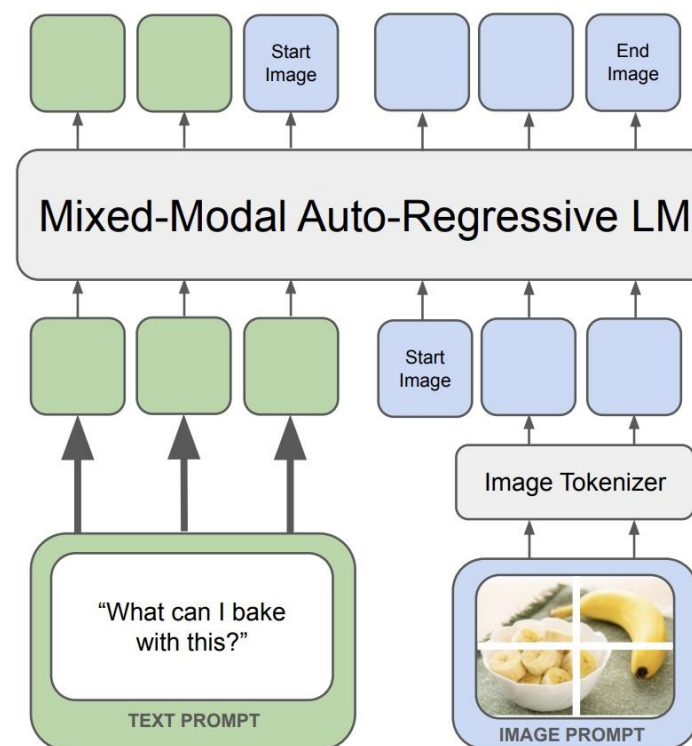
视觉生成部分

What I can create, I do understand !

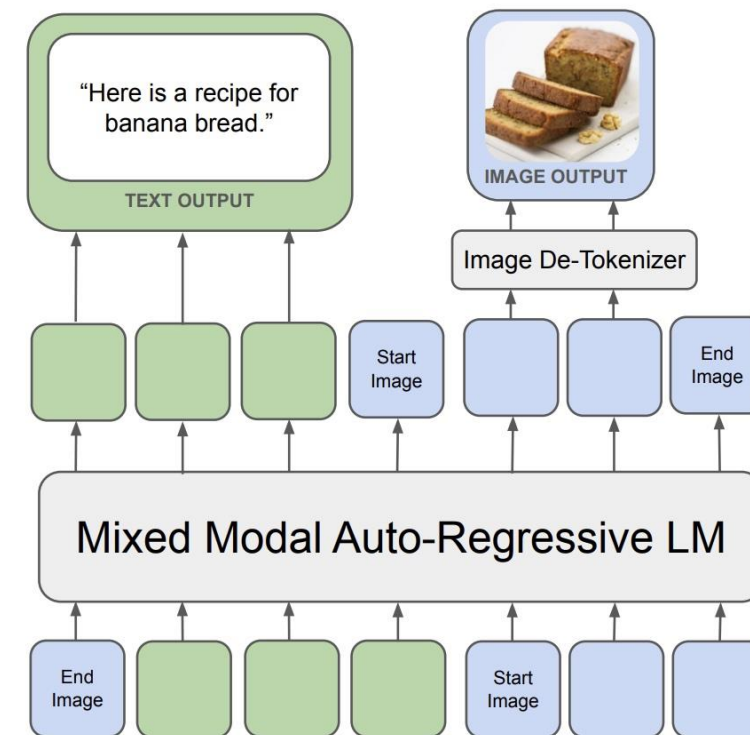
如何实现生成和理解统一的原生框架？



Chameleon	Concurrent GPUs	GPU Hours
7B	1024	856481
34B	3072	4282407

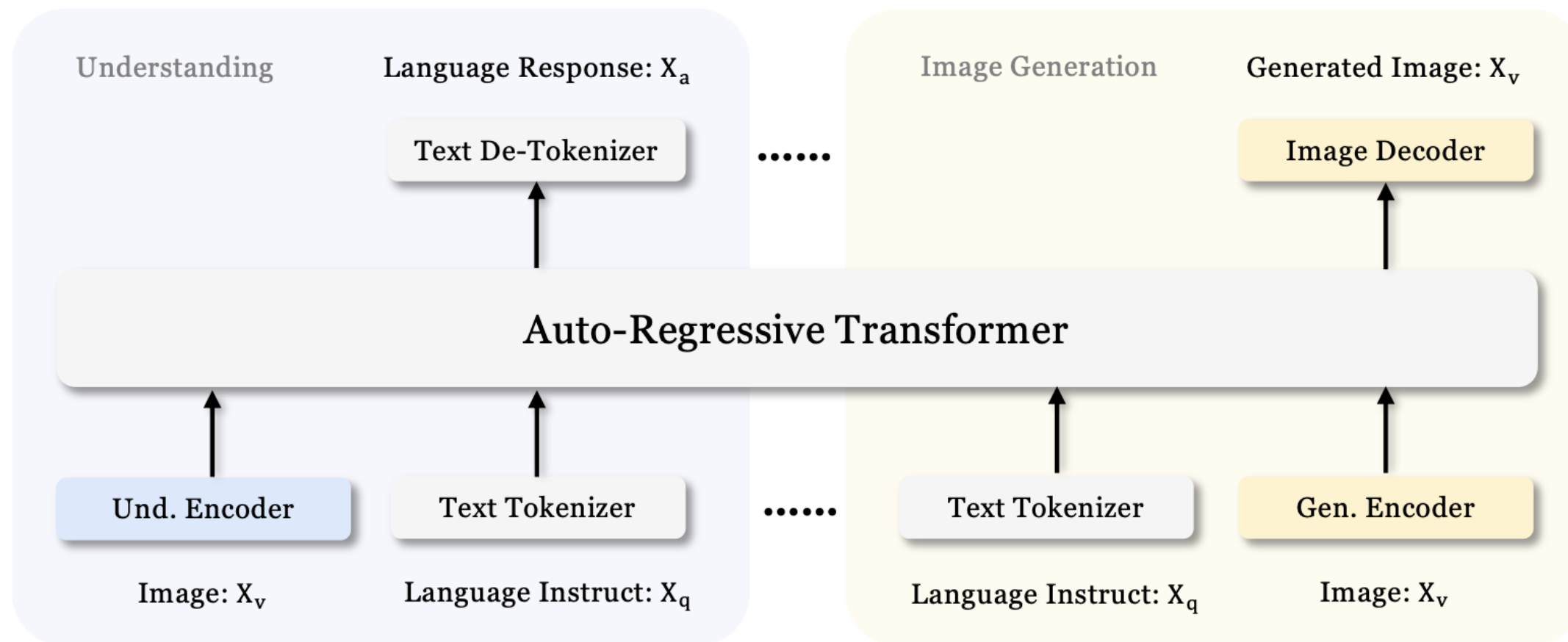


(a) Mixed-Modal Pre-Training

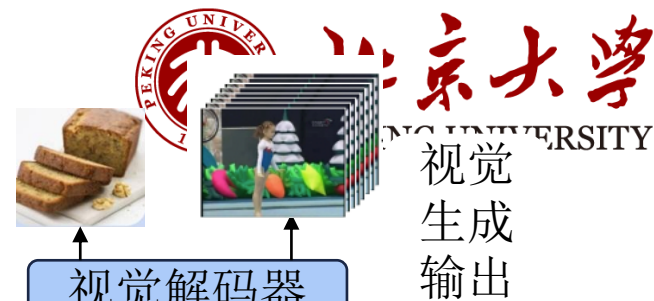


(b) Mixed-Modal Generation

如何实现生成和理解统一的原生框架？



如何实现生成和理解统一的原生框架？



理解输出：文本

X_a



自回归Transformer统一架构：理解专家部分

生成专家部分

Projection W

视觉理解编码器

Z_v

H_v

X_v Image

H_q

X_q

文本指令或文本提示词

视觉生成编码器

1. 视觉理解和生成的Encoder是否需要统一？短期不统一，长期统一该如何将压缩和对齐做到同一个Encoder中
2. 离散还是连续？离散的话Vocabulary Size多大合适？
2. 单纯增加Vocabulary Size可能比较低效，离散的基础上设计不同层次的字典作为残差补充，将Quantization Loss补回来？
3. 理解和生成的Task Conflict、视觉和语言模态的Gradients Conflict怎么办？
4. 视觉CoT如何加到统一架构，让视觉生成和理解都更加精准？LLaVA-o1
5. Visual Decoder是否可以是Diffusion，或者基于Diffusion后处理做Enhancer？



如何实现生成和理解统一的原生框架？

理解和生成在同一主干网络中端到端训练网络，
实现同一个模型能理解和生成（国产训练中）

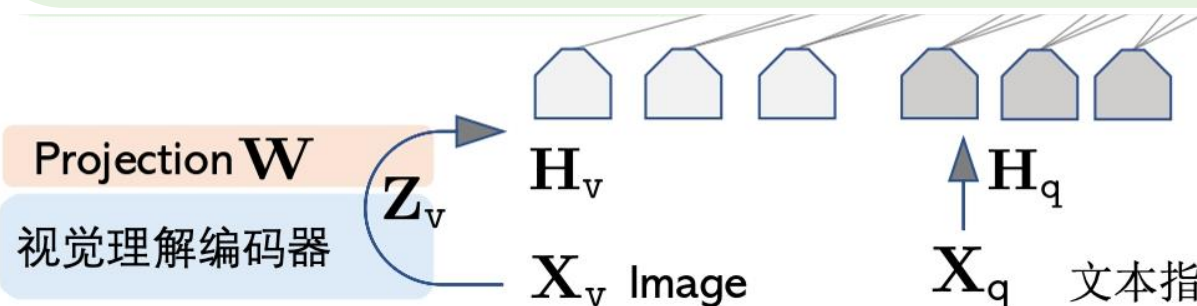
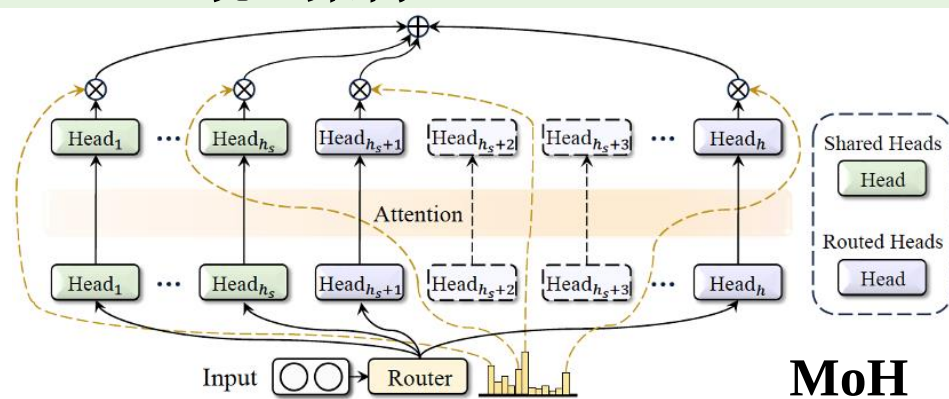
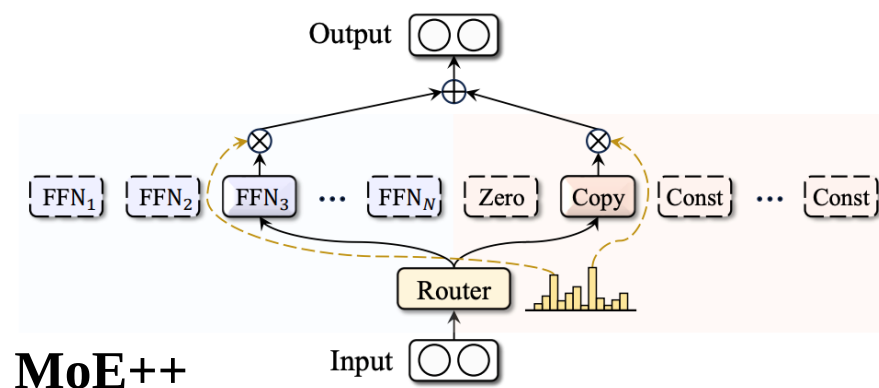
理解输出：文本



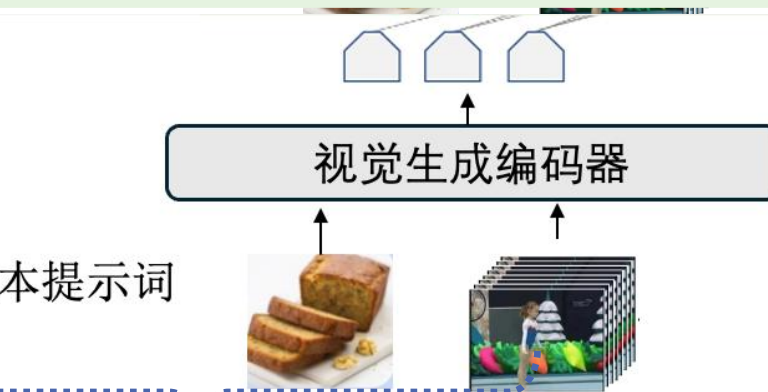
视觉生成输出

视觉解码器

基于混合注意力专家和混合专家是自回归Transformer统一架构



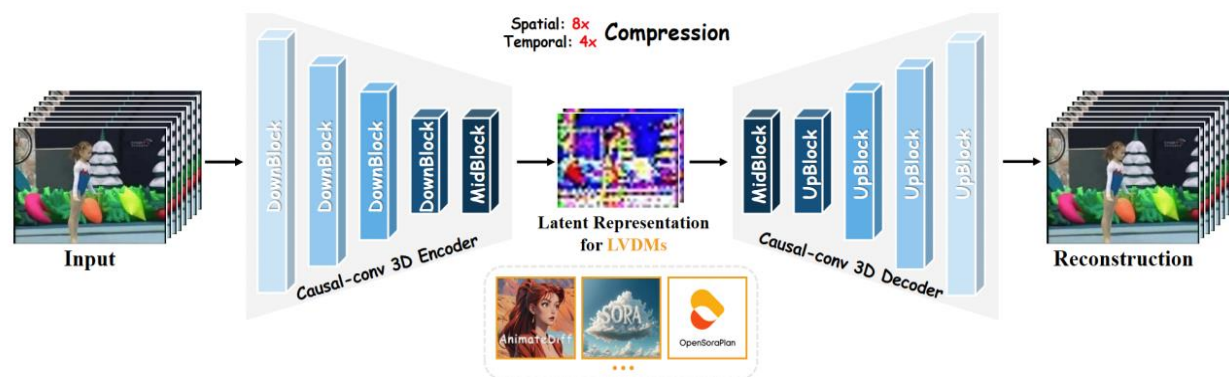
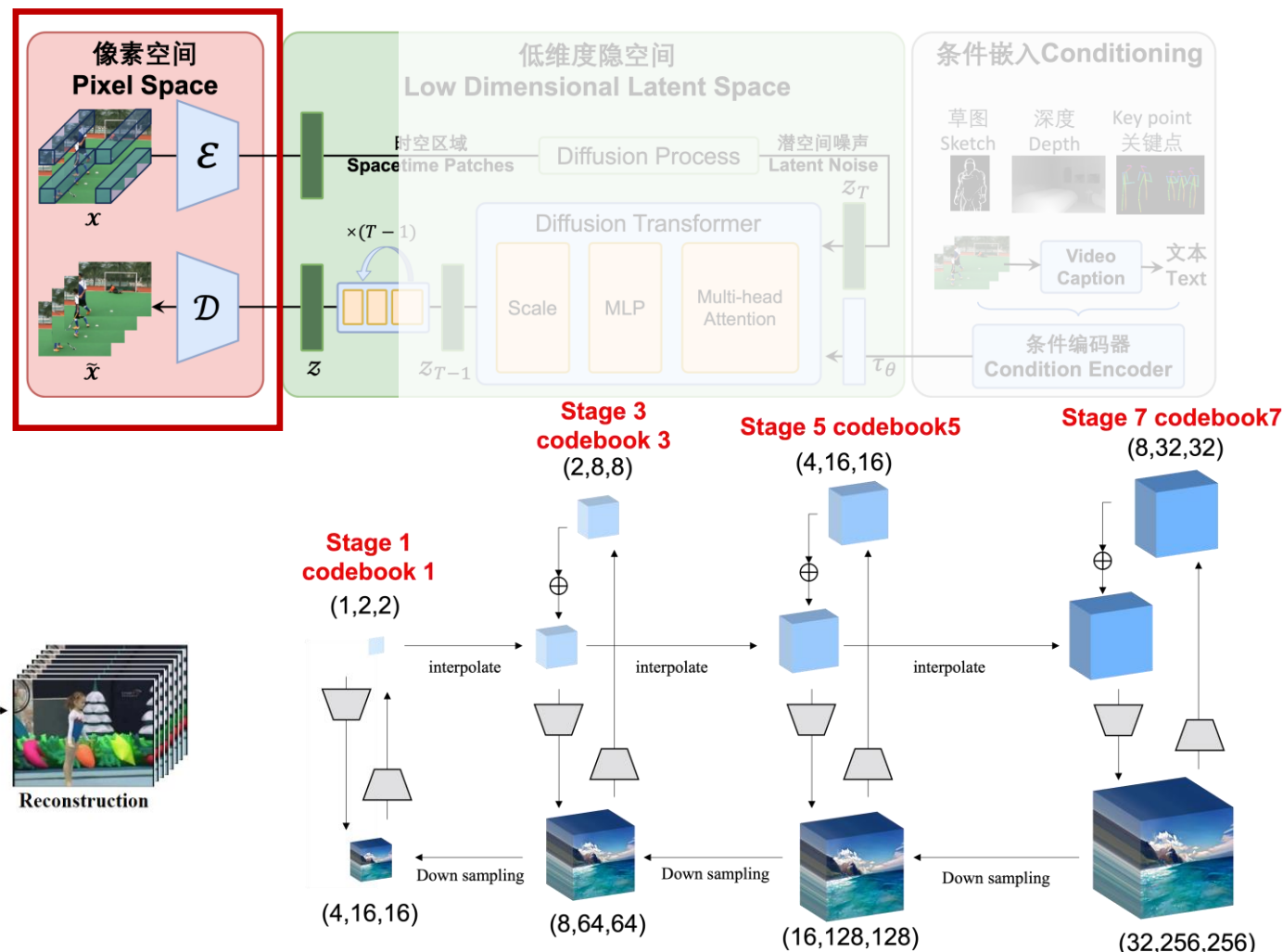
多模态理解任务



多模态生成任务

如何实现生成和理解统一的原生框架？

Compression as Intelligence (压缩即智能)



单 GPU/NPU 实现长视频压缩

Deepspeed	60s, 1080P Video			30s, 720P Video		
	Batch Size	GPU Memory	Step/sec	Batch Size	NPU Memory	Step/sec
Zero2	1	73	0.04	3	60	0.05
Zero3	1	71	0.04	3	59	0.05



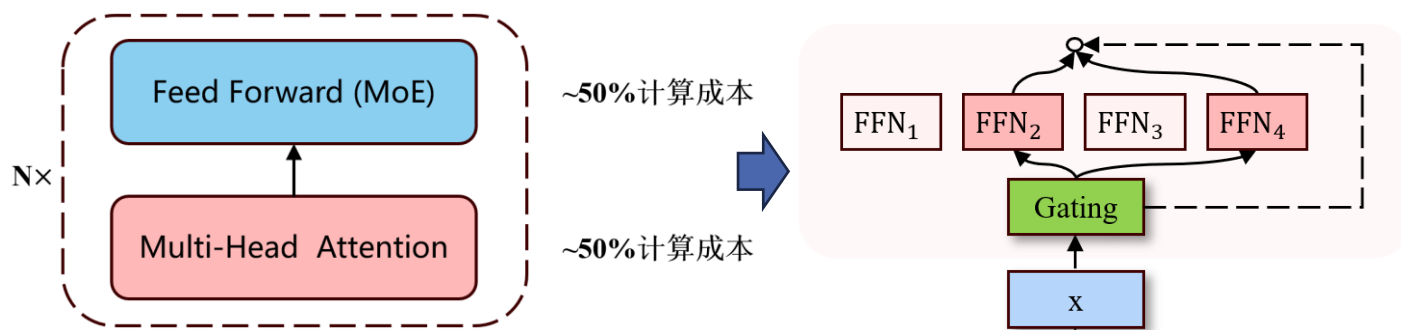
Origin



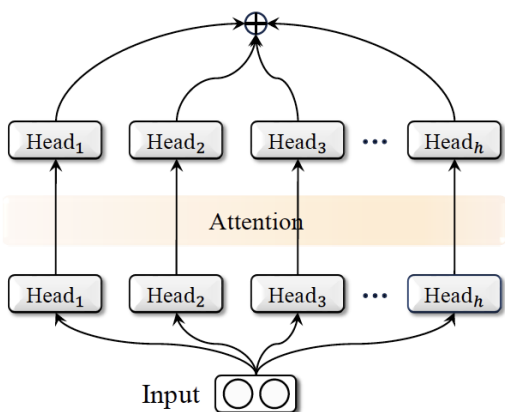
Reconstruction

如何实现生成和理解统一的原生框架?

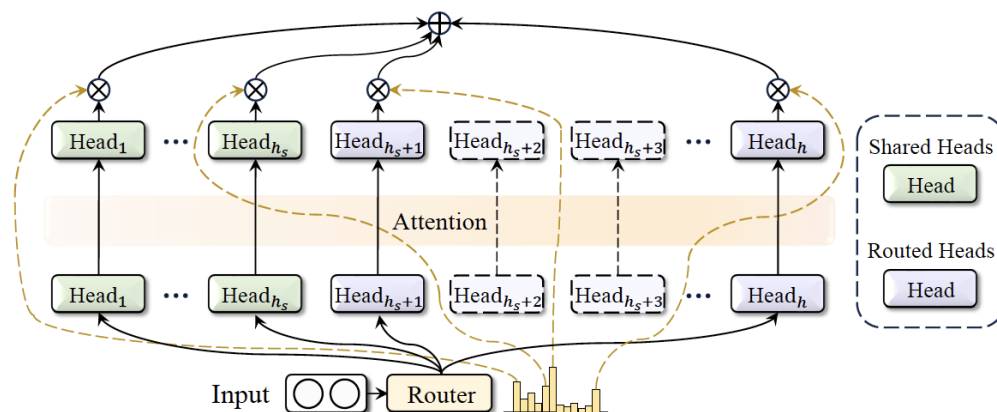
MoE、MoH架构对于原生统一也至关重要



Methods	#Activated Heads (%)	Language Tasks						Avg.
		SciQ	PIQA	WinoGrande	OpenbookQA	LogiQA	TruthfulQA	
LLM-S 100B	100	63.0	63.1	51.1	27.4	26.9	31.6	43.9
MoH-LLM-S 100B	75	64.7	62.0	50.6	28.8	26.4	35.2	44.6
MoH-LLM-S 100B	50	67.0	62.2	51.5	29.2	26.7	35.6	45.4
LLM-B 100B	100	73.1	69.7	52.0	31.8	28.4	29.5	47.4
MoH-LLM-B 100B	75	74.7	69.2	52.8	30.0	28.1	32.2	47.8
MoH-LLM-B 100B	50	75.2	67.0	52.0	29.0	26.9	32.8	47.2
LLM-B 200B	100	73.1	70.3	53.3	32.4	29.0	29.5	47.9
MoH-LLM-B 200B	75	76.0	69.2	52.7	30.4	29.8	32.6	48.5
MoH-LLM-B 200B	50	75.6	66.9	53.5	29.4	26.7	32.7	47.5



(a) Multi-Head Attention

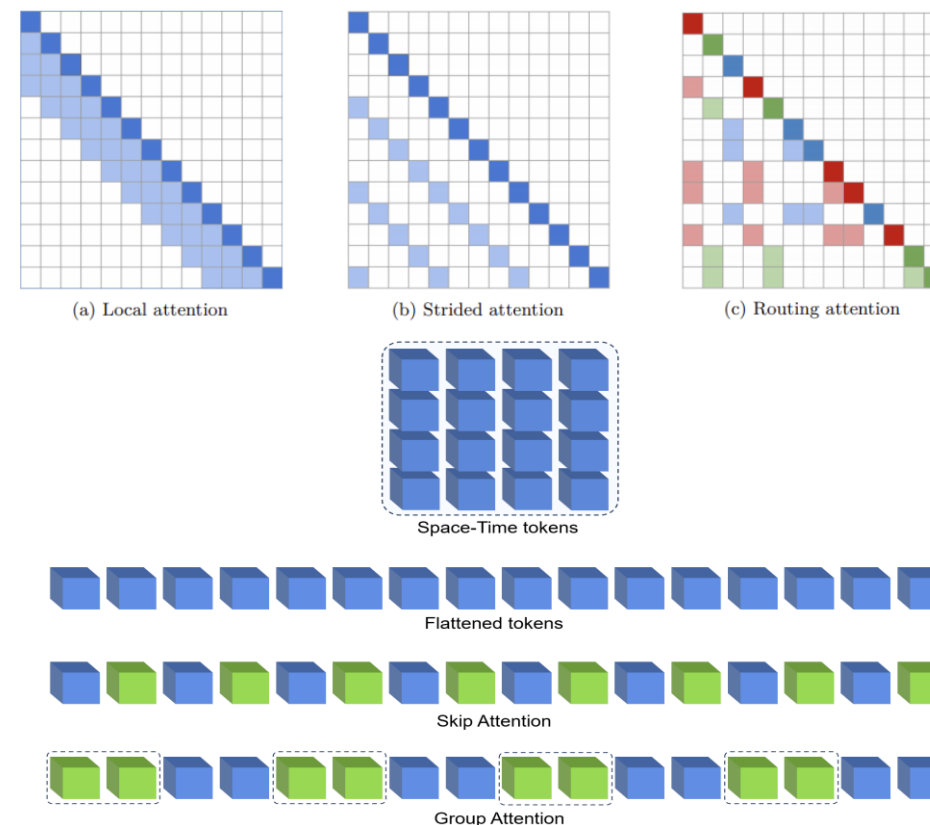
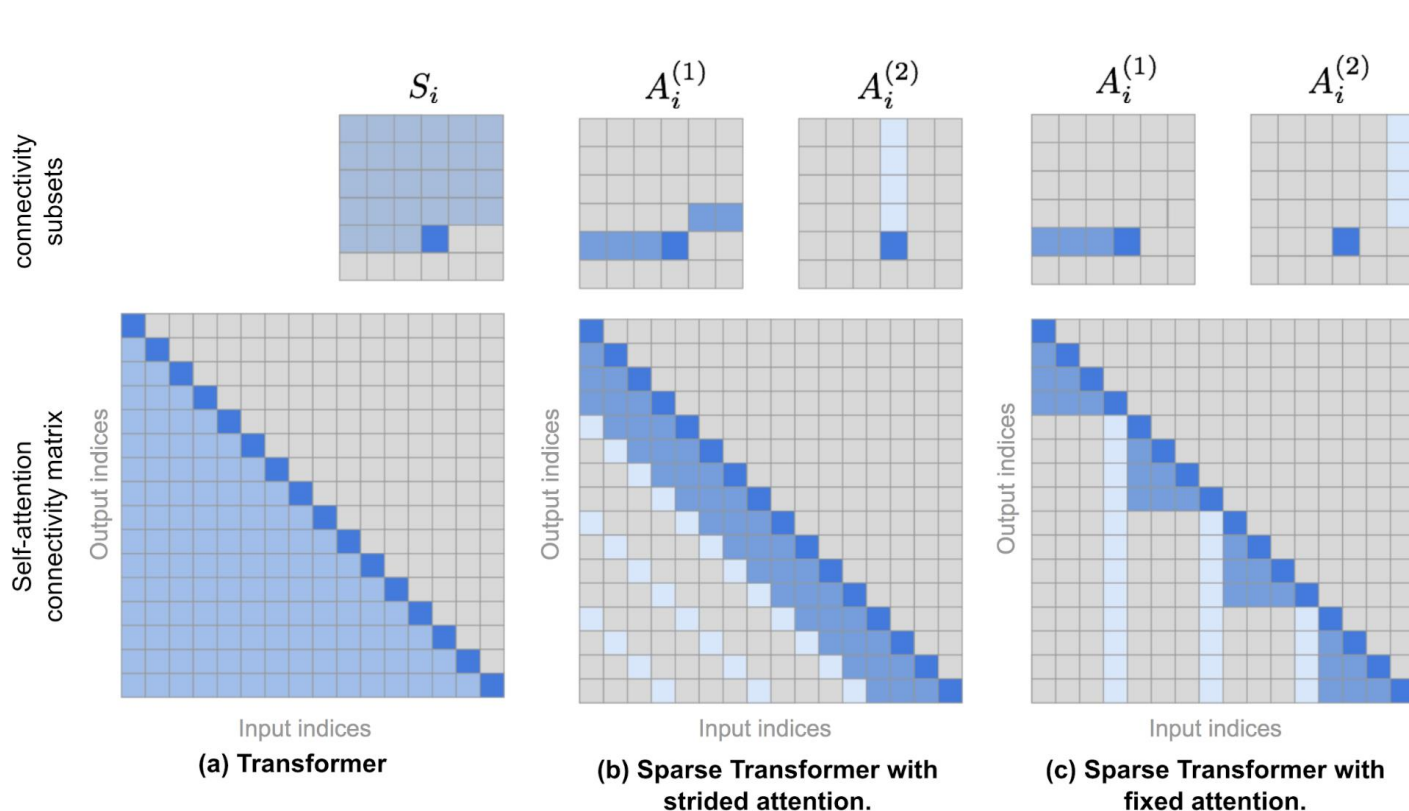


(b) Our proposed Mixture-of-Head Attention

Methods	#Params (M)	#Activated Heads (%)	FID↓	sFID↓	IS↑
DiT-S/2-400K (Peebles & Xie, 2023)	33	100	68.40	-	-
MoH-DiT-S/2-400K	33	90	67.25	12.15	20.52
MoH-DiT-S/2-400K	33	75	69.42	12.85	19.96
DiT-B/2-400K (Peebles & Xie, 2023)	130	100	43.47	-	-
MoH-DiT-B/2-400K	130	90	43.40	8.40	33.51
MoH-DiT-B/2-400K	130	75	43.61	8.48	33.43
DiT-L/2-400K (Peebles & Xie, 2023)	458	100	23.33	-	-
MoH-DiT-L/2-400K	458	90	23.17	6.16	58.92
MoH-DiT-L/2-400K	458	75	24.29	6.38	57.75

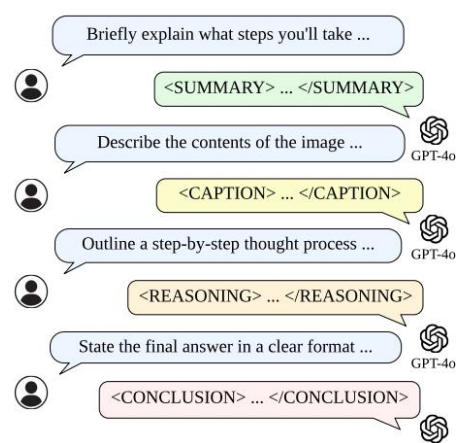
如何实现生成和理解统一的原生框架？

高效注意力机制永远都很重要

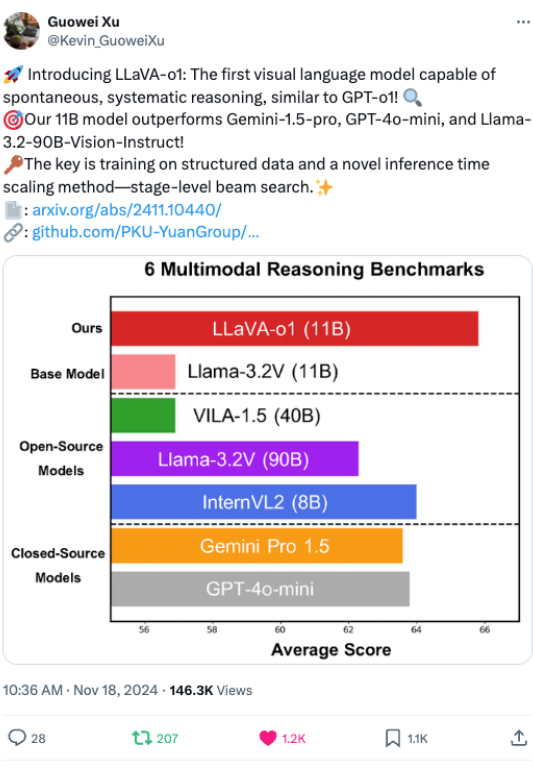
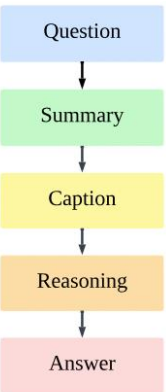


多模态CoT增强原生统一框架？

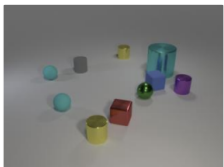
LLaVA-CoT/o1模型开源，首个视觉多模态慢思考模型



Dataset	Type	Size
ShareGPT4V [8]	General VQA	31.3k
ChartQA [38]	General VQA	17.2k
A-OKVQA [45]	General VQA	16.1k
AI2D [23]	Science-Targeted VQA	11.4k
GeoQA+ [7]	Science-Targeted VQA	11.4k
ScienceQA [34]	Science-Targeted VQA	5.6k
DocVQA [39]	General VQA	4.0k
PISC [28]	General VQA	1.0k
CLEVR [22]	General VQA	0.5k
CLEVR-Math [13]	Science-Targeted VQA	0.5k



Example 1: Reasoning Problems



Question: Subtract all tiny shiny balls. Subtract all purple objects. How many objects are left?

Options: A. 4, B. 8, C. 2, D. 6

Answer from Llama-3.2-11B-Vision-Instruct:

To determine the correct answer, we need to follow the instructions provided.

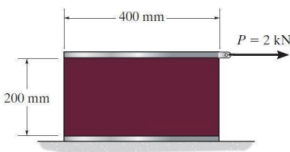
The first instruction is to subtract all tiny shiny balls. Since there are **three tiny shiny balls** in the image, we would be left with no shiny balls after performing this action. The second instruction is to subtract all purple objects. However, upon closer inspection, it becomes apparent that there are **no shiny purple objects present in the image**. Therefore, we cannot subtract any purple objects.

Given that we have already removed all the tiny shiny balls and there were no purple objects to begin with, the correct answer is C. 2, as initially, **there were 2 purple objects, but they are not to be subtracted** according to the instruction.

Answer from LLaVA-o1:

Summary stage: What's the problem? What should I do?
<SUMMARY> I will examine the image to identify and count all the objects, **focusing on the tiny shiny balls and purple objects**. Then, I will subtract these objects from the total count and **determine the number of remaining objects**. </SUMMARY>
Caption stage: What can I know from the image?
<CAPTION> The image shows various 3D geometric shapes on a flat surface, including spheres, cubes, and cylinders in different colors. Notably, there are **small shiny green spheres and a purple cylinder**. </CAPTION>
Reasoning stage: How to solve the problem step-by-step?
<REASONING> First, I will count all the objects in the image: there are **10 objects in total**. Next, I will identify and subtract the tiny shiny balls. There is **one small shiny green sphere**. Then, I will identify and subtract the purple objects. There is **one purple cylinder**. By subtracting the two objects (the shiny ball and the purple cylinder) from the total, I find that **8 objects remain**. </REASONING>
Conclusion stage: What is the final answer?
<CONCLUSION> B </CONCLUSION>

Example 2: Science Problems



Question: An acetel polymer block is fixed to the rigid plates at its top and bottom surfaces. If the top plate displaces 2 mm horizontally when it is subjected to a horizontal force P=2 kN, determine the shear modulus of the polymer. The width of the block is 100 mm. Assume that the polymer is linearly elastic and use small angle analysis.

Options: A: 15 MPa, B: 5 MPa, C: 20 MPa, D: 2 MPa

Answer from Llama-3.2-11B-Vision-Instruct:

To determine the shear modulus of the polymer, we can use the formula for shear modulus in a small angle analysis:
 $G = \frac{F}{A \cdot \gamma}$
where γ is the shear strain, F is the applied force, A is the area of the block, and γ is the shear strain.

Given that the top plate of the block displaces horizontally by 2 mm when subjected to a horizontal force of 2 kN, we can calculate the shear strain γ . Since the displacement is horizontal and the original length of the block is 400 mm, the shear strain γ is:
 $\gamma = \frac{\Delta x}{L} = \frac{2 \text{ mm}}{400 \text{ mm}} = 0.005$

Next, we calculate the area A of the block. The width of the block is given as 100 mm, and assuming the height of the block is also 200 mm (as given in the image analysis), the area A is:
 $A = 100 \text{ mm} \times 200 \text{ mm} = 20000 \text{ mm}^2$

Now, we can substitute these values into the formula for shear modulus:
 $G = \frac{F}{A \cdot \gamma} = \frac{2000 \text{ N}}{20000 \text{ mm}^2 \cdot 0.005} = \frac{2000}{100} = 20 \text{ MPa}$

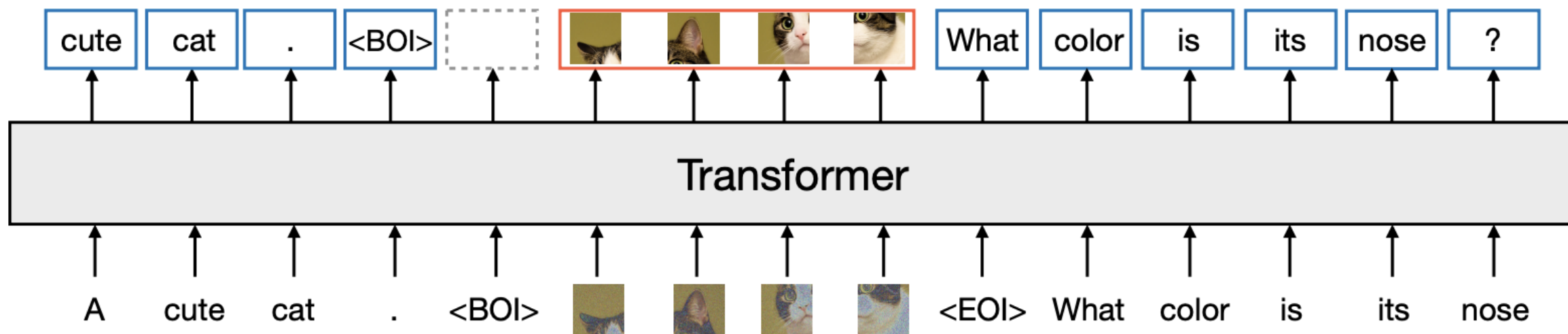
Therefore, the correct answer is C: 20 MPa.

Answer from LLaVA-o1:

Summary stage: What's the problem? What should I do?
<SUMMARY> To solve the problem, I will **determine the shear modulus** of the acetel polymer block using the relationship between the applied force, displacement, and dimensions of the block. I will use the **formula for shear modulus** and calculate the result. </SUMMARY>
Caption stage: What can I know from the image?
<CAPTION> The image shows a rectangular block of acetel polymer fixed between two rigid plates. The top plate moves horizontally by 2 mm when a force $P = 2 \text{ kN}$ is applied. The block's dimensions are given as **400 mm in length, 200 mm in height, and 100 mm in width**. </CAPTION>
Reasoning stage: How to solve the problem step-by-step?
<REASONING> The shear modulus G of the polymer can be calculated using the formula:
 $G = \frac{F}{A \cdot \gamma}$
where:
- F is the applied force (2 kN),
- A is the area of the block's top surface (100 mm x 400 mm = 40,000 mm²),
- γ is the horizontal displacement (2 mm),
- L is the height of the block (200 mm).
First, convert the force to Newtons: $F = 2 \text{ kN} = 2000 \text{ N}$.
Next, calculate the shear modulus:
 $G = \frac{2000}{40000 \cdot \frac{2}{400}} = \frac{2000}{200} = 10 \text{ MPa}$
 $G = \frac{2000}{40000 \cdot 0.01} = 5 \text{ MPa}$
</REASONING>
Conclusion stage: What is the final answer?
<CONCLUSION> B. 5 MPa </CONCLUSION>

如何实现生成和理解统一的原生框架？

争议点：统一Loss还是统一建模方式？



$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{y_i} \left[-\log P_{\theta}(y_i | y_{<i}) \right] \quad \mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon} \left[\|\epsilon - \epsilon_{\theta}(\mathbf{x}_t, t, c)\|^2 \right] \quad \mathcal{L}_{\text{Transfusion}} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}$$



总结

- 生成未必理解、理解无法视觉生成
- 理解基本收敛于自回归、而生成架构仍未收敛到Diffusion
- 自回归Autoregressive Transformer也许是大一统的关键
- 类Sora架构对于生成理解统一模型的实现仍然意义重大
- 路线之争：
 1. 究竟完全押注于自回归，还是在Loss层面统一？
 2. 视觉模态到底选择连续还是离散？

Thank you!