QCon
全球软件开发大会

InfoQ 极客传媒

理解与生成一体化：统一多模态大模型的自回归路线

王曜明

美团 Staff Engineer

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播

参会咨询

6月 上海 1000人

AiCon 全球人工智能开发与应用大会

会议时间：6月26-27日

- AI Infra 系统工程
- 多 Agent 协作与实践
- 多模态融合
- 模型训练与推理创新
- 数据平台与特征服务

8月 深圳 1000人

AiCon 全球人工智能开发与应用大会

会议时间：8月21-22日

- Agentic AI
- 轻量化与高效推理
- 多模态应用
- AI + IoT 场景实践
- AI 工业化落地

10月 上海 1200人

QCon 全球软件开发大会

会议时间：10月22-24日

- AI Agent
- Vibe Coding
- 智能可观测
- 推理基建
- 模型攻防
- AI x 创造力

12月 北京 1000人

AiCon 全球人工智能开发与应用大会

会议时间：12月18-19日

- 大模型架构创新
- 多模态 AI 产业融合
- 具身智能
- AI for Science
- 大模型安全

QCon
全球软件开发大会

InfoQ 极客传媒

目录

01 生成理解一体化模型的动机

02 行业主流路线梳理

03 Decoder-only纯自回归路线

04 全链路训练范式

05 工程实践与典型挑战

06 总结与未来展望

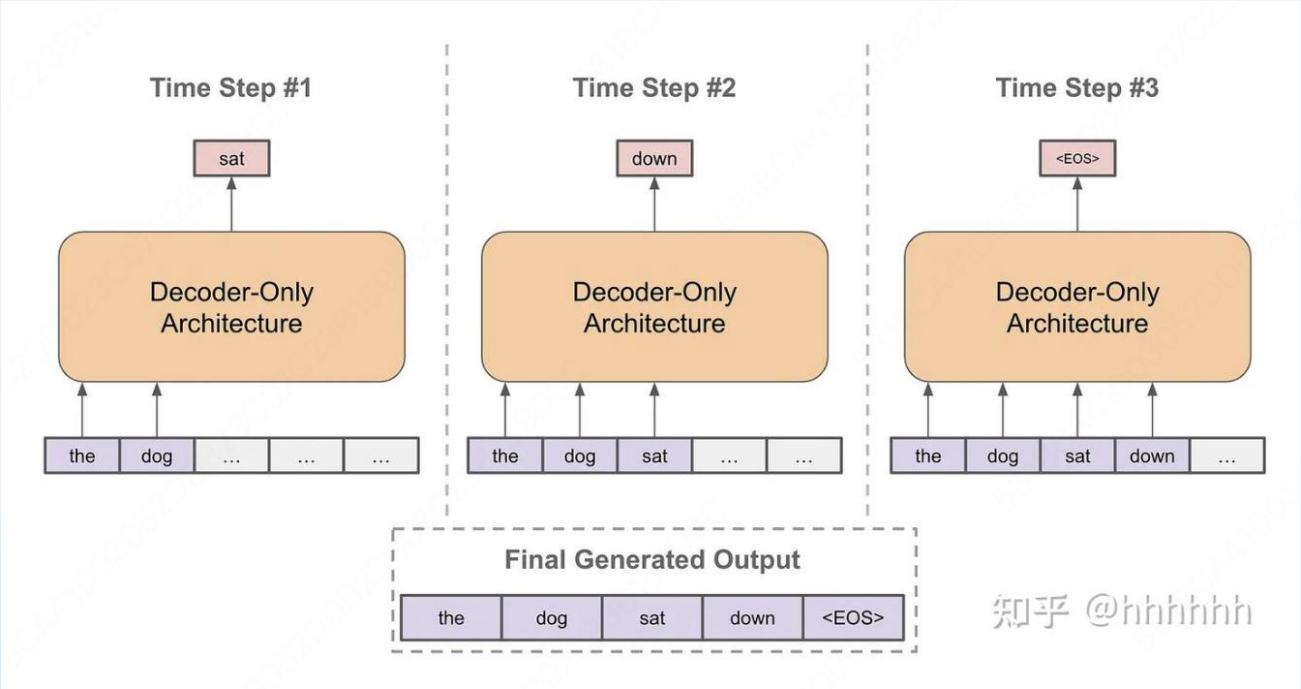
01

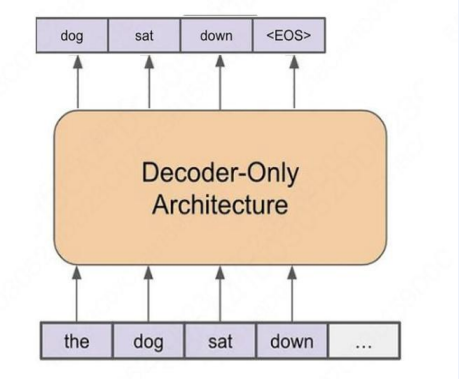
理解生成一体化模型

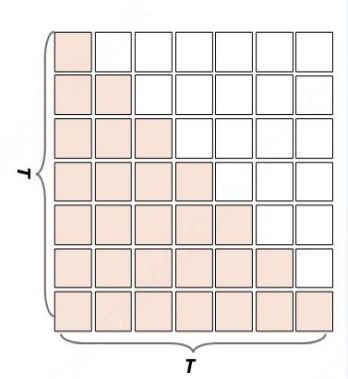
动机与演进趋势

LLM and Tokenizer

- 推理阶段: token by token进行解码输出
- 训练阶段: 是否也需要序列化串行输出？如何并行训练？







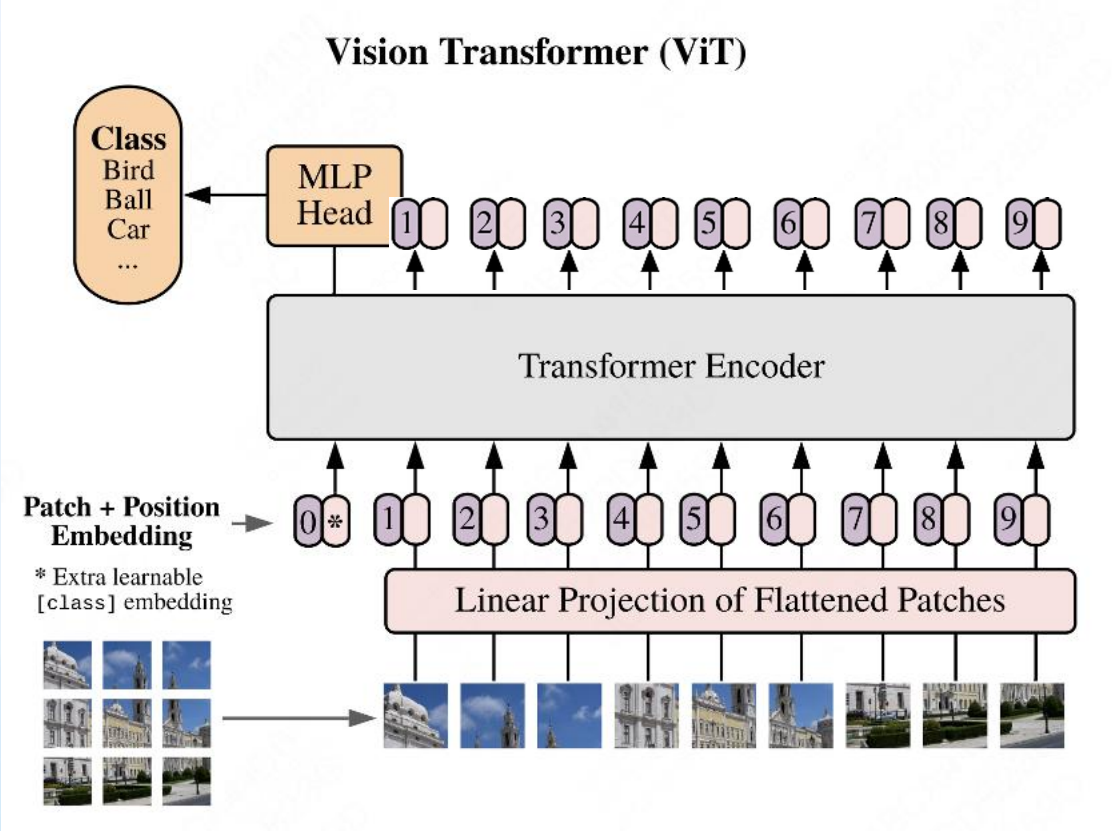
teacher forcing + causal attention mask

Next Token Prediction (NTP) Loss

- LLM如何处理非文本数据（如图像）?
- 多模态大模型(MLLM)需要一种“图像tokenizer”来将图像转化为可处理的序列

Vision Transformer and Patch Embedding

图像信息密度低，512x512图像包括两万个像素



Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

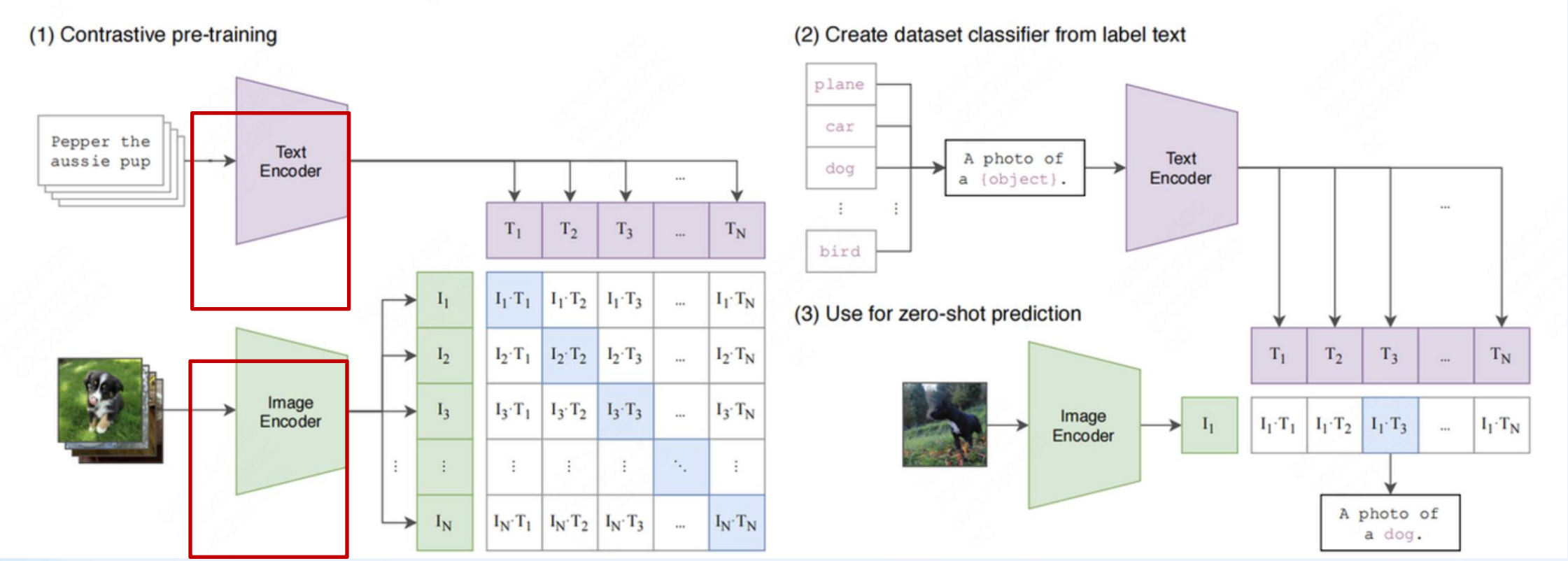
- **Visual Patch as Token**
  将一张图像切割成同样大小 (14 x14) 的小块，

  每个小块（patch）被展平，通过一个线性层 (patch embedding) 转换为一个向量，作为一个token，传递给Transformer作为输入。

- 文本Tokenizer： 句子-> 单词/词元
- 图像Tokenizer： 图像-> 图像块

ViT只能处理图像，如何将图像与语言模型结合？需要一种新的模型来实现视觉与语言的融合

CLIP 图文对比学习预训练



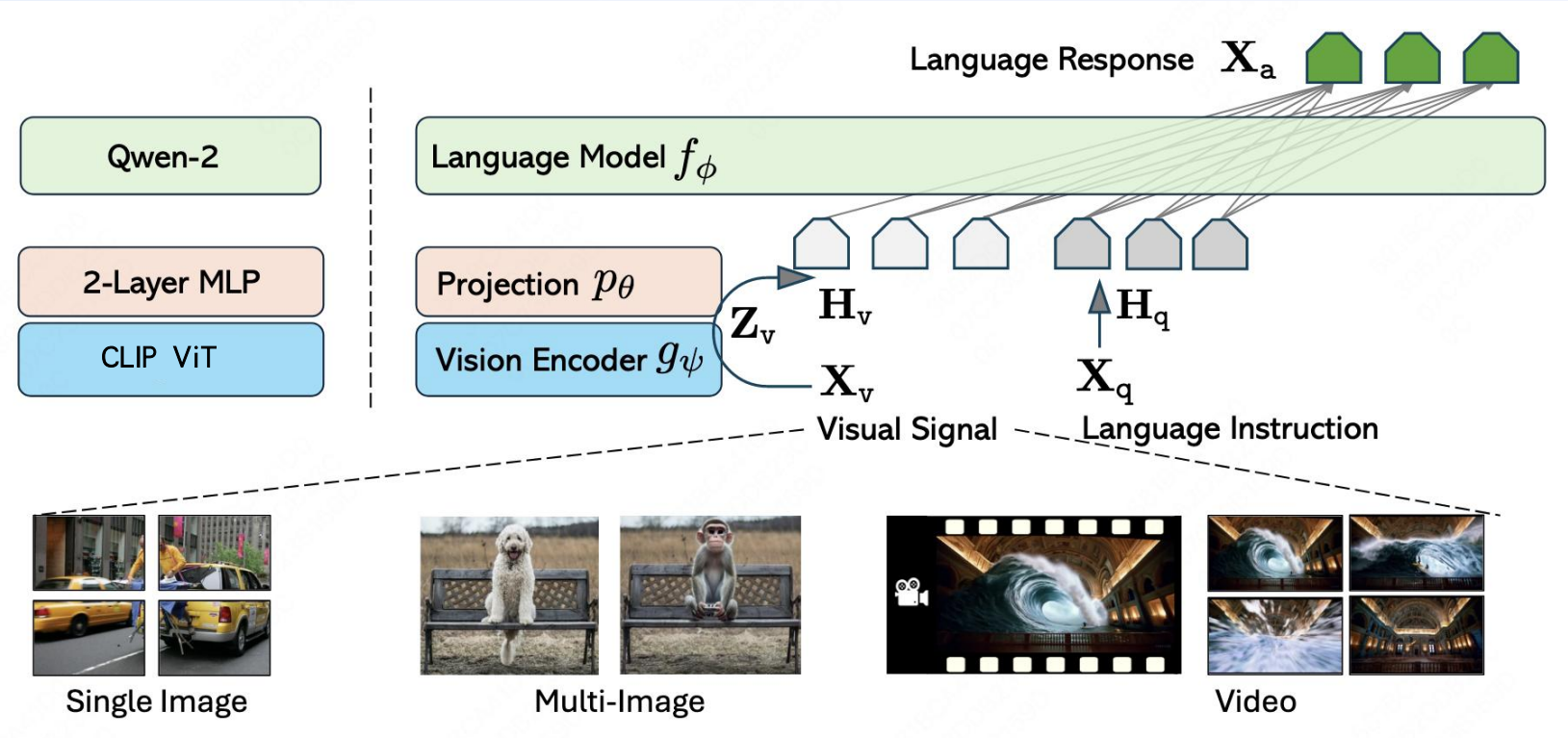
CLIP开启了视觉预训练和Vision Foundation Model时代，在4亿图文数据对上进行预训练。**这个模型不仅仅是为了微调迁移到其他纯视觉任务，更是为了能够**与语言模型无缝对接

CLIP虽然实现了视觉与语言的理解，但**还不能称之为大模型**，其能力主要集中在检索和判别式任务

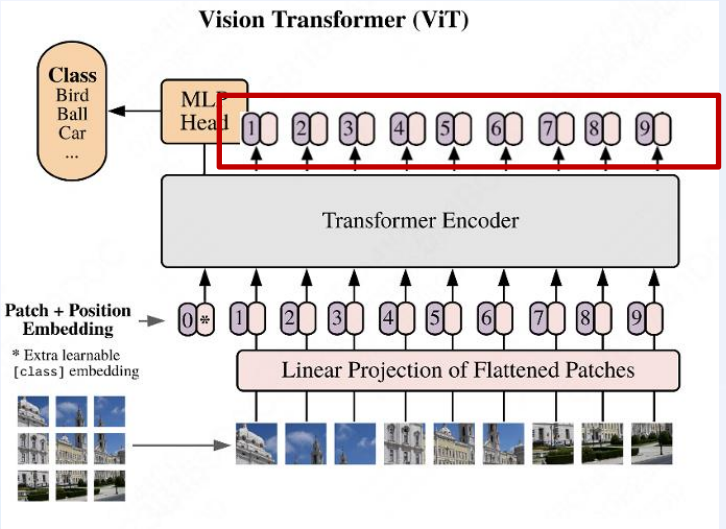
如何进一步和LLM结合扩展其文本生成式能力和交互能力？

从CLIP到LLaVA多模态大模型

整体结构：预训练好的LLM + 预训练好的ViT + 可训练的MLP Projection



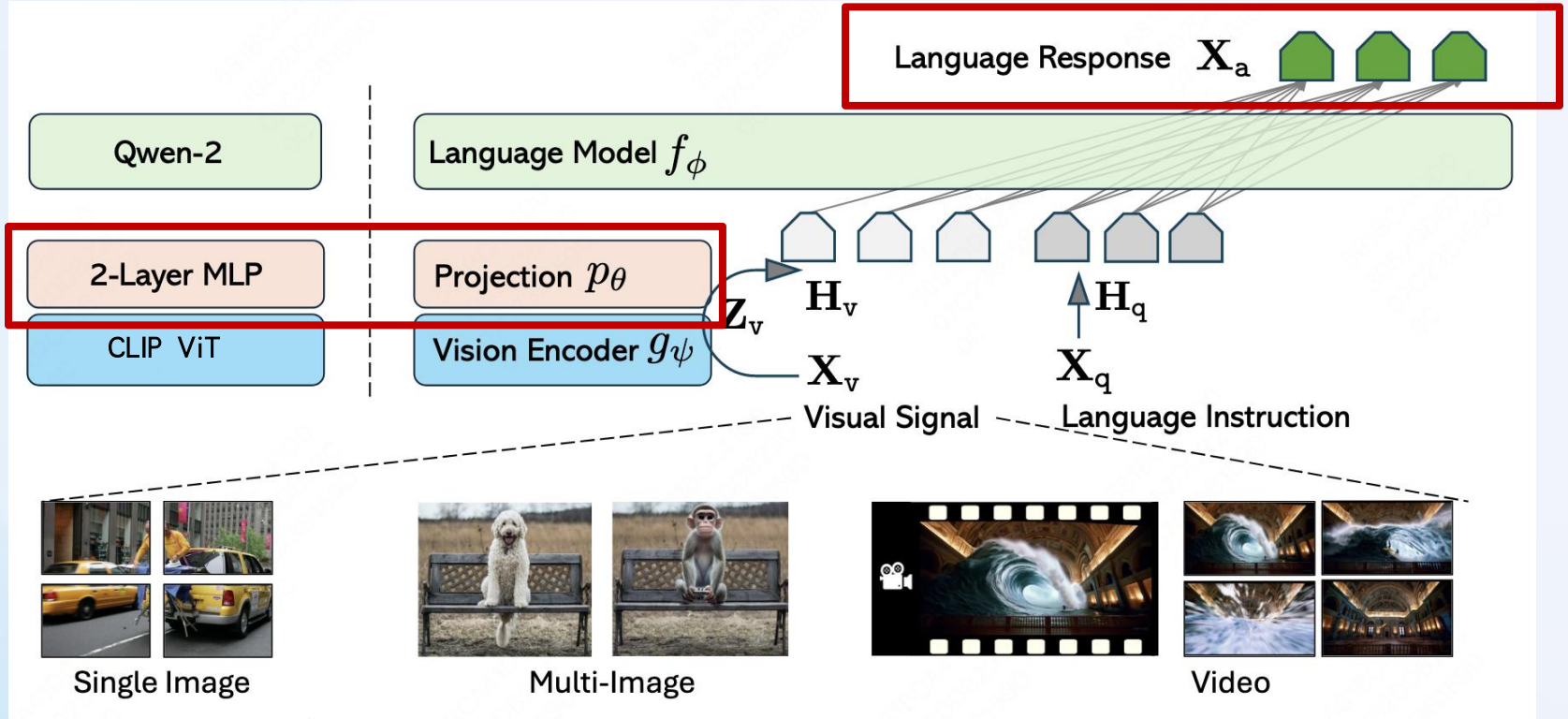
- 引入CLIP ViT作为视觉编码器来提取图像特征
- 整个视觉编码器作为一个语义级Tokenizer 实现信息压缩 (H,W,3) -> (H/14, W/14, C)



Liu, Haotian, et al. "Improved baselines with visual instruction tuning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024.

从CLIP到LLaVA多模态大模型

整体结构：预训练好的LLM + 预训练好的ViT + 可训练的MLP Projection



- 引入CLIP ViT作为视觉编码器来提取图像特征
- 整个视觉编码器作为一个语义级Tokenizer 实现信息压缩 (H,W,3) -> (H/14, W/14, C)
- MLP Connector：进一步的图文特征对齐
- 将图像特征嵌入到LLM的语境中，通过提示（prompt）的方式引导LLM生成与图像相关的文本。

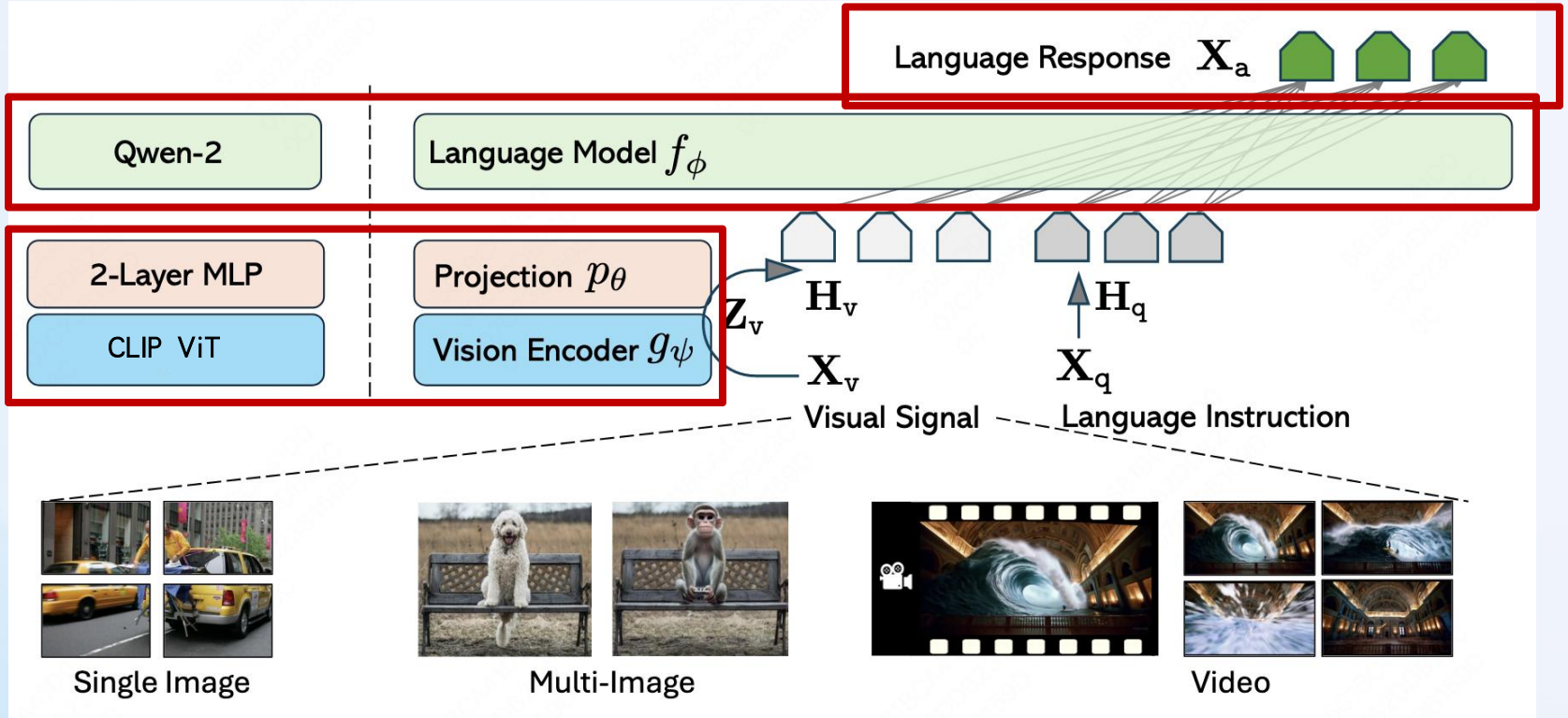
第一阶段通过大量图文描述对进行 vision-language alignment训练

Liu, Hao [illegible] rovec baselines with visual instruction tuning." *Proceedings of the IEEE [illegible] ce on computer vision and pattern recognition*. 2024.

这张图片展示了一位穿着黄色上衣的人在一辆黄色出租车后方，正站在一个架起来的梯子旁边。这个人似乎在晾晒衣物，衣物挂在梯子上。背景中可以看到城市的街道和建筑物，还有其他出租车在路上行驶。图片给人一种即兴生活场景的感觉。

从CLIP到LLaVA多模态大模型

整体结构：预训练好的LLM + 预训练好的ViT + 可训练的MLP Projection



- 引入CLIP ViT作为视觉编码器来提取图像特征
- 整个视觉编码器作为一个语义级Tokenizer实现信息压缩 (H,W,3) -> (H/14, W/14, C)
- MLP Connector：进一步的图文特征对齐
- 将图像特征嵌入到LLM的语境中，通过提示（prompt）的方式引导LLM生成与图像相关的文本。

第二阶段通过大量Instruction Data （图文问答）来加强整体的指令遵循，对话能力

Q: 图中有几个车辆？ A: 两个

Q: 图中右边车辆的车牌号是多少？ A: XXXXXXX

LLaVA多模态大模型的延续：QwenVL，InternVL系列

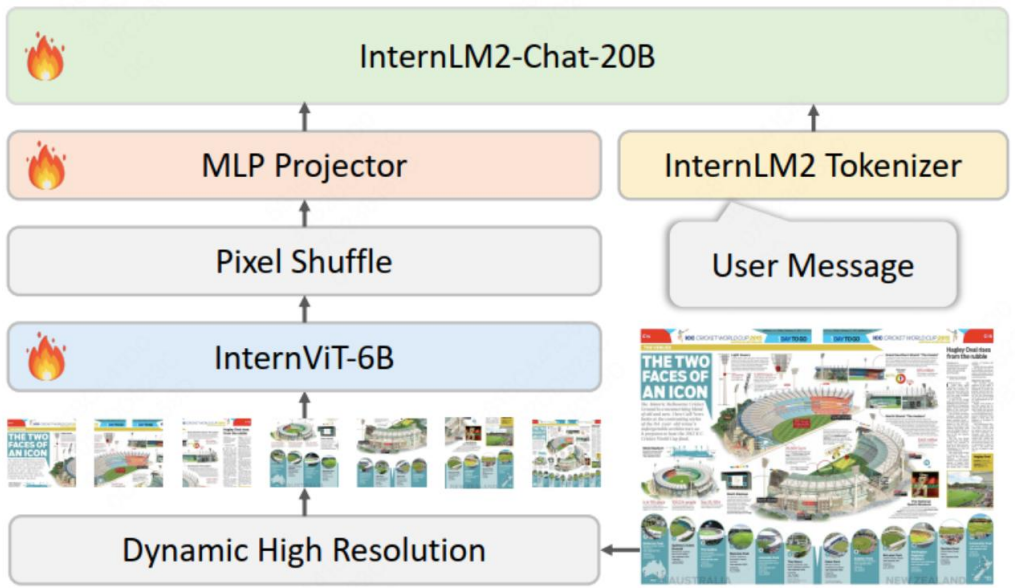


Figure 3. **Overall Architecture.** InternVL 1.5 adopts the ViT-MLP-LLM architecture similar to popular MLLMs [62, 64], combining a pre-trained InternViT-6B [18] with InternLM2-20B [11] through a MLP projector. Here, we employ a simple pixel shuffle to reduce the number of visual tokens to one-quarter.

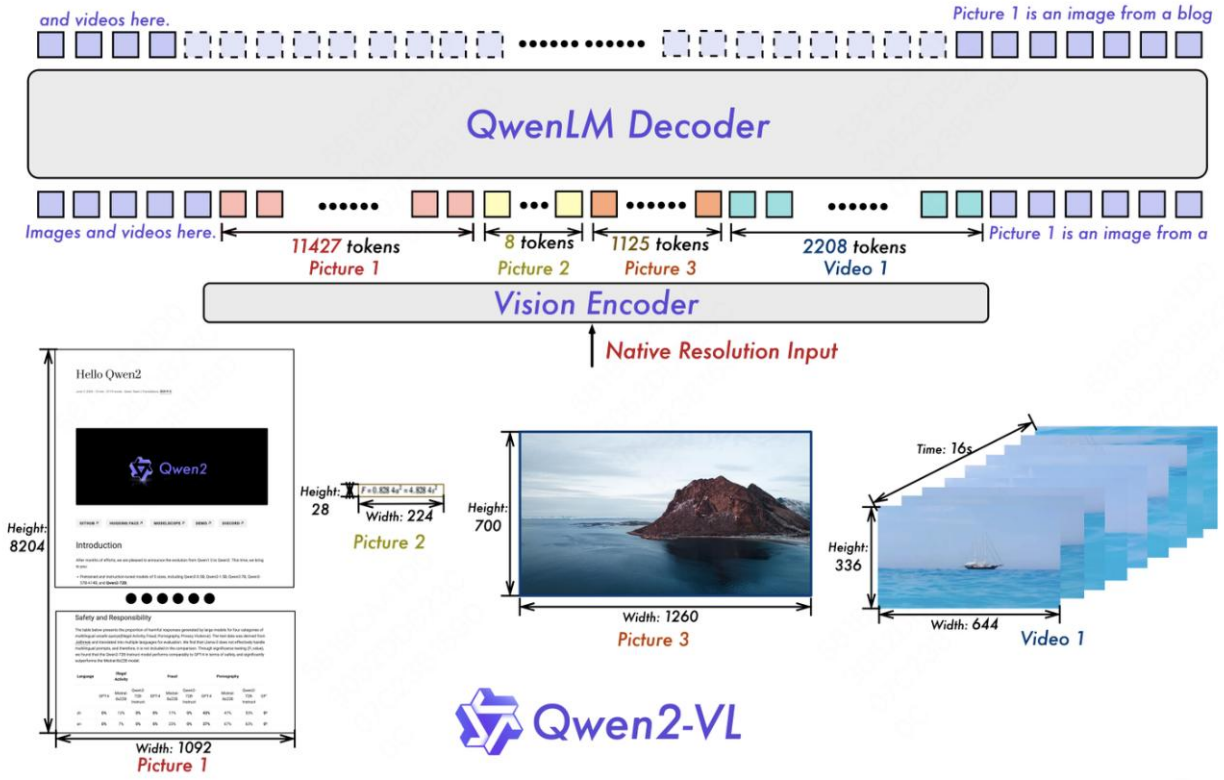
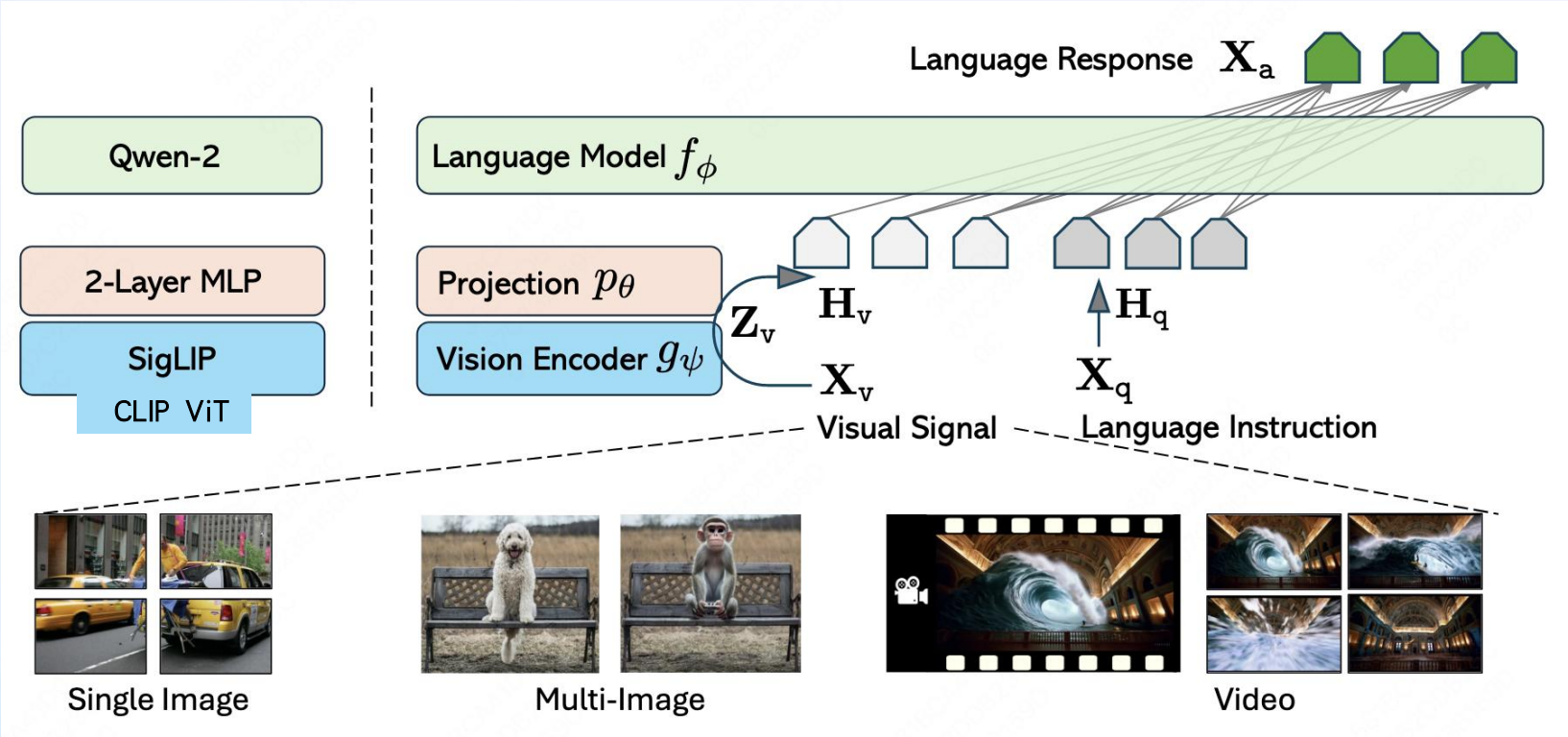


Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.

后续主流方法进一步在 ViT结构、原生分辨率、位置编码、长上下文、视觉token压缩、训练策略等多个维度进一步优化

LLaVA式的多模态大模型的弊端：



1.ViT的必要性？

ViT本身作为一个信息瓶颈，主要包含语义high-level的信息，损失掉了浅层low-level信息，对于OCR、Grounding等细粒度任务不利

Encoder-Free结构

2.将视觉特征空间投影到文本空间是否合理？

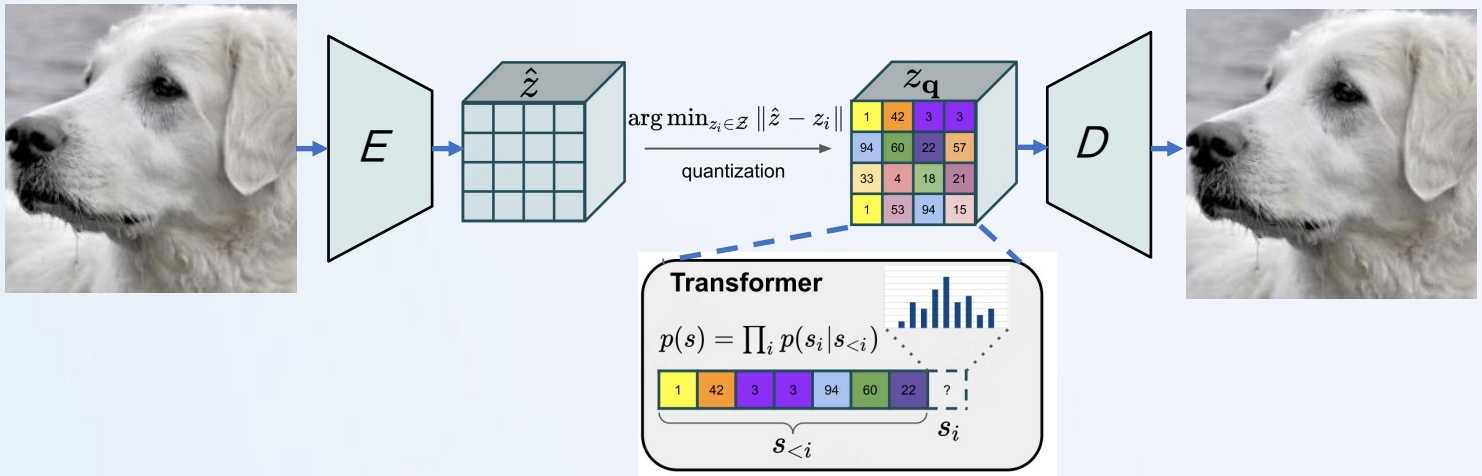
原生多模态训练

3.图像仅作为输入来执行VQA图文问答，仅支持文本生成，如何实现图像生成？原生多模态生成理解一体化模型

生成端：Diffusion vs AutoRegressive

1.基于Auto Regressive (AR) 的图像生成 （VQVAE、VQGAN、LlamaGen）

离散的图像Tokenizer+离散Codebook空间下的自回归生成



自编码器Tokenizer差异：

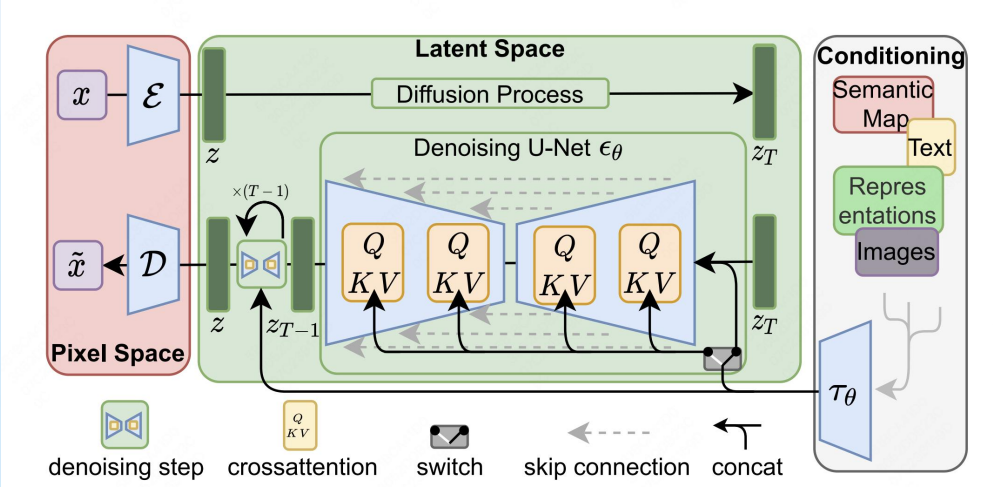
AR： 离散Tokenizer、矢量量化

Diffusion： 连续Tokenizer、无需量化

分布建模方式差异

AR： 序列预测

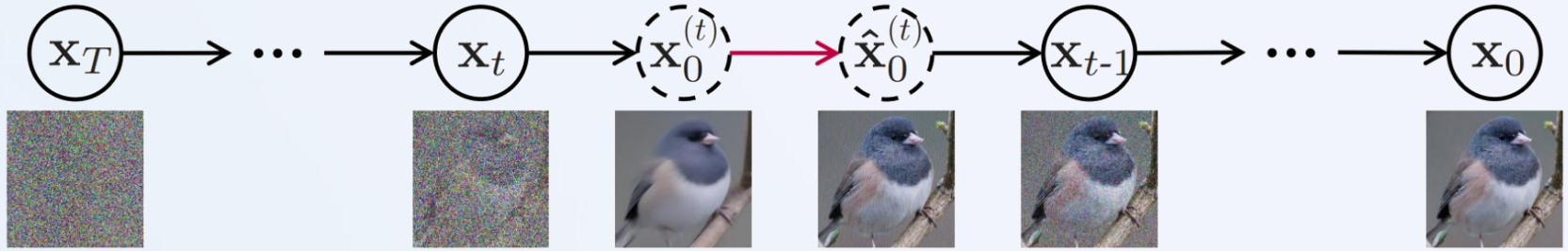
Diffusion： 去噪过程（ODE、SDE）



2.基于Diffusion的图像生成 (Stable Diffusion、DiT）

连续的图像Tokenizer+连续Latent空间下的Diffusion去噪生成

生成端：Diffusion的问题



传统DDPM需要多步denoising，计算复杂

蒸馏训练

预训练扩散模型

$$\nabla_{x_t} \log q(x_t|y_\sigma) = t^{-2}\left(\mathbb{E}\left[x_0 \,\middle|\, y_{\sigma_{eff}} = \frac{\sigma^{-2}y_\sigma + t^{-2}x_t}{\sigma^{-2}+t^{-2}}\right] - x_t\right)$$

生成式去噪器

$$\nabla_\theta D_{KL}(p_\theta(x_t|y_\sigma) \,||\, q(x_t|y_\sigma))$$

加噪图片 y_σ

$z \sim N(0, I)$

生成式去噪器

$x_t \rightarrow \cdots \rightarrow x_T$

前向扩散过程

梯度下降

优化参数 θ

蒸馏 预训练扩散模型的知识

一步推理

优于 预训练扩散模型 1000步结果

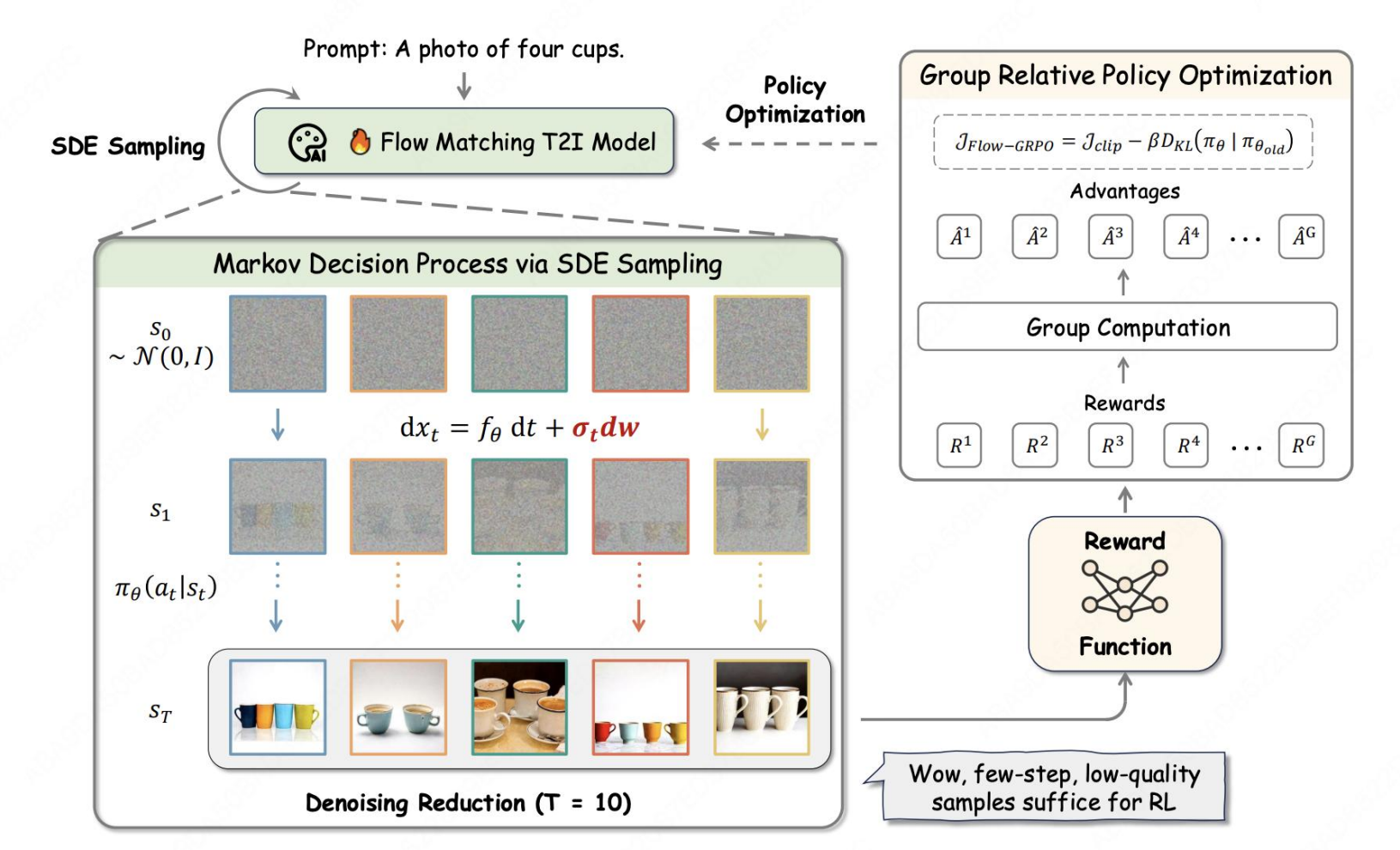
纯噪声/带噪图片

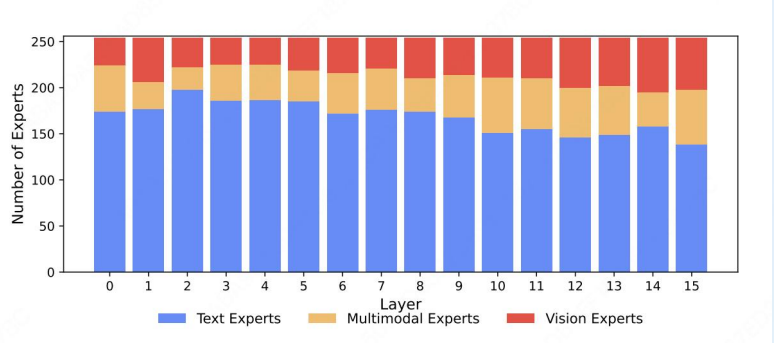
生成式去噪器

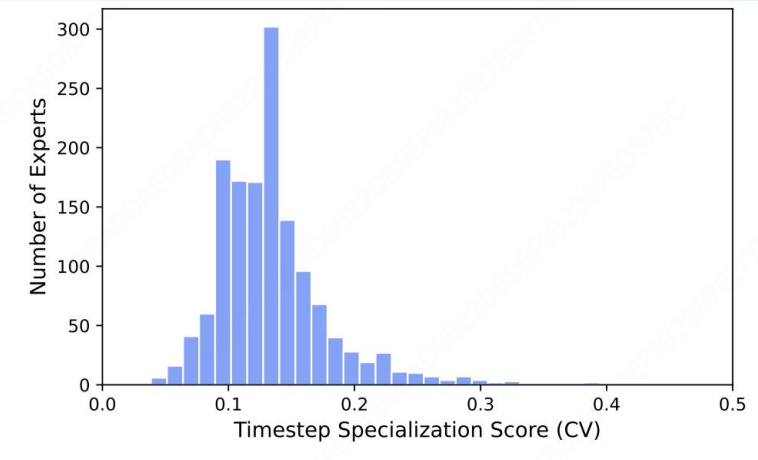
生成结果

生成端：Diffusion的问题

Flow Matching等ODE方法缺乏随机性，需要额外操作才能随机采样。

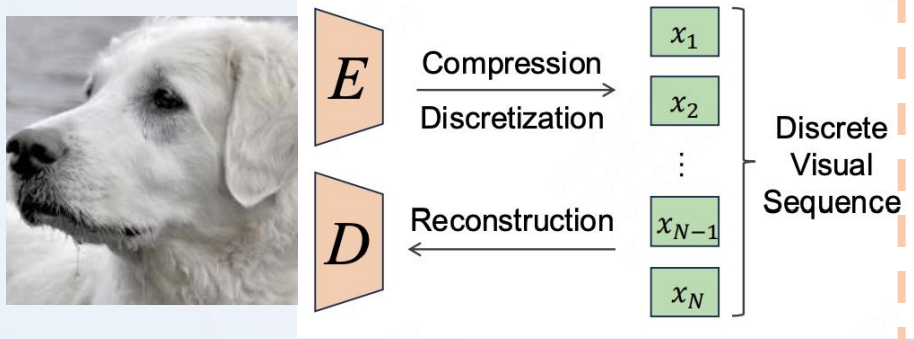






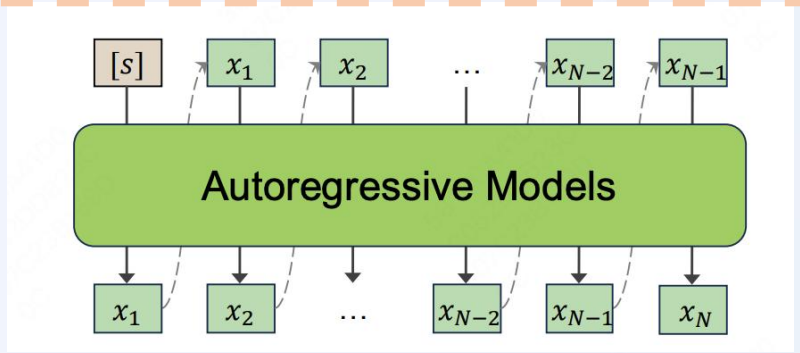
Diffusion模型没有很好的scaling能力，尤其在MoE架构上

自回归图像生成发展脉络



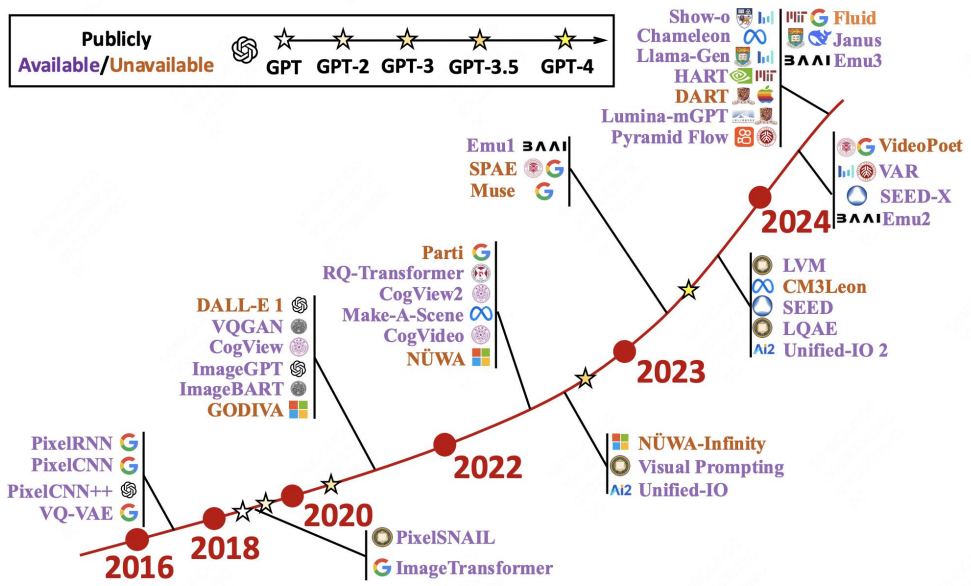
自编码器表征学习

+



自回归上下文序列建模

学习



年份	方法	自编码器	自回归建模	自回归模型参数量
2016	VQVAE	CNN—based tokenizer	CNN	O（10M）
2020	VQGAN	CNN—based tokenizer	Transformer	O（100M）
2024	LlamaGen	CNN—based tokenizer	LLM	O（1B）
2025	Infinity	CNN—based multiscale tokenizer	LLM	O（10B）

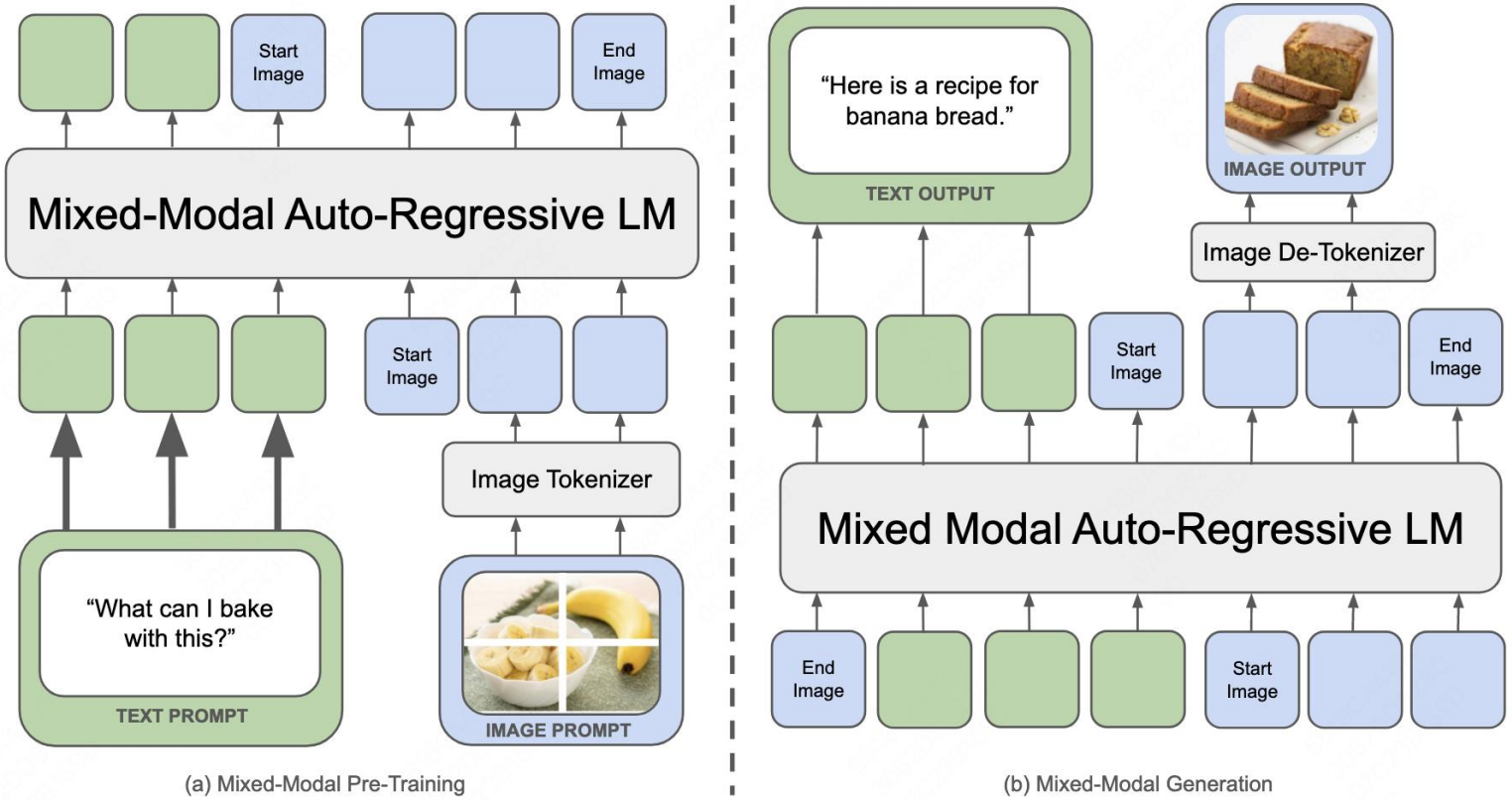
02

行业主流路线梳理

多模态图像生成和图像理解统一的早期尝试

Meta *Chameleon: Mixed-modal early-fusion foundation models*

- 自回归序列建模统一：文本生成(图->文的图像理解)和图像生成（文->图）
- 图像编码器的统一：对理解和生成统一用VQVAE实现图像编码作为tokenizer



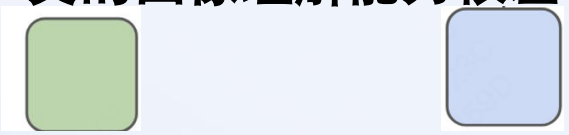
(a) Mixed-Modal Pre-Training (b) Mixed-Modal Generation

基于LLaMa-2： 7B/34B模型参数
+ 大规模文生图和图生文数据
+ 大规模联合训练（百万GPU hours）

- 离散tokenizer 信息损失严重
- VQVAE 面向重构，几乎没有语义表征能力

文->图的图像生成能力尚可

图->文的图像理解能力很差！



离散文本Token 离散图像Token

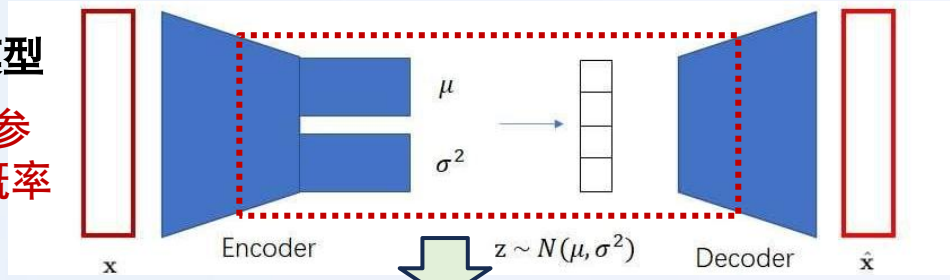
多模态图像生成和图像理解统一的早期尝试

■ 生成模型：连续隐空间到离散

➢ VAE生成模型与VQVAE模型：

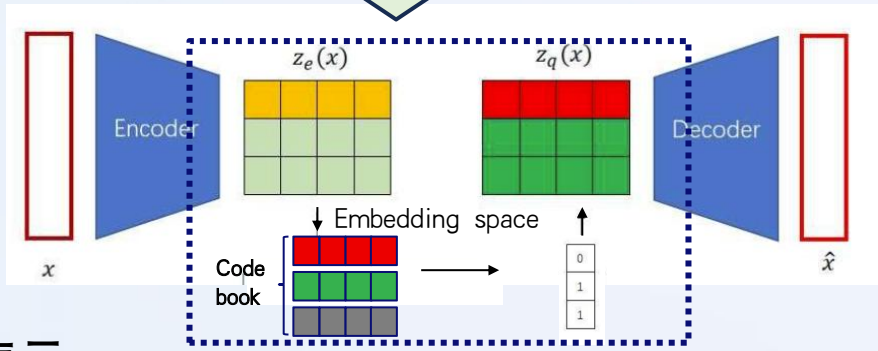
VAE生成模型

变分推断参数化连续概率

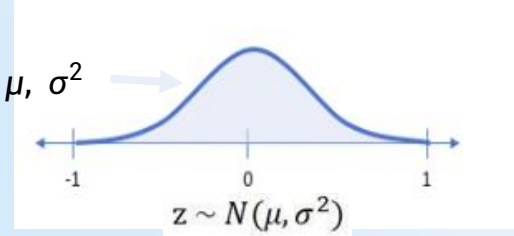


VQVAE模型

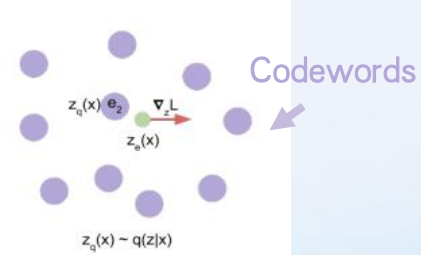
可学习码本构建离散空间



➢ 隐空间表示：



VAE：参数化高斯分布



VQ-VAE：码本选择离散码字

■ VQVAE的离散化、重构与生成

➢ 离散化：码字选择

Codebook

$$E = [e_1, e_2, \ldots, e_K]$$

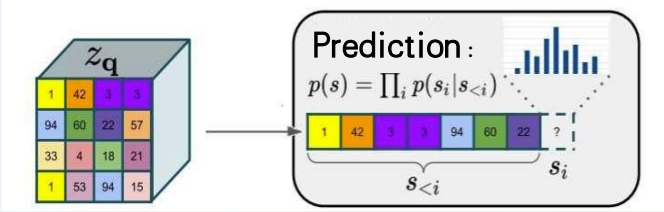
$$z \to e_k = z_q, \quad k = \underset{j}{\operatorname{argmin}} \|z - e_j\|_2$$

最邻近码字

➢ 重构：码字索引在解码端映射对应码字输入解码器

➢ 自回归生成：

自回归模型预测下一个离散码字



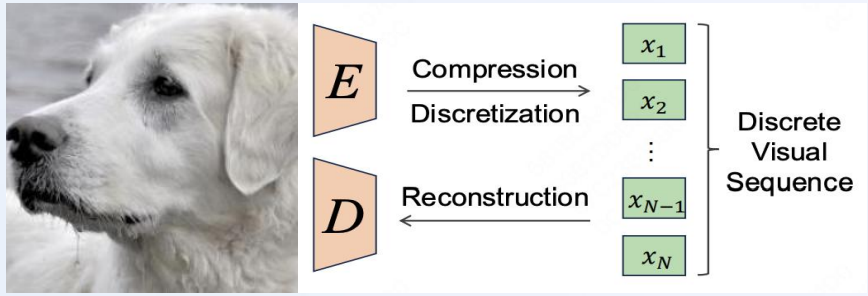
■ 离散化优势

➢ **符号对齐**：现实数据本质是离散的，离散的隐变量更适配人类对数据符号化的理解

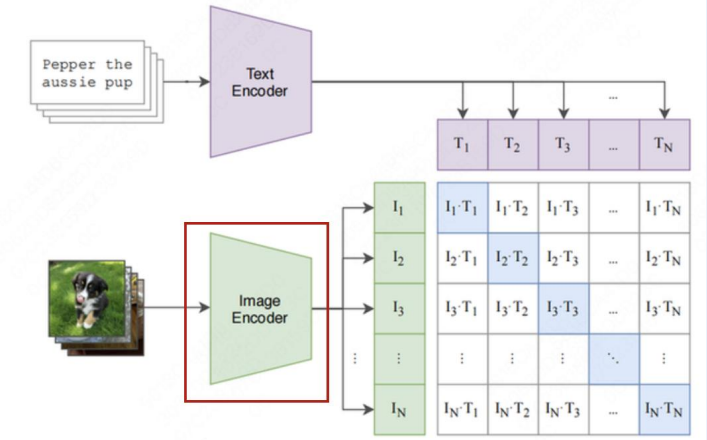
➢ **避免模糊**：连续空间导致生成结果过度平滑，而离散化能保留清晰结构

➢ 离散索引天然适配Transformer输入

图像生成和图像理解Tokenizer的异同



图像生成Tokenizer： VQVAE



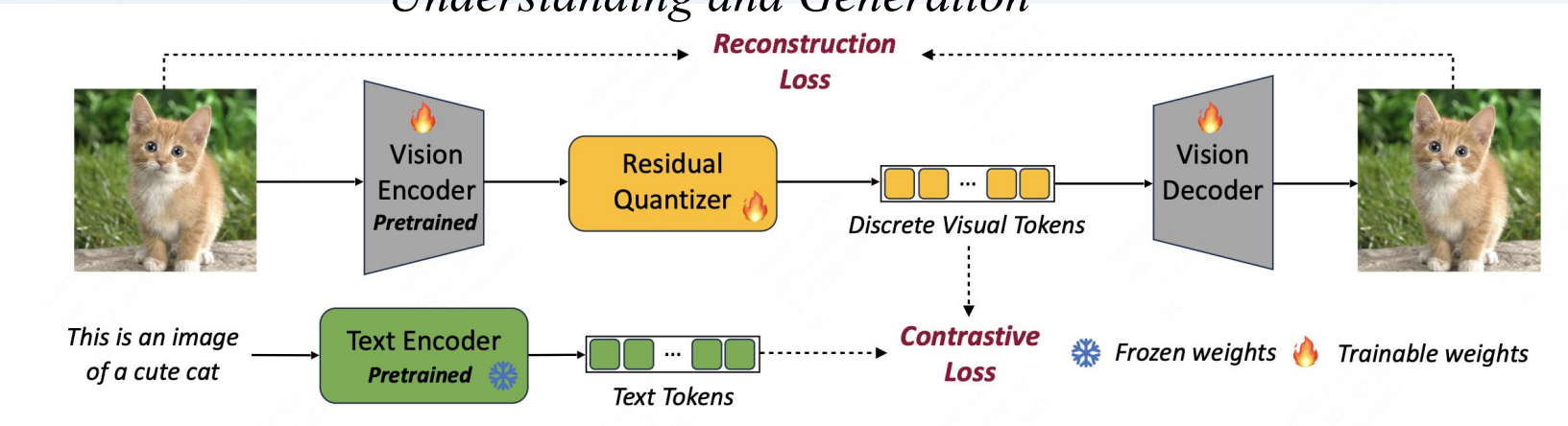
图像理解Tokenizer： ViT

	图像生成	图像理解
Tokenizer	VQVAE	ViT
loss	重构损失loss	对比学习loss
结构	编码+解码	仅编码
量化	有量化+离散特征	无量化+连续特征
粒度	Low-Level 像素级	High-level 语义级

面向重构Lowlevel特征的VQVAE tokenzier无法满足图像理解的语义特征需求

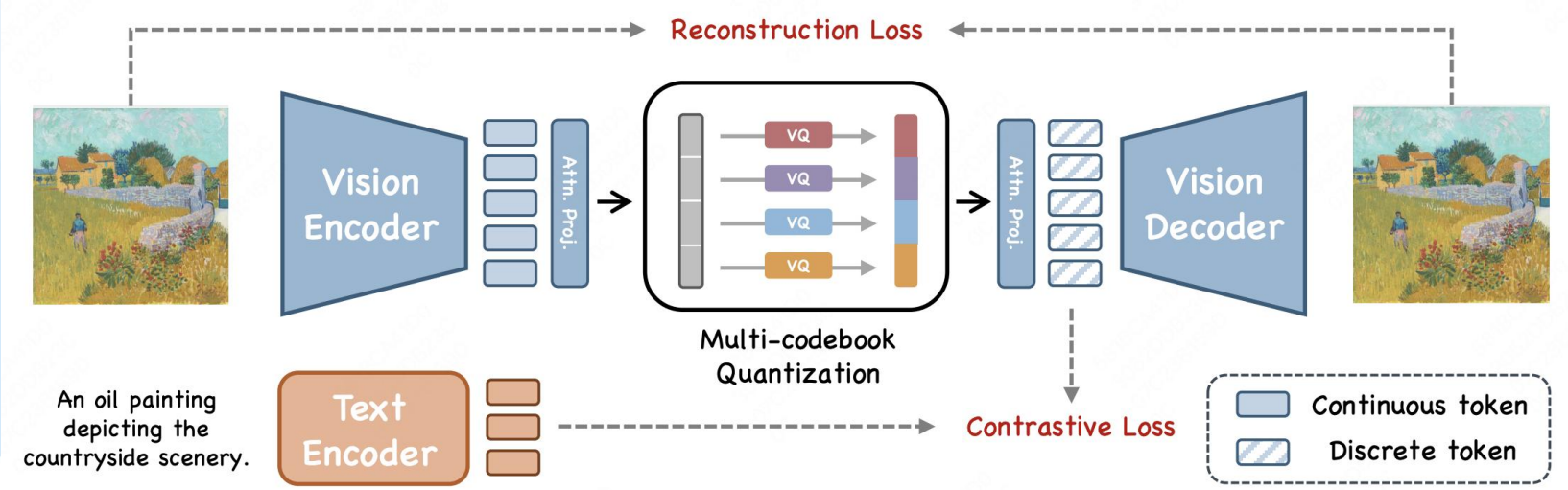
理解生成统一的图像Tokenizer

2025ICLR MIT *VILA-U: A Unified Foundation Model Integrating Visual Understanding and Generation*



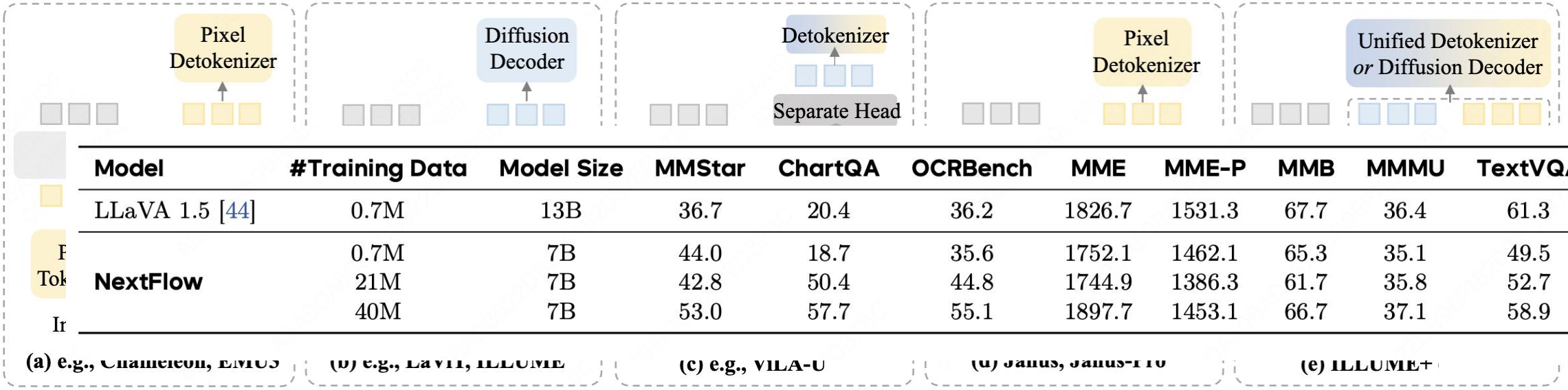
- 为预训练ViT进行VQ量化
- 同时进行图文对比和重建学习

2025NuerIPS 字节 *UniTok: A Unified Tokenizer for Visual Generation and Understanding*



- VQ量化仍然会带来信息损失
- Low-Level和High-Level表征提取能力存在冲突

理解生成统一的图像Tokenizer

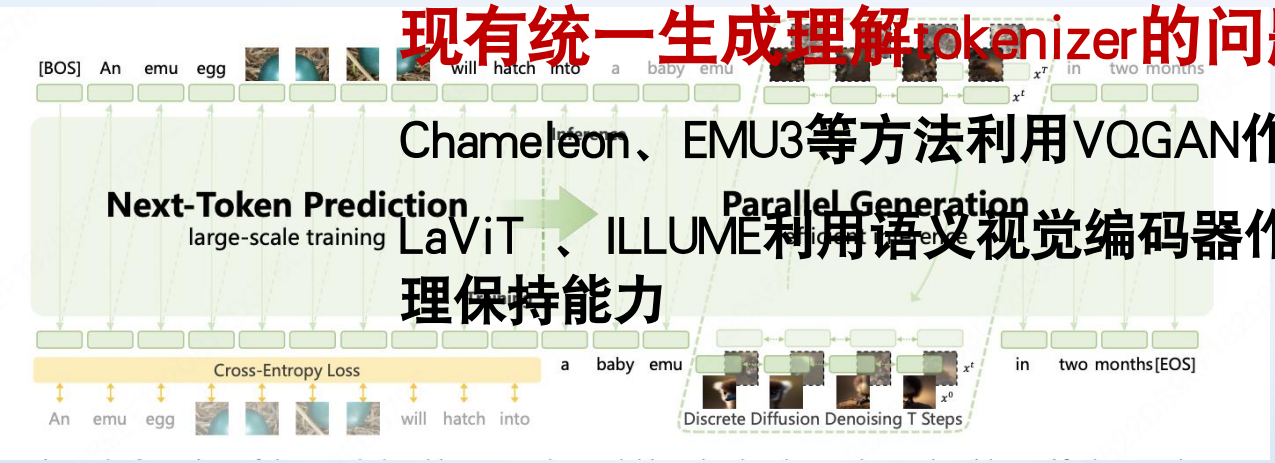


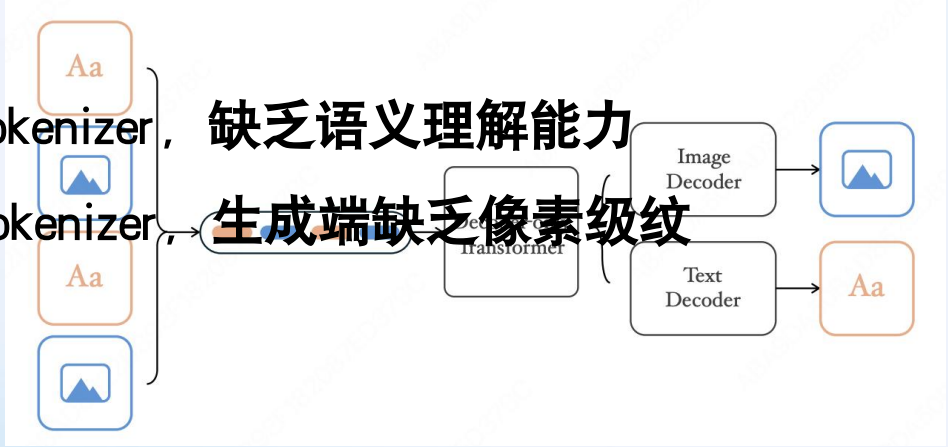
Model	#Training Data	Model Size	MMStar	ChartQA	OCRBench	MME	MME-P	MMB	MMMU	TextVQA
LLaVA 1.5 [44]	0.7M	13B	36.7	20.4	36.2	1826.7	1531.3	67.7	36.4	61.3
NextFlow	0.7M	7B	44.0	18.7	35.6	1752.1	1462.1	65.3	35.1	49.5
	21M	7B	42.8	50.4	44.8	1744.9	1386.3	61.7	35.8	52.7
	40M	7B	53.0	57.7	55.1	1897.7	1453.1	66.7	37.1	58.9

现有统一生成理解tokenizer的问题：

Chameleon、EMU3等方法利用VQGAN作为tokenizer，缺乏语义理解能力

LaViT 、ILLUME利用语义视觉编码器作为tokenizer，生成端缺乏像素级纹理保持能力

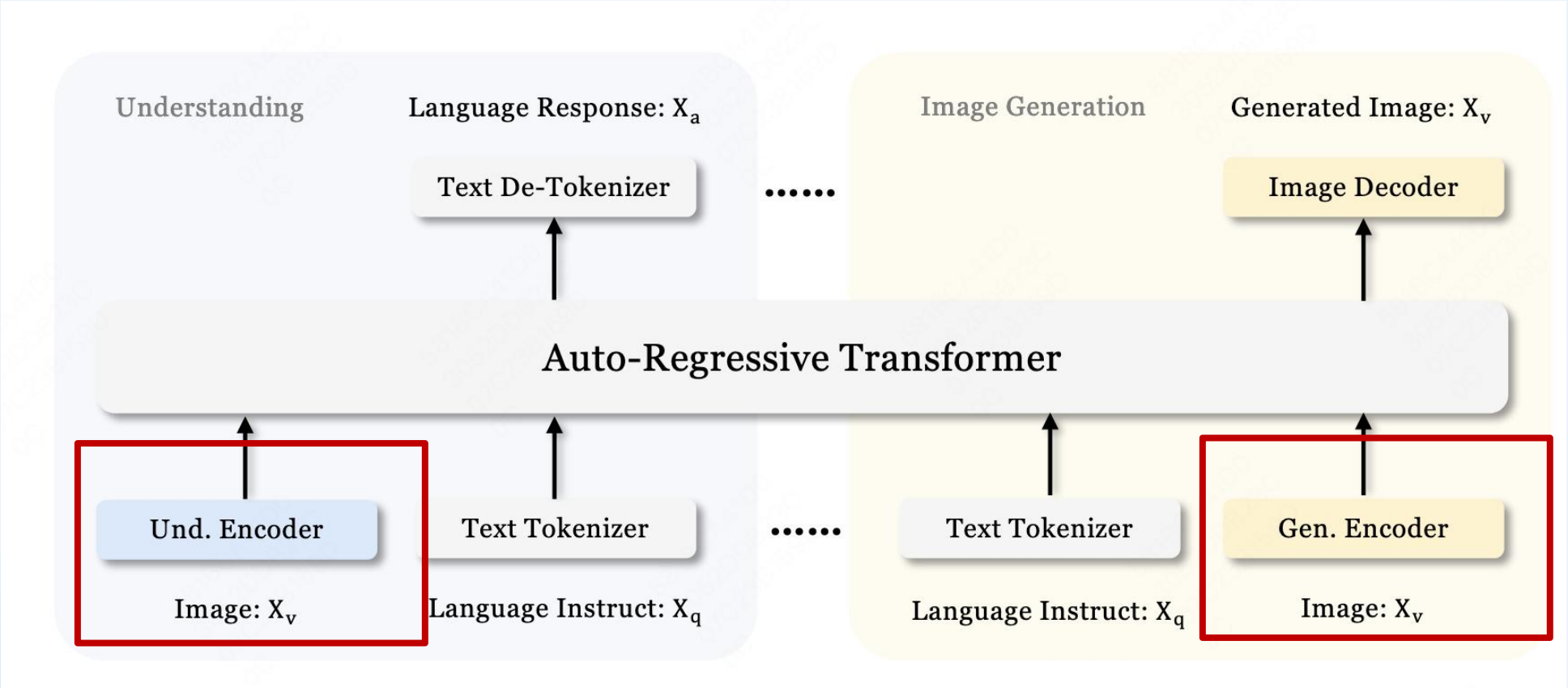




双编码器结构： NTP路线

2025 DeepSeek *Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation*

- 共享的LLM 自回归Backbone ＋ 解耦的两个Encoder

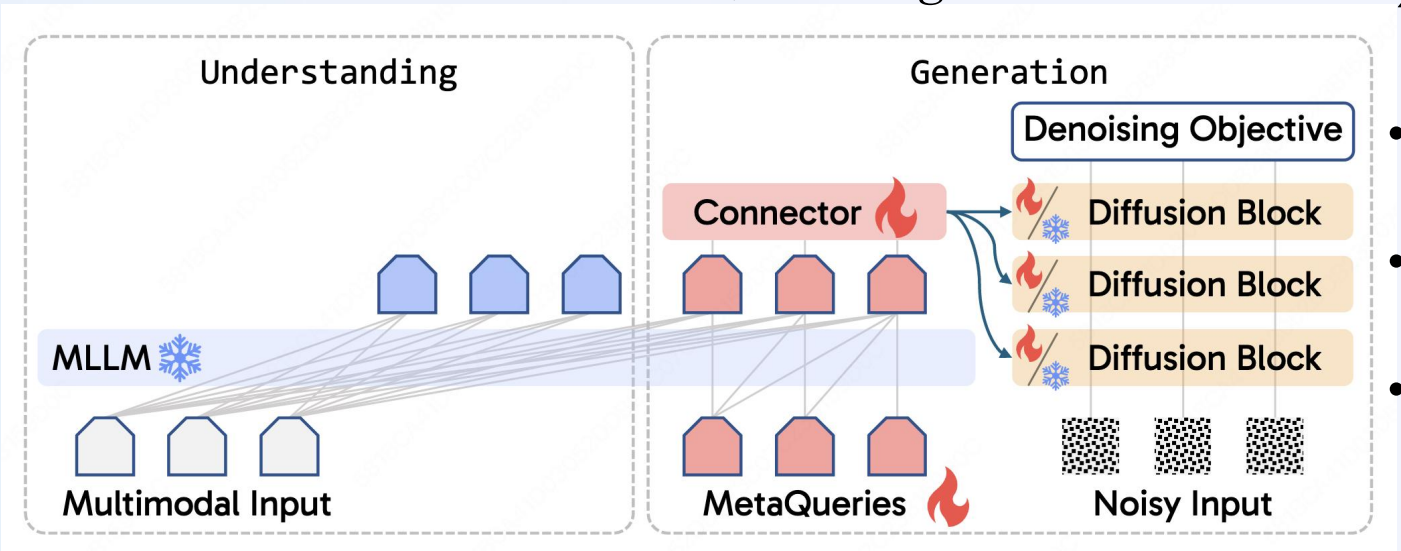


CLIP ViT
语义级HighLevel编码

VQVAE
像素级LowLevel编码

双编码器结构：AR+Diffusion

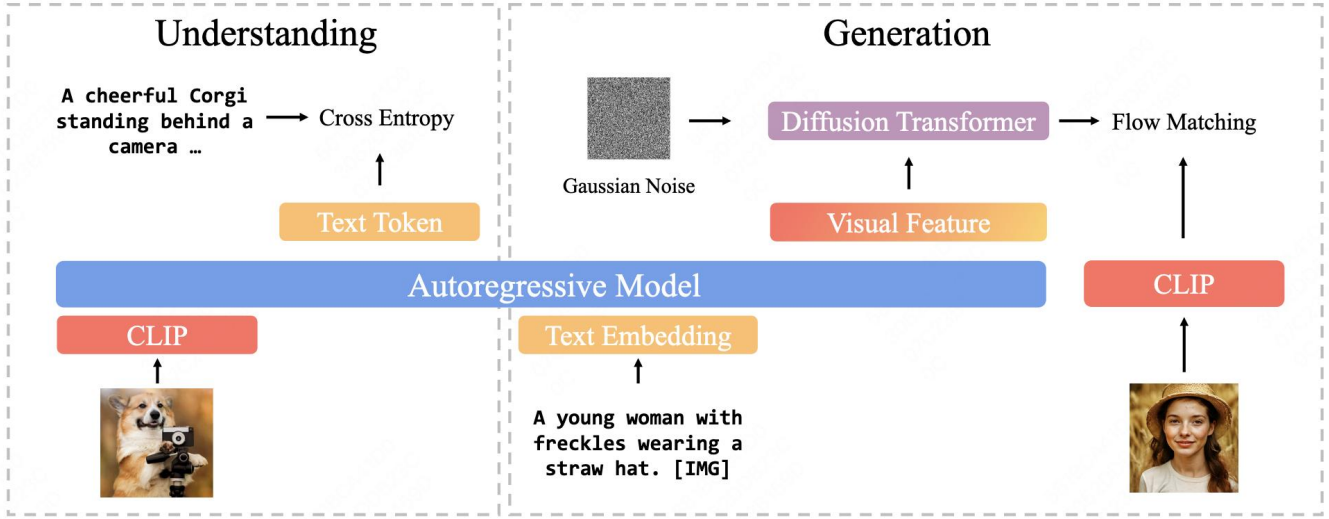
2025 Saleforce *BLIP3-o: A Family of Fully Open Unified Multimodal Models—Architecture, Training and Dataset*



2025 Qwen *Qwen-Image*
2025 Google *Nano-Banana*
……

- 利用预训练好的多模态大模型(MLLM)作为文本/图像特征提取器
- 将丰富的文本/图像语义特征作为条件信息送到Diffusion中实现图像生成
- 可借助预训练好的Diffusion实现高质量图像生成

2025 Meta *Transfer between Modalities with MetaQueries*

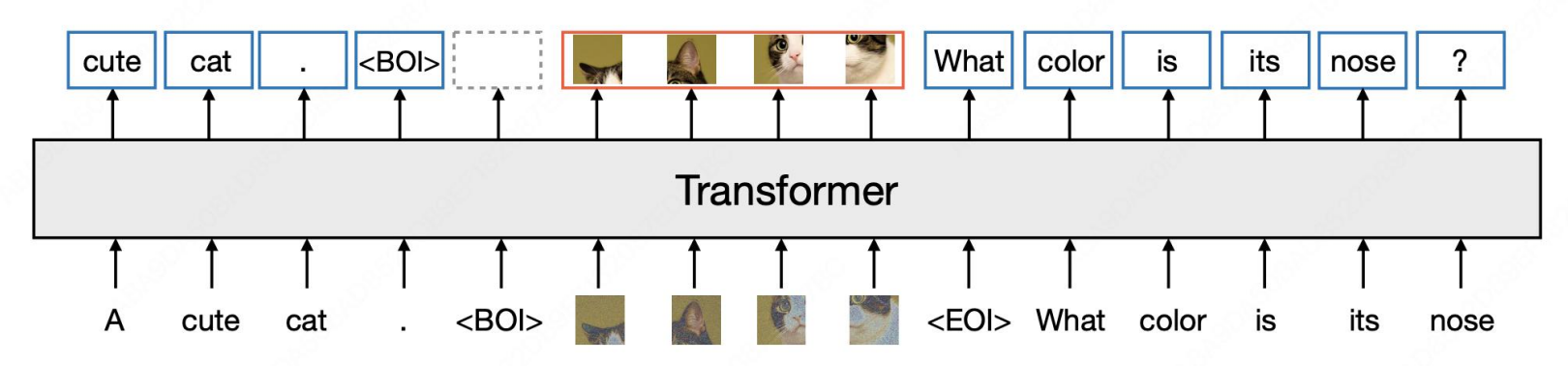


- 结构复杂不原生、MLLM并不参与训练
- 本质上是更换了离线的特征提取器

是否可以完全不采用任何外部元件，如Diffusion，ViT，实现一个尽可能简洁、更加符合第一原则性的一体化的结构？

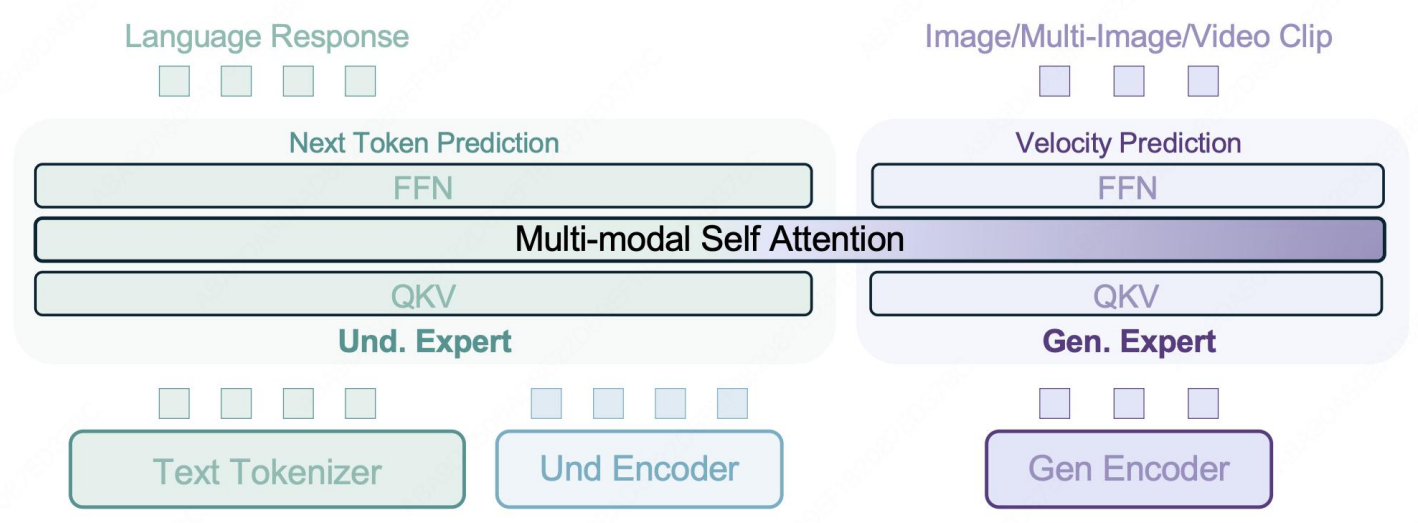
InfoQ 极客传媒

双编码器结构：AR+Diffusion



$$\mathcal{L}_{\text{Transfusion}} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}$$

联合训练两种截然不同的目标函数（NTP与扩散）具有较高的工程挑战



- MoT，生成理解完全分离的参数，只适用于dense模型，难以scaling到MoE大尺寸模型。
- Flow Matching相比DDPM更快收敛，更容易训练。但却遇到确定性限制，难以直接进行RL训练

03

Decoder-only纯自回归路线

Decoder-only纯自回归路线

- 多模态大模型提取图像/文本特征作为扩散生成控制条件
- 外接预训练扩散生成模型或ViT：**难以联合优化实现多模态深度**

➢ **外接扩散生成模型**

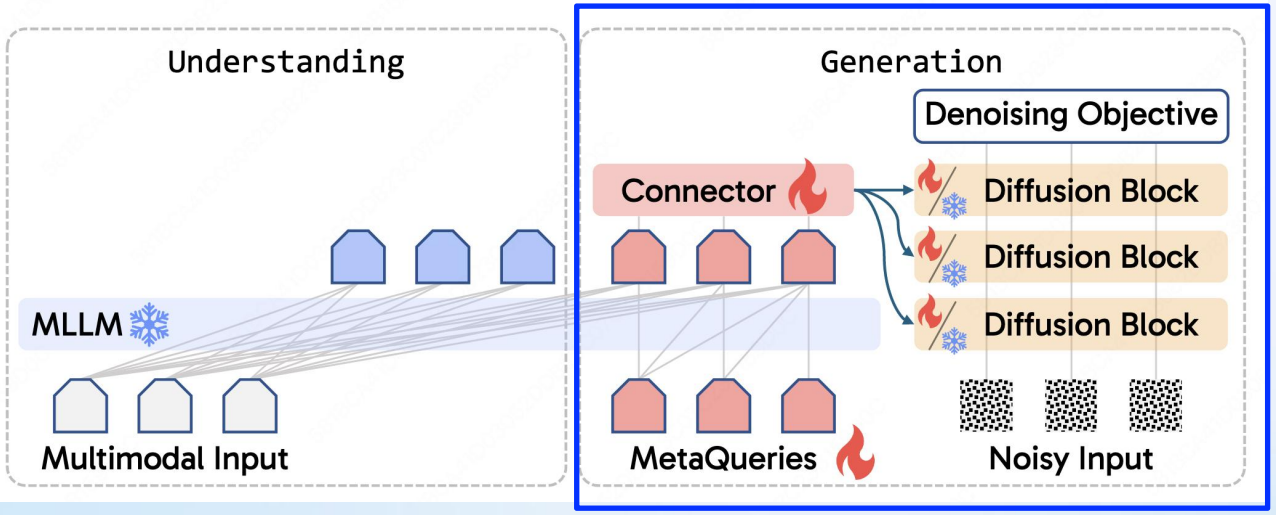
MetaQueries [Meta 2025]

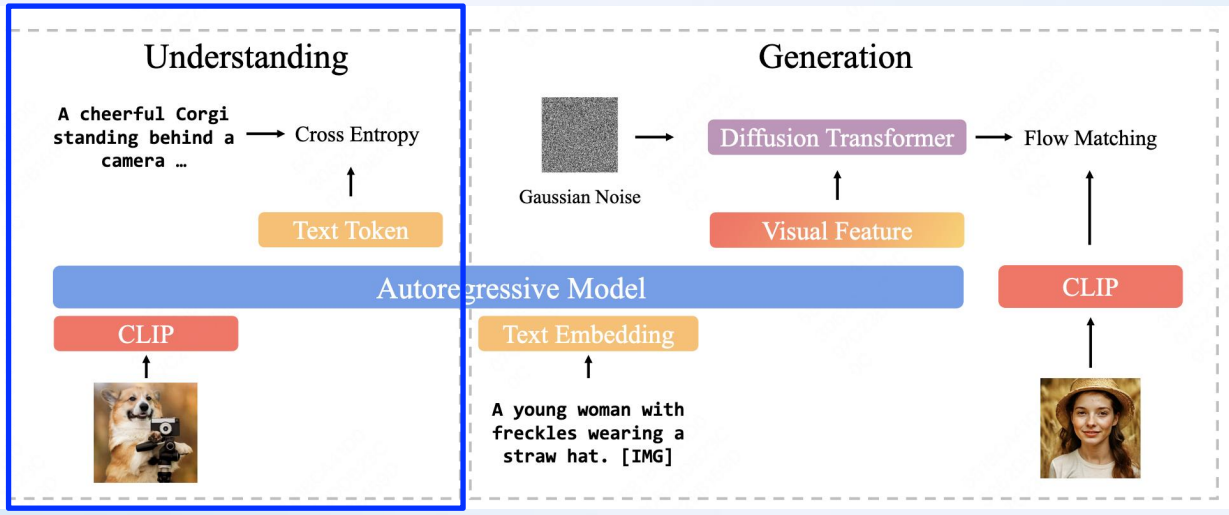
可学习查询（Queries）连接多模态大模型与扩散生成模型

➢ **外接ViT用于视觉理解**

BLIP3-o [Salesforce 2025]，Qwen-Image [Qwen 2025]，Nano-Banana [Google 2025]

额外CLIP ViT用于视觉理解





实现单一多模态大模型统一理解与生成

Decoder-only纯自回归路线

- 统一架构：多模态大模型的单一解码器
- 无需视觉编码器，支持图像原生分辨率，实现多模态特征早期融合
- 难点：如何统一图像理解与生成Token的训练、预测

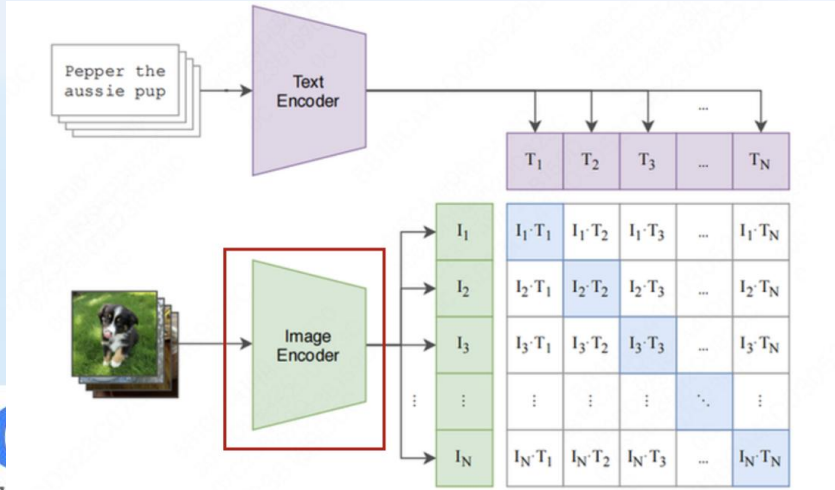
图像理解 Token

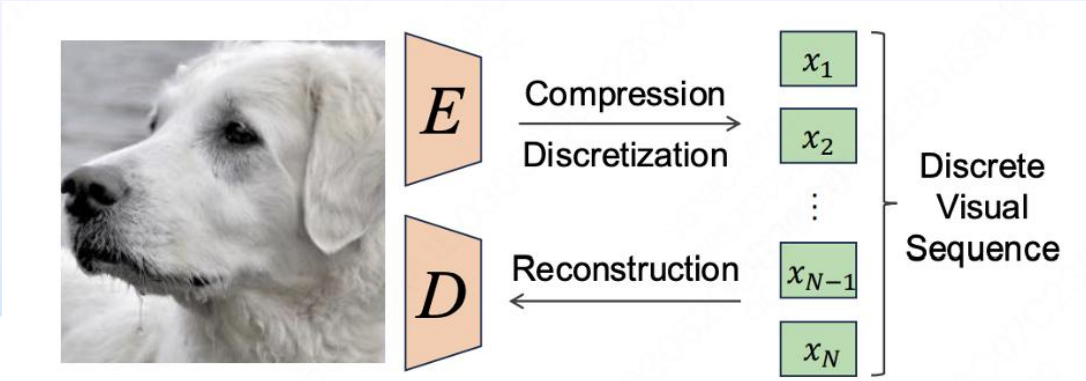
- 连续特征
- 高层语义信息，不关注精细尺度细节
- 下一个Token预测 （Next—Token Prediction）

矛盾！

图像生成 Token

- 离散特征
- 重建精细尺度细节，不关注高层语义信息
- 下一个尺度预测（Next—Scale Prediction）

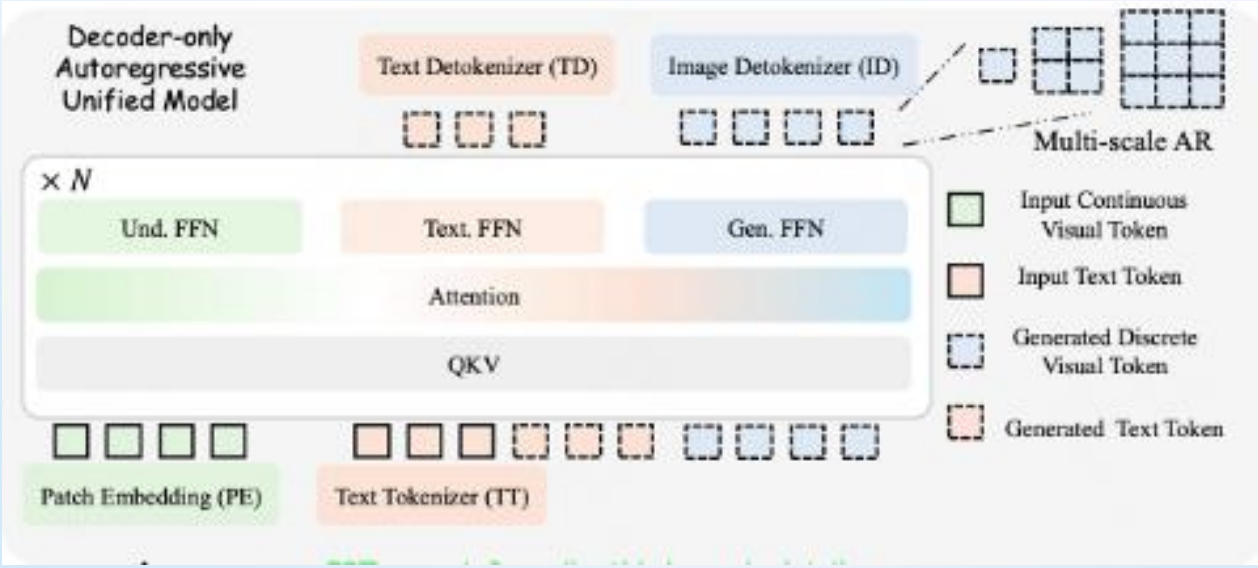


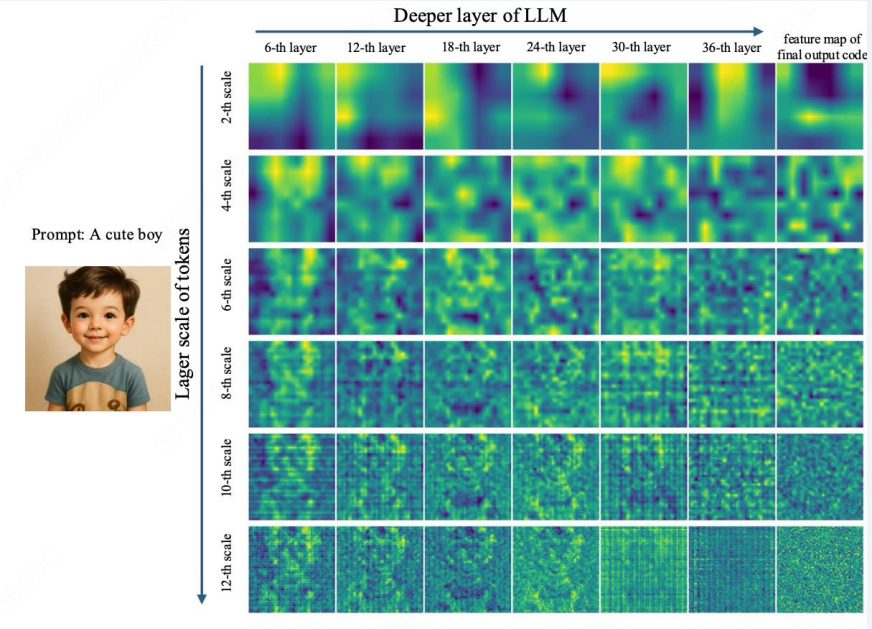


全球软件开发大会

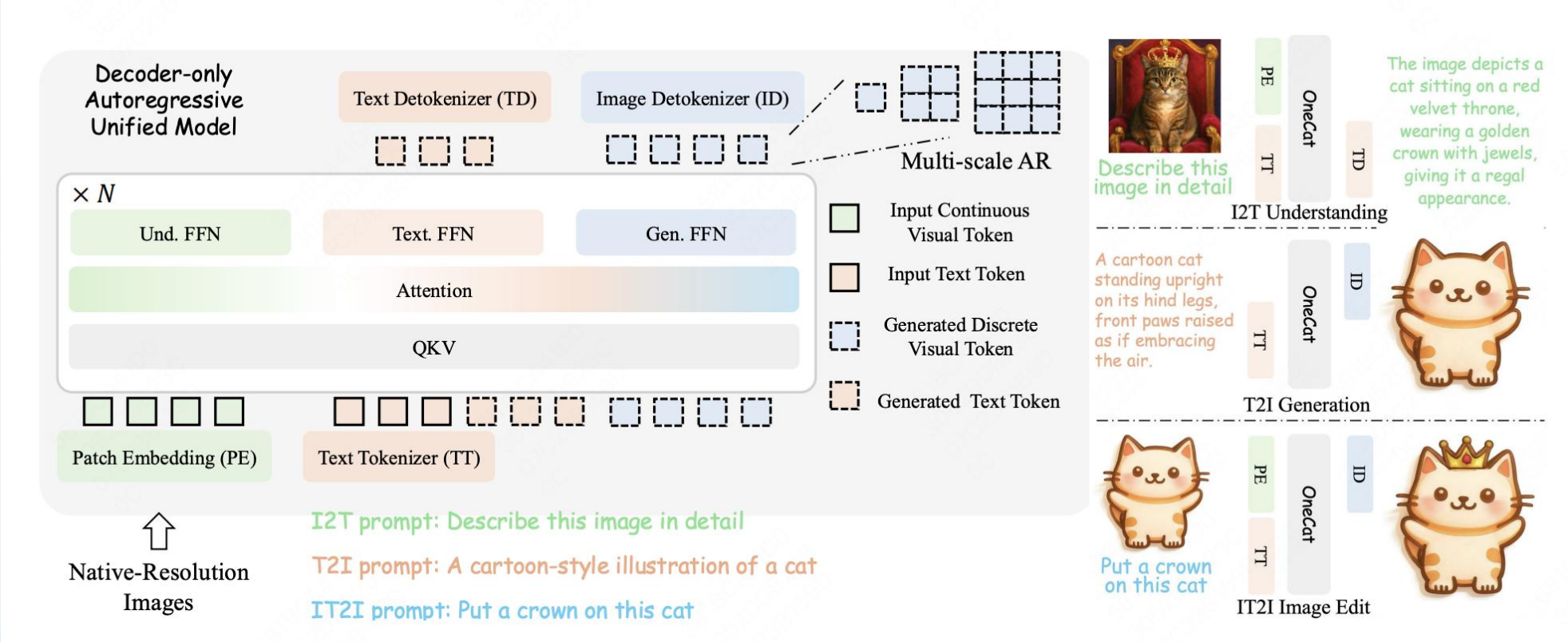
Decoder-only纯自回归路线

- **统一架构：** 多模态大模型的单一解码器
- **混合模态专家模型：** 共享QKV和自注意力机制，仅采用独立FFN专家提取连续图像理解、离散图像生成、文本Token
- **统一下一Token预测与下一尺度预测，尺度自适应适配实现多尺度生成**



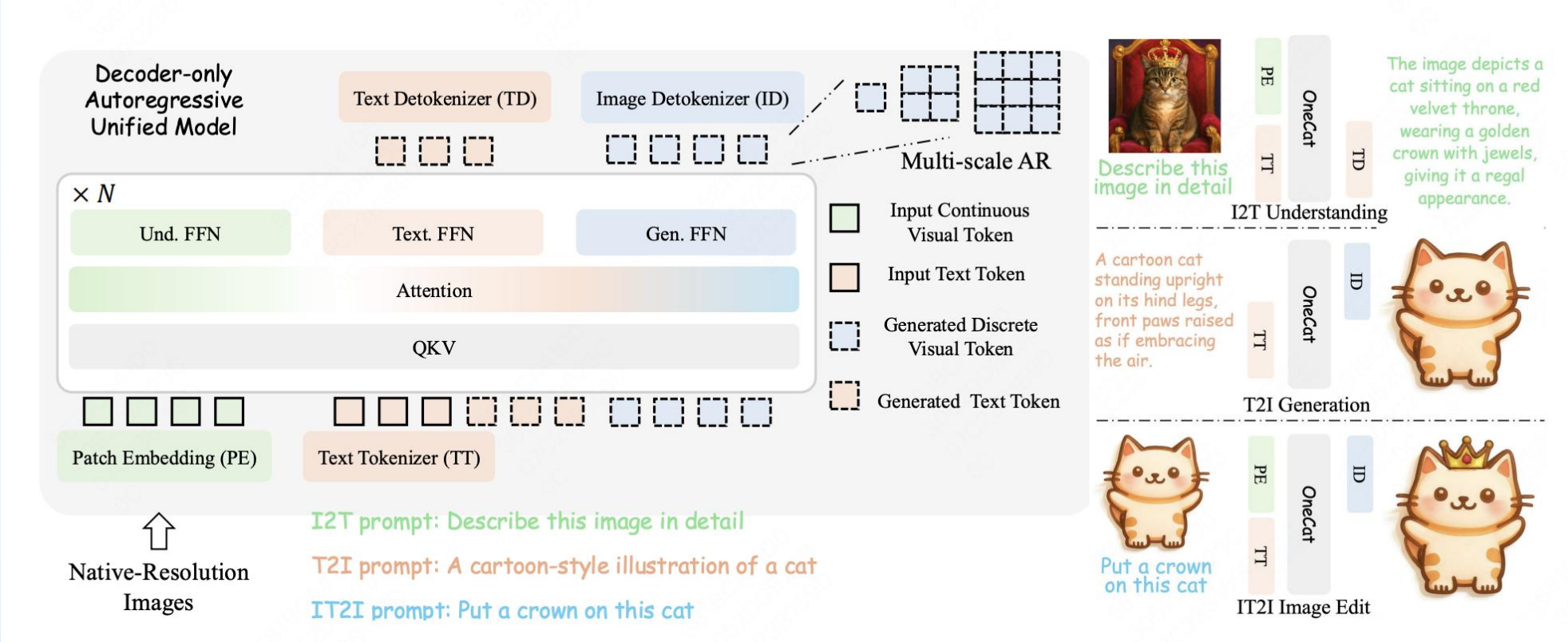


Decoder-only纯自回归路线



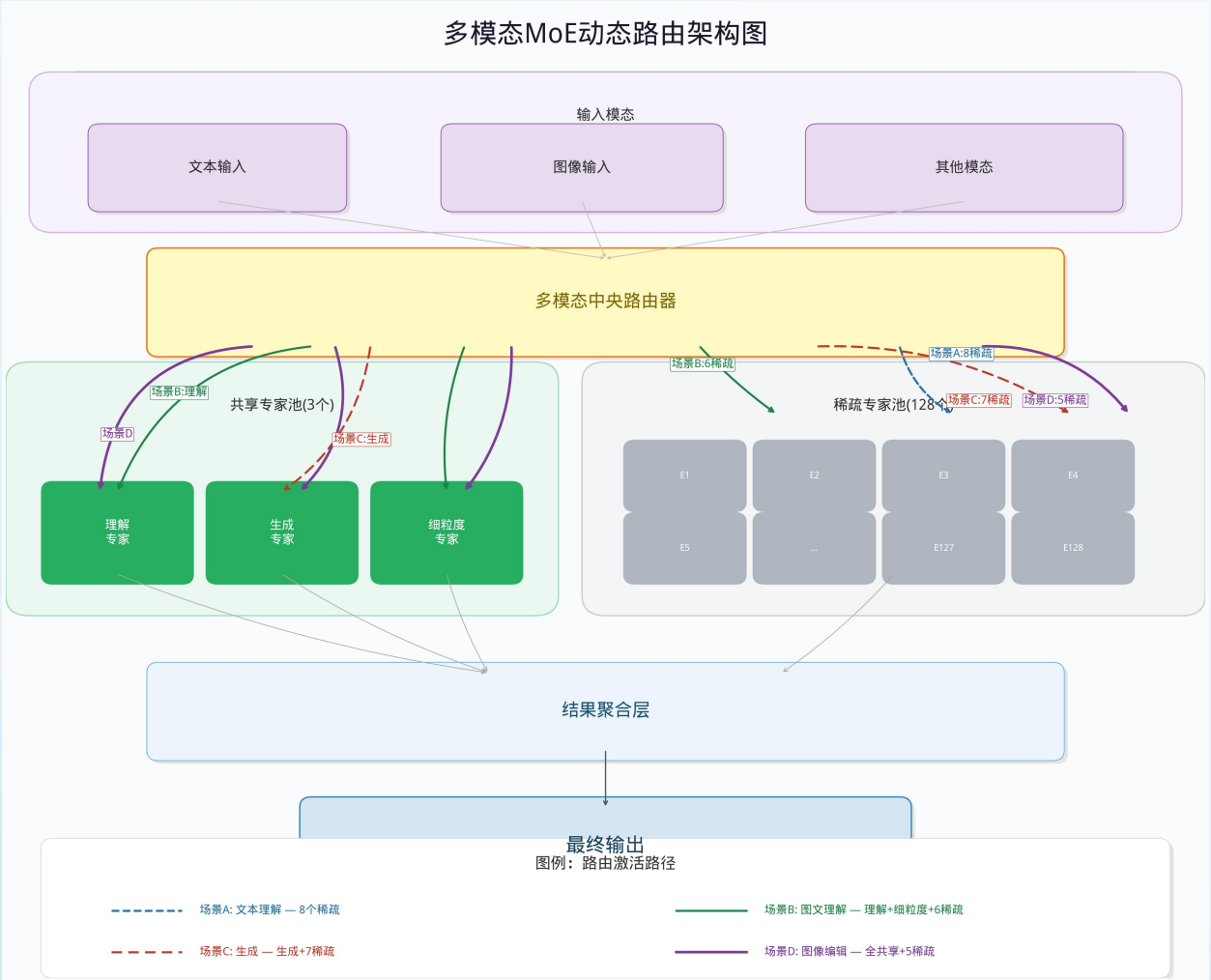
- 1.推理无需视觉编码器，及视觉tokenizer的原生多模态模型，推理速度提升
- 2.模态MoE架构：文本、视觉理解、图像生成共享QKV层及MHSA层，实现准确的指令遵循。但各自采用独立的FFN Expert保证模态独立特征的有效提取
- 3.统一自回归+统一loss：文本输出采用标准的NTP进行训练，图像生成采用NSP O (HW^2)->O (L)

Decoder-only纯自回归路线



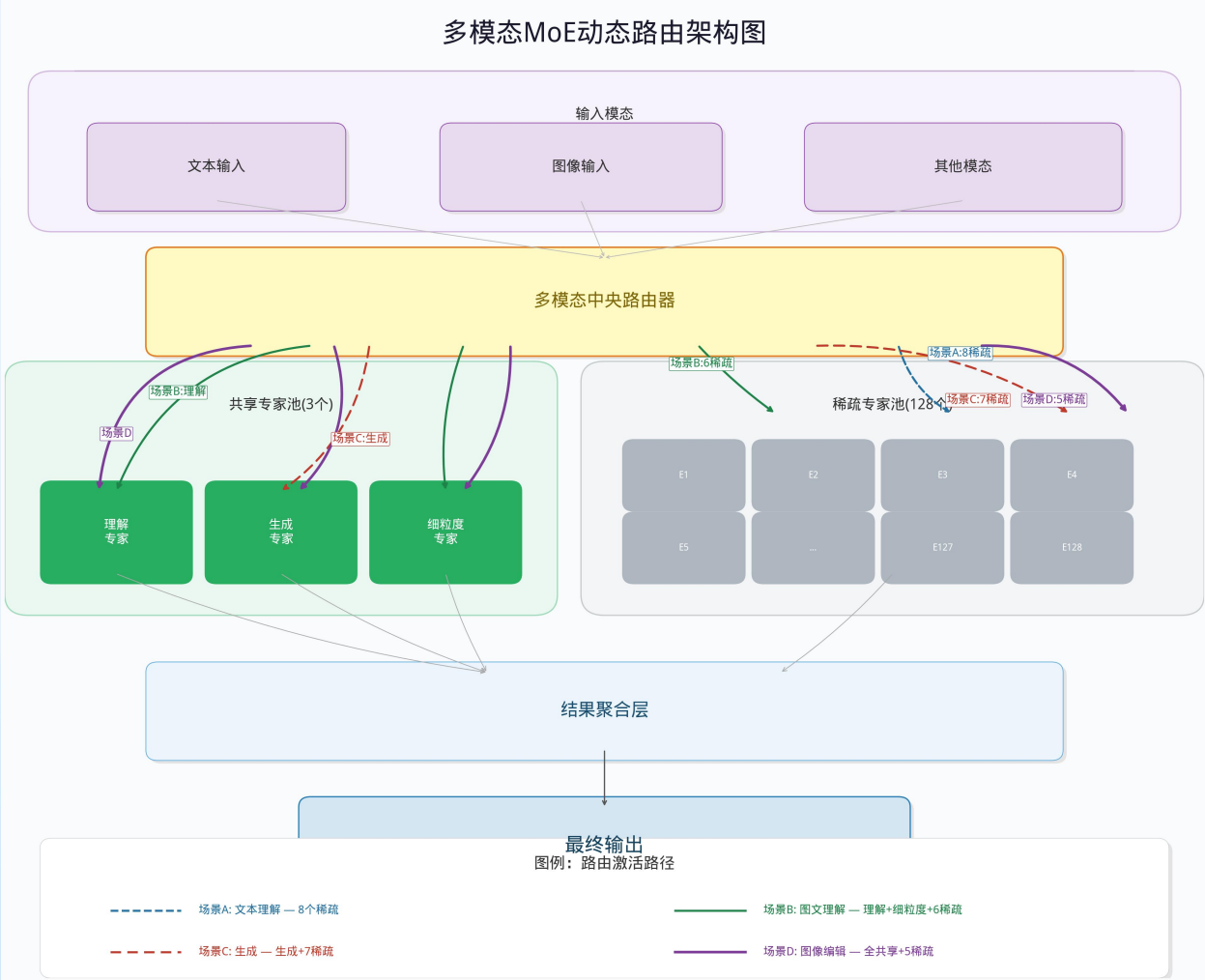
- 4.图像生成端，对于不同scale的token，在Gen Expert的基础上，引入额外的scale-aware lora层提取不同尺度特有的特征 （7～13个scale对应256px～1024px）
- 5.原生分辨率理解、原生分辨率生成　生成最高支持1024px

分模态共享MoE架构



- 本架构采用多模态混合专家（Mixture of Experts, MoE）设计，核心思想是通过一个中央路由器（Gating Network）根据输入模态的不同，动态选择并激活不同的专家组合。
- 架构包含两类专家池：共享模态专家池（3个常驻专家）和稀疏专家池（128个专家），通过差异化的路由策略实现高效的多模态处理。
- 所有被激活的专家输出经过结果聚合层（Aggregation Layer）进行加权融合，最终产生统一的输出结果。聚合层根据路由器分配的权重对各专家的输出进行动态组合（分组加权），确保不同专家的贡献被合理整合。

分模态共享MoE架构



- 共享专家池包含3个常驻专家：理解专家、生成专家和细粒度专家。理解专家侧重于语义理解与特征提取，生成专家负责内容生成与重建，细粒度专家则专注于精细化的特征对齐与融合。
- 稀疏专家池包含128个独立的专家网络，每次推理时仅激活其中一小部分（5–8个）。

四种典型的路由策略：

– 场景A（文本理解）：仅激活稀疏专家池中的8个专家，无需调用共享专家，适用于纯文本语义理解任务。

– 场景B（图文理解）：同时激活共享专家池中的理解专家和细粒度专家，以及稀疏专家池中的6个专家，用于图文联合理解。

– 场景C（生成任务）：激活共享专家池中的生成专家和稀疏专家池中的7个专家，专注于内容生成类任务。

– 场景D（图像编辑）：激活全部3个共享专家加上稀疏专家池中的5个专家，以支持复杂的多模态编辑任务。

交错理解生成

- **多模输出**：故事/海报创作时，模型可以先规划文案，再生成首图，再根据反馈继续补文字或补第二张图。
- **多轮编辑**：参考图编辑时，模型先理解输入图，再直接生成修改结果，并在下一轮继续接收局部参考图迭代。
- **增量生成**：复杂场景可以先定大结构，再逐步补细节，而不是一次性完成。
- **推理需求**：复杂视觉/空间感知任务中，模型可以生成多模态COT辅助推理，例如视角变换，辅助线等。

User：

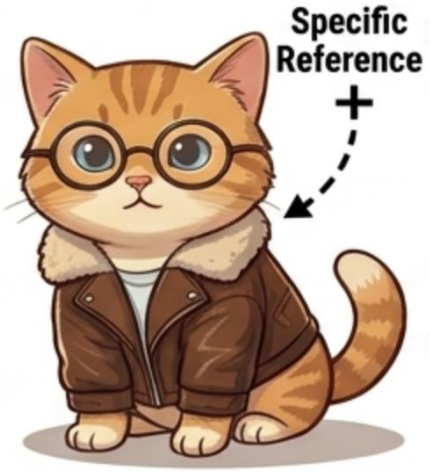
帮我生成一张戴着眼镜的猫咪图片

AI：<BOI>

<EOS>

User：帮我给这只猫穿上@夹克和@帽子，并把背景换成书房

AI：好的<BOI>



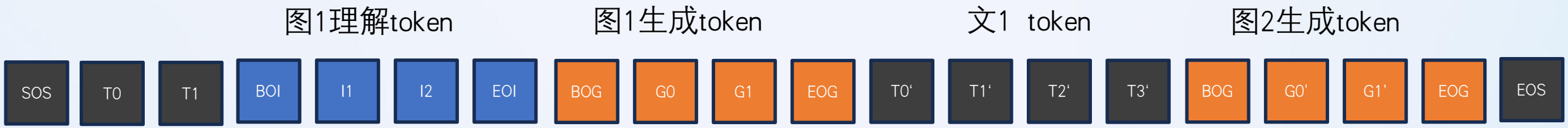
Wait，我忘记了帽子和书房背景，现在加上<BOI>

<EOS>

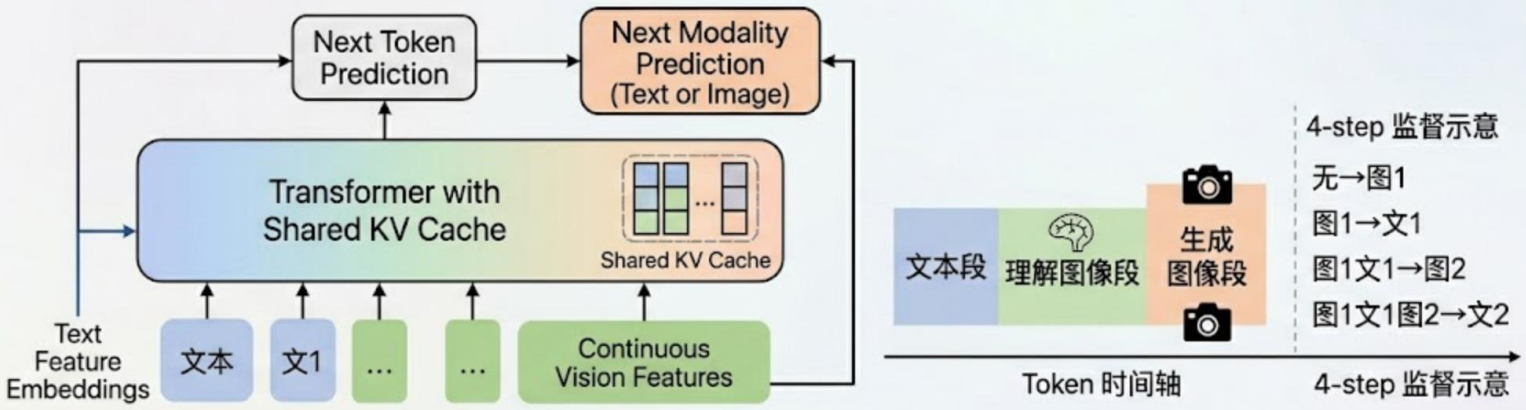
多模态交错上下文

多模态交错生成

生成、理解、编辑统一上下文与并行训练

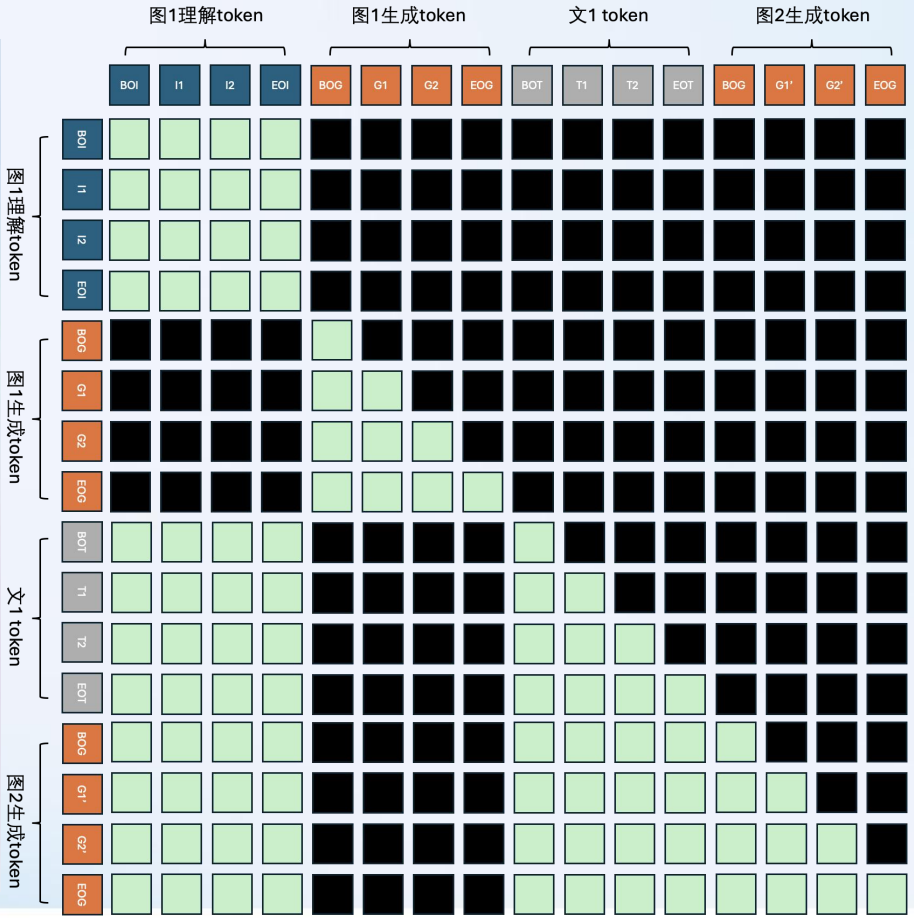


OneCat2：统一交错序列的自回归决策架构



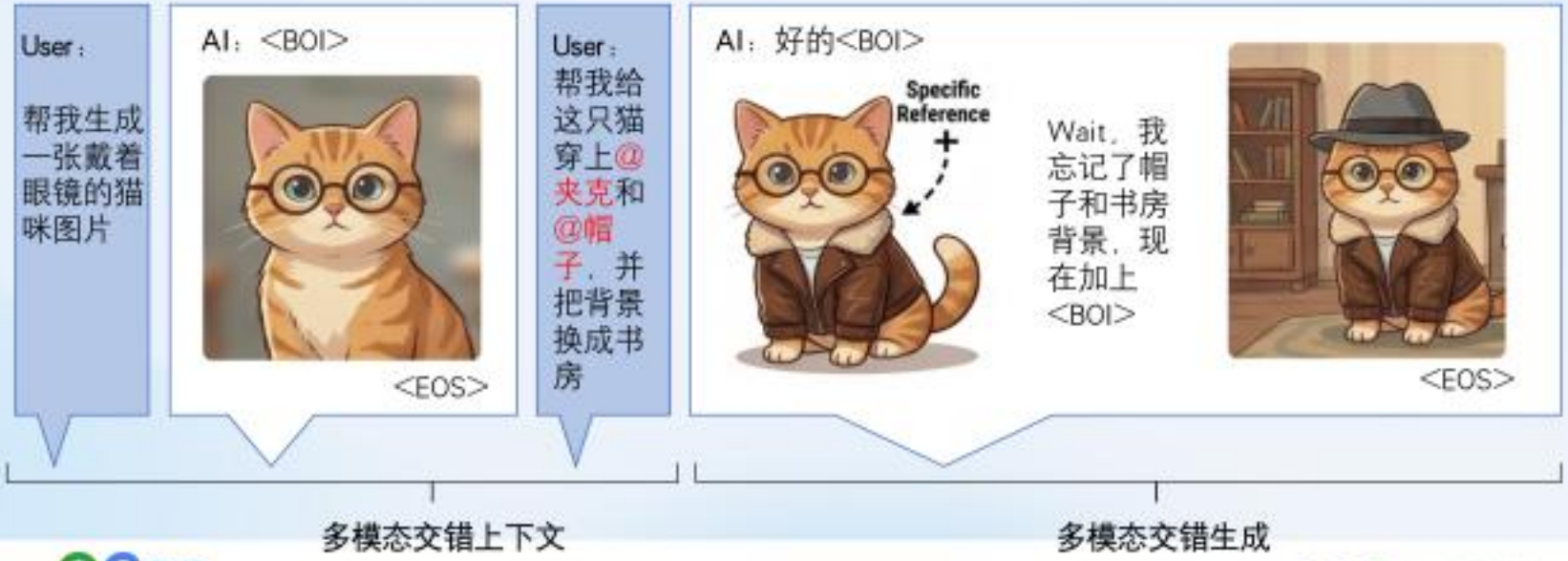
- 模态决策：模型不仅预测下一个 token，还预测下一个生成模态，从而决定下一步是回答还是出图。
- 迁移潜力：跨模态正向迁移未必能直接体现在单指标上，但统一训练至少为共享 world model 提供了条件。

Shared KV cache + Next modality prediction



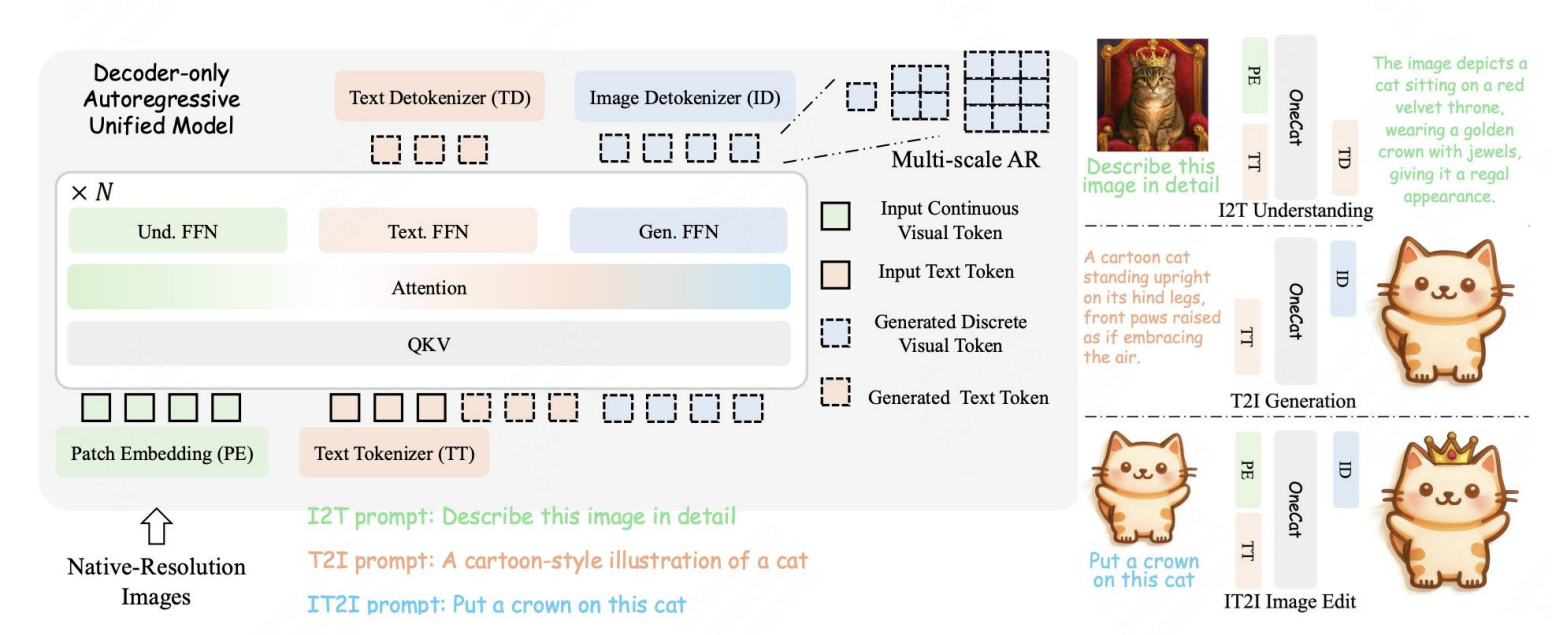
为什么交错数据理解/生成是统一模型的关键能力

- **多模输出**：故事/海报创作时，模型可以先规划文案，再生成首图，再根据反馈继续补文字或补第二张图。
- **多轮编辑**：参考图编辑时，模型先理解输入图，再直接生成修改结果，并在下一轮继续接收局部参考图迭代。
- **增量生成**：复杂场景可以先定大结构，再逐步补细节，而不是一次性完成。
- **推理需求**：复杂视觉/空间感知任务中，模型可以生成多模态COT辅助推理，例如视角变换，辅助线等。



式

生成理解联合的多阶段预训练范式



- 基座：LLM采用Qwen2.5 3B的文本基座模型，引入额外Expert、Head进行多模态训练，总参9B，激活3B

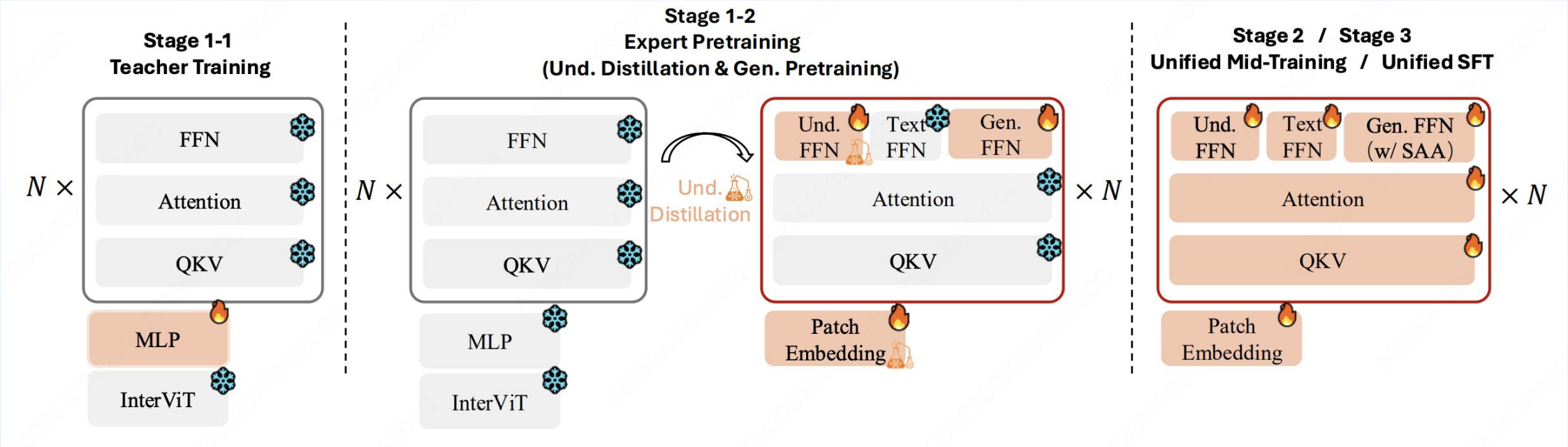
- 训练策略：三阶段训练
  1.冻结LLM、针对理解expert、生成expert进行预热训练
  2.网络全开、多任务联合训练
  3.进一步多任务联合高质量supervised fine-tuning

- 阶段一、图像生成、理解独立预训练：**让文本模型拥有基本的视觉感知和生成能力**

生成端：基于50M文生图数据预训练，让LLM能够从token by token的NTP生成范式拓展到多尺度自回归生成

理解端：基于436M图文描述数据进行蒸馏，让LLM文本模型快速获取基础的视觉感知能力
如何设计可用于蒸馏的教师teacher模型？ 如何设计蒸馏loss？

生成理解联合的多阶段预训练范式



如何设计蒸馏teacher？

Qwen2.5VL < Qwen2.5+InternViT+MLP

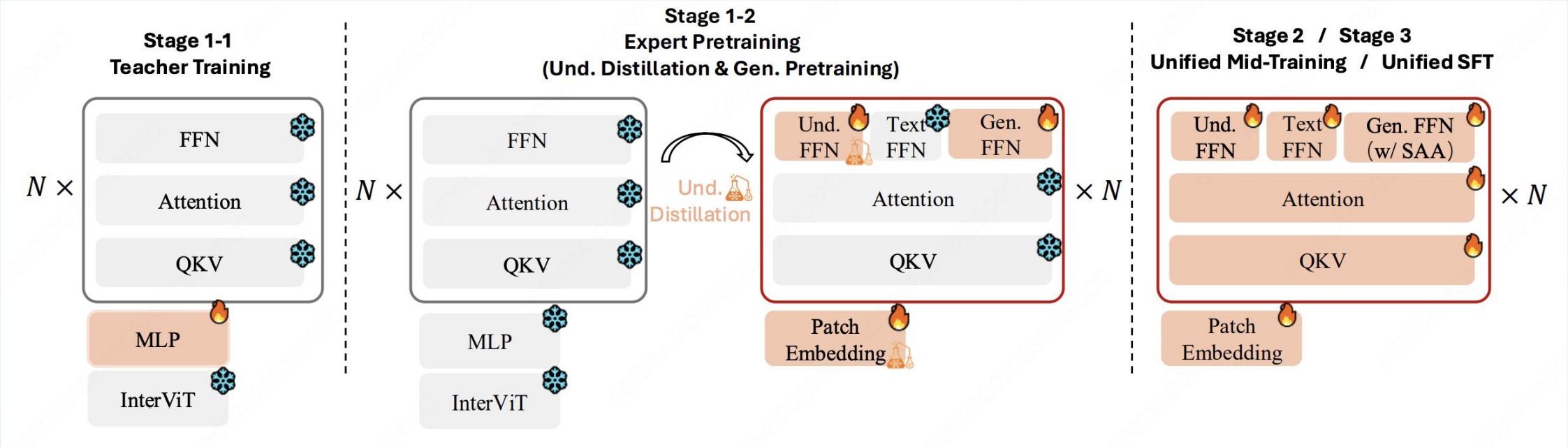
如何设计蒸馏loss？

特征蒸馏 > logits蒸馏

最后一层图文token 蒸馏loss> 仅文本token 蒸馏loss

所有层蒸馏loss >最后一层蒸馏loss

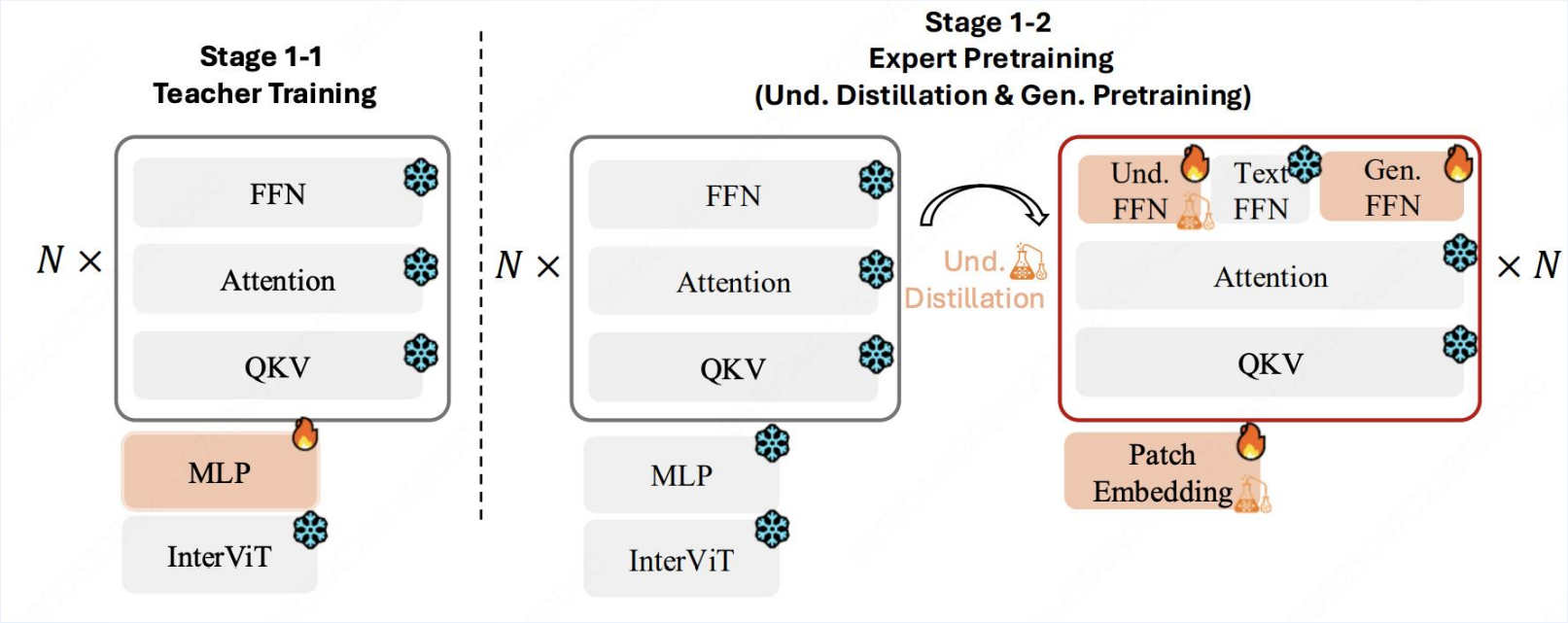
生成理解联合的多阶段预训练范式



- 阶段二、模型参数全开进行图像生成、编辑、理解、纯文本多任务联合训练
- 阶段三、在约总共14M高质量图像理解、生成、编辑数据下进行高质量SFT

生成理解联合的多阶段预训练范式

➢ 稀疏MoE架构下训练全流程



➢ **阶段一、专家预训练**

理解专家训练，生成专家单独训练

➢ **阶段一、专家预训练**

激活理解和生成专家，只打开细粒度专家进行理解端特征重建和生成端像素重建

➢ 阶段二、CPT
限制最大分辨率，2k以下长度，图文caption拓展到各类图文

➢ 阶段三、Mid Train，
8k数据，推理，视频，生成理解交错，编辑数据，再接32k数据

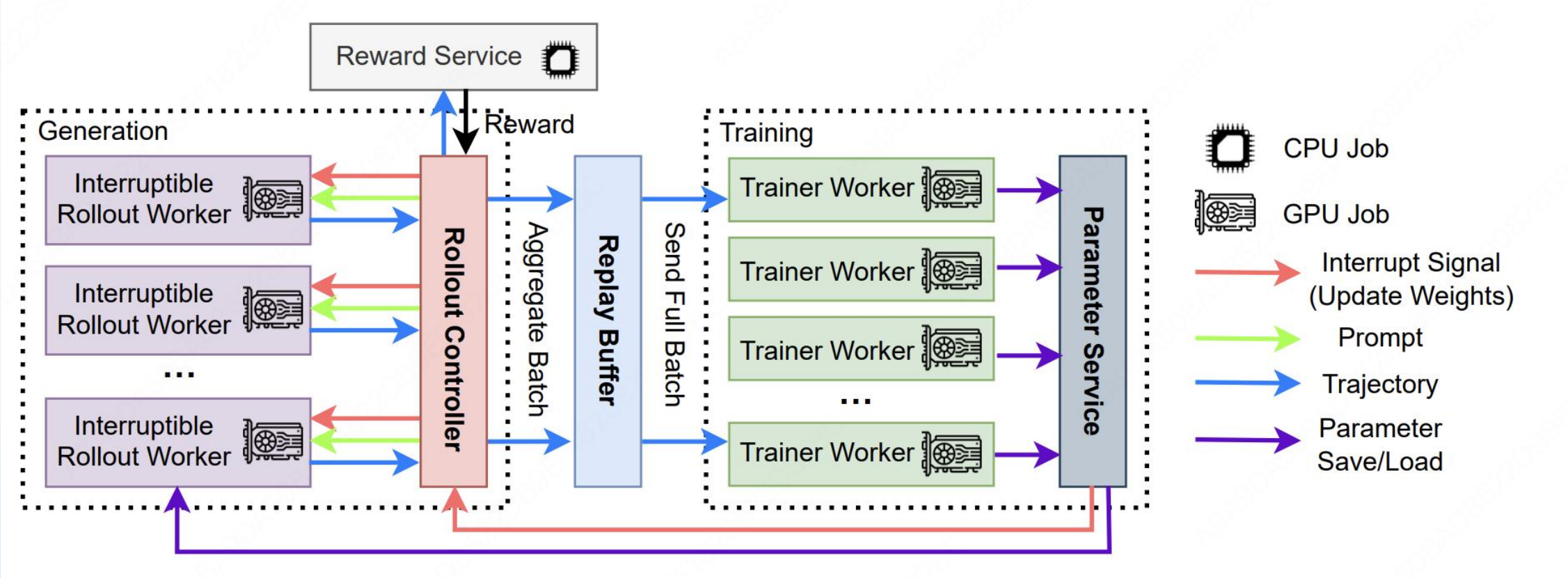
➢ 阶段四、后训练

轻量化SFT+统一RL

任意分辨率+参数全打开训练

一体化RL后训练范式

框架选择：Areal框架 Sglang部署加FSDP2训练

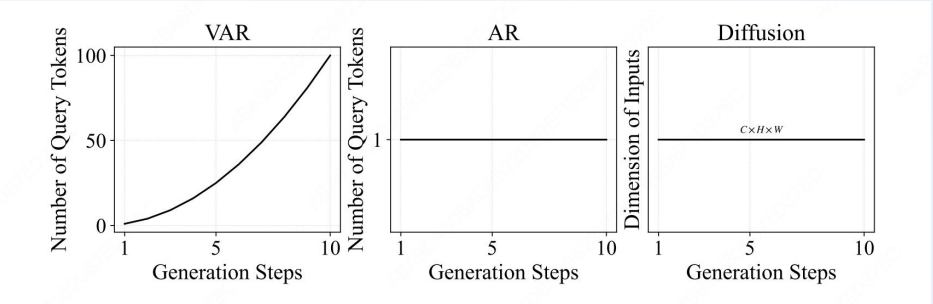


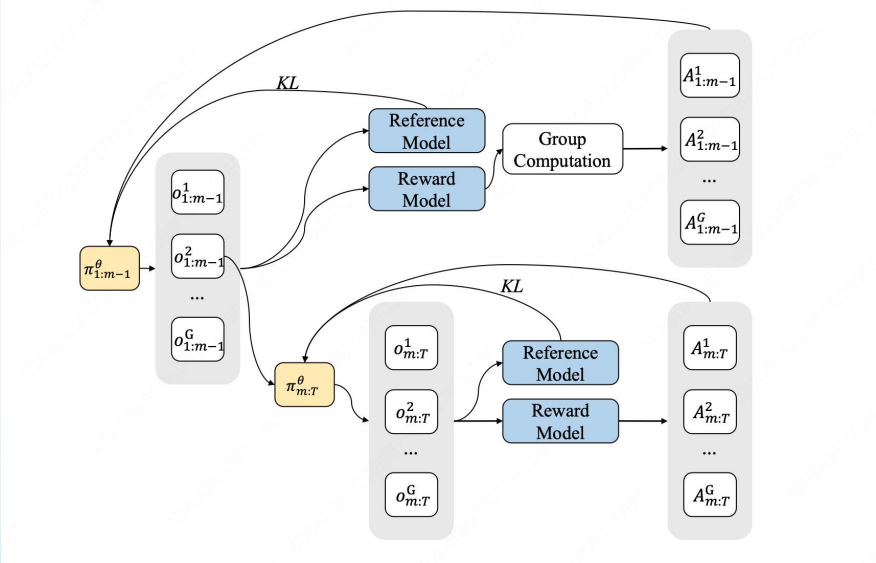
一体化RL后训练范式

VAR Rollout异步

- 同 Prompt Batch 化：同 prompt 的 16 个样本共享 input_ids/mask，仅 seed 不同，无需 padding，B=16 一次 forward。
- VMR 分段训练：在第 6 个 scale 处切分，用 K=2 次 suffix rollout 估计中间价值作为 prefix reward，3:1 交替 GRPO，缓解跨 scale 梯度冲突。
- 异步 Rollout/Train：Rollout 组和 Train 组各占一半 GPU，通过共享内存通信，rollout 和 train 时间重叠，1.6x 加速。
- Prefix/Suffix 异步拆分：Prefix 组做 prefix rollout+train，Suffix 组用保存的 KV cache 做 suffix rollout+train，尤其在高分辨率生成收益更高

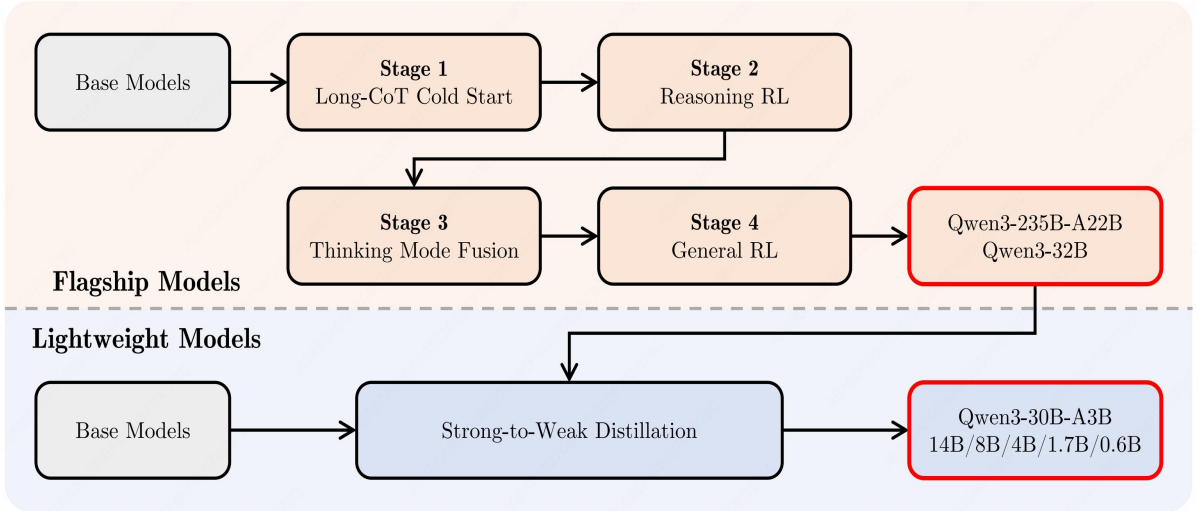
RL训练稳定性





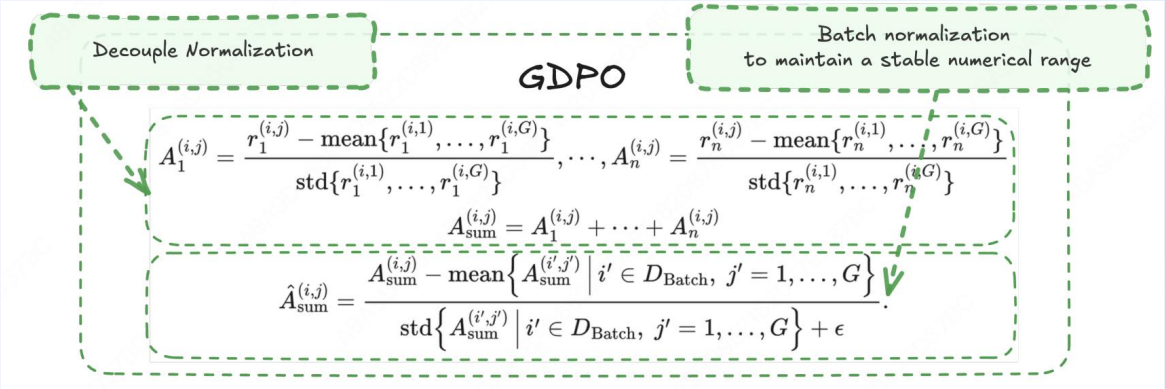
一体化RL后训练范式

理解端 OPD



- Qwen3-235B-A22B作为Teacher
- Dense且平稳的KL约束
- 生成OPSD+理解OPD

生成端 DPPO+GDPO



$$A_1^{(i,j)} = \frac{r_1^{(i,j)} - \text{mean}\{r_1^{(i,1)}, \ldots, r_1^{(i,G)}\}}{\text{std}\{r_1^{(i,1)}, \ldots, r_1^{(i,G)}\}}, \cdots, A_n^{(i,j)} = \frac{r_n^{(i,j)} - \text{mean}\{r_n^{(i,1)}, \ldots, r_n^{(i,G)}\}}{\text{std}\{r_n^{(i,1)}, \ldots, r_n^{(i,G)}\}}$$

$$A_{\text{sum}}^{(i,j)} = A_1^{(i,j)} + \cdots + A_n^{(i,j)}$$

$$\hat{A}_{\text{sum}}^{(i,j)} = \frac{A_{\text{sum}}^{(i,j)} - \text{mean}\left\{A_{\text{sum}}^{(i',j')} \,\middle|\, i' \in D_{\text{Batch}},\ j' = 1, \ldots, G\right\}}{\text{std}\left\{A_{\text{sum}}^{(i',j')} \,\middle|\, i' \in D_{\text{Batch}},\ j' = 1, \ldots, G\right\} + \epsilon}.$$

$$L_\mu(\pi) = \mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{|y|} M_t \cdot \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} \cdot \hat{A}_t\right],$$

$$M_t^{\text{PPO}} = \begin{cases} 0, & \text{if } \hat{A}_t > 0 \text{ and } \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} > 1 + \epsilon_{\text{high}}, \\ 0, & \text{if } \hat{A}_t < 0 \text{ and } \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} < 1 - \epsilon_{\text{low}}, \\ 1, & \text{otherwise}, \end{cases}$$

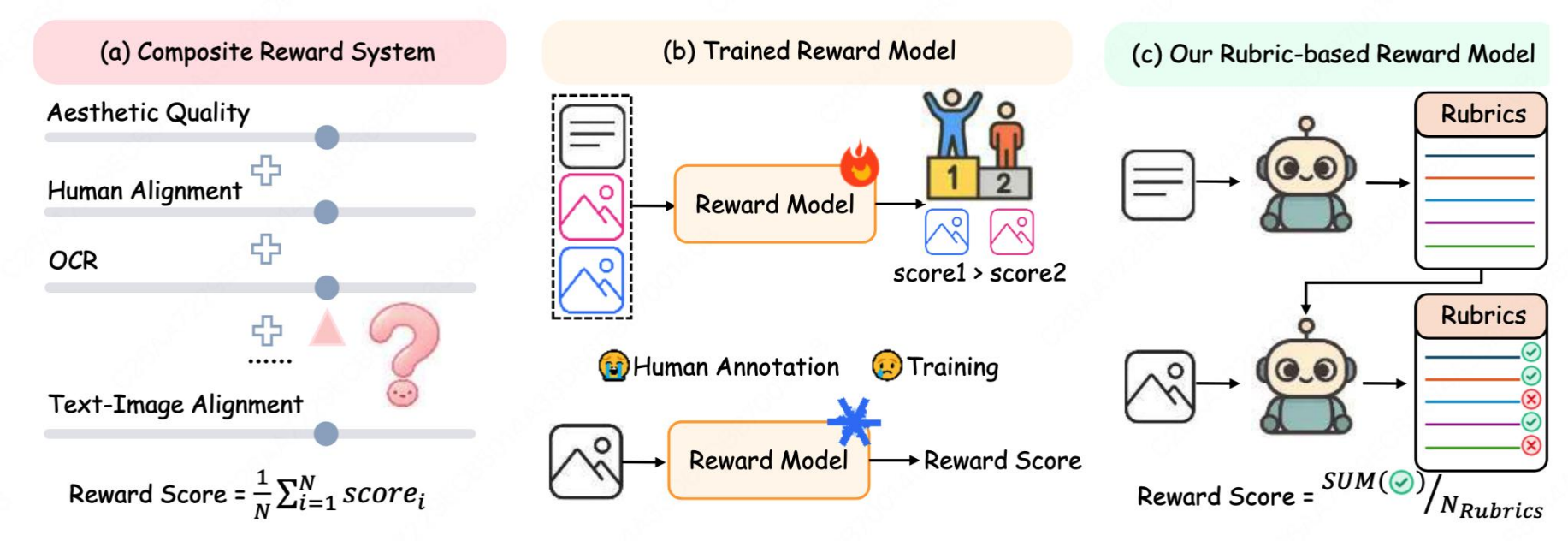
$$M_t^{\text{DPPO}} = \begin{cases} 0, & \text{if } \hat{A}_t > 0 \text{ and } \pi(y_t|s_t) - \mu(y_t|s_t) > \delta, \\ 0, & \text{if } \hat{A}_t < 0 \text{ and } \mu(y_t|s_t) - \pi(y_t|s_t) > \delta, \\ 1, & \text{otherwise}. \end{cases}$$

- GDPO：OCR，美学，遵循，人体姿态等多种reward结合
- DPPO：为生僻字生成这种小场景提供更高的clip range

生成Reward Modeling

- 传统生图 reward model 将所有偏好压缩为单一标量 reward，缺乏可解释性和 scaling 能力
- Rubric Reward Model: 利用 VLM 的预训练优势，跨维度评估生成质量。结合 thinking 能力和 rubric，降低 reward hacking，让奖励具备 prompt 自适应的分配，更具可解释性

- 结合 Boolean 和 Likert reward，在指令遵循、视觉质量、编辑一致性维度进行评估
- 通过 SFT 和 Multi-Agent RL 实现 rubric-generation 和 rubric-based-judge 联合优化



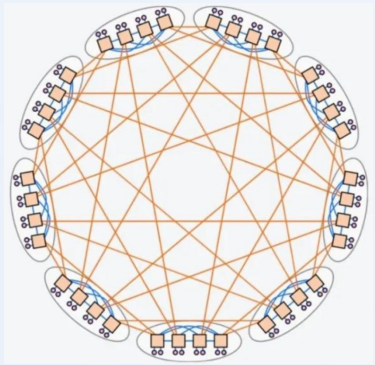
RubricRL：Simple Generalizable Rewards for Text—to—Image Generation

05

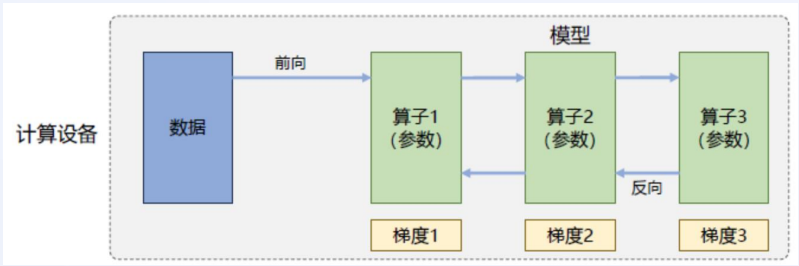
工程实践与典型挑战

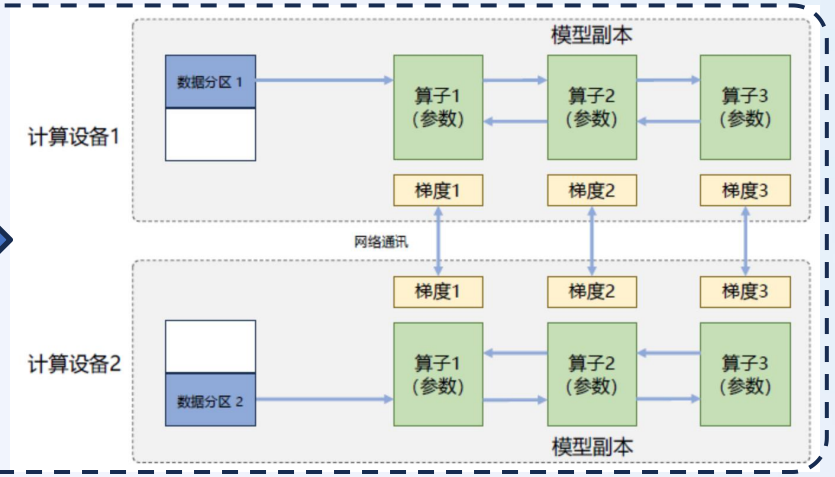
训练能效优化

- 大模型的**数据规模**决定了其训练过程需要在**计算集群上分布式执行**



大规模计算集群

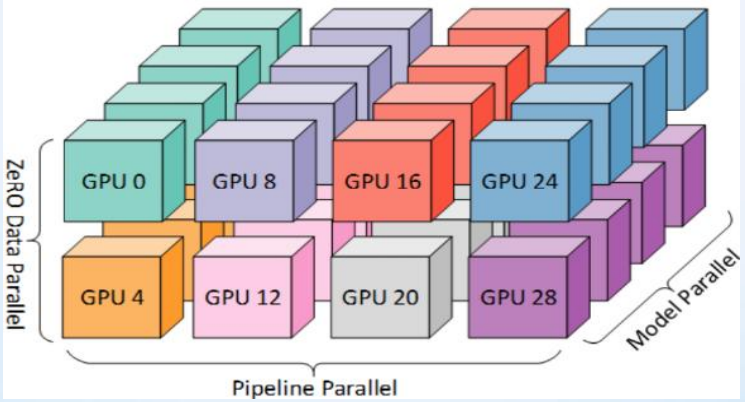


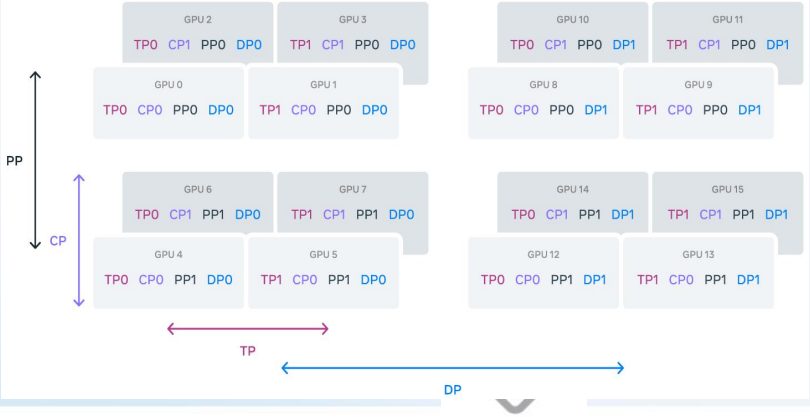


分布式训练并行策略（以数据并行为例）

- 支撑更大模型的关键：构建高效、稳健的**混合并行模式**

Megatron的
3D混合并行



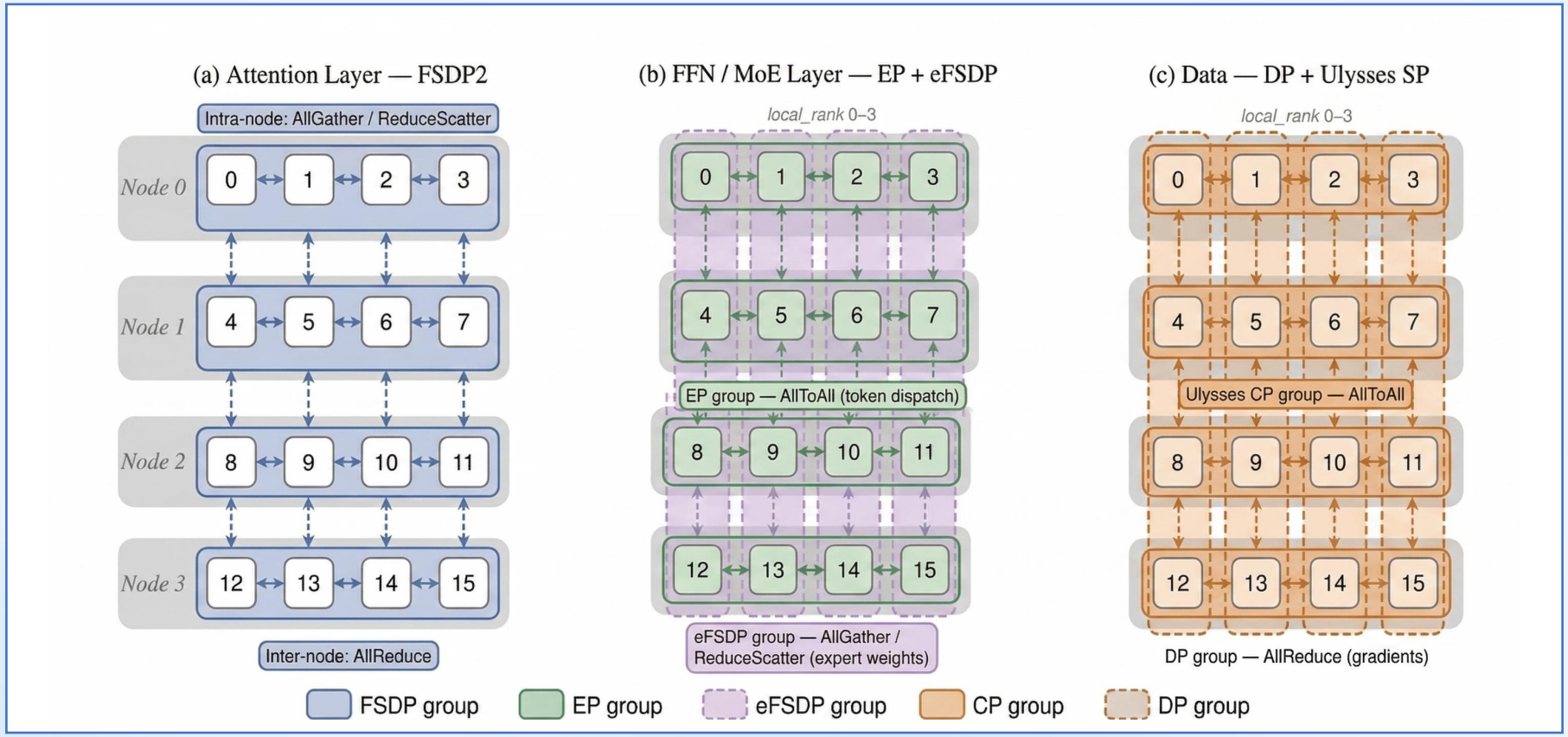


LLaMA3的
4D混合并行

训练能效优化

基于底层硬件物理拓扑特征，构建三级正交DeviceMesh 支撑高效的**混合并行策略**

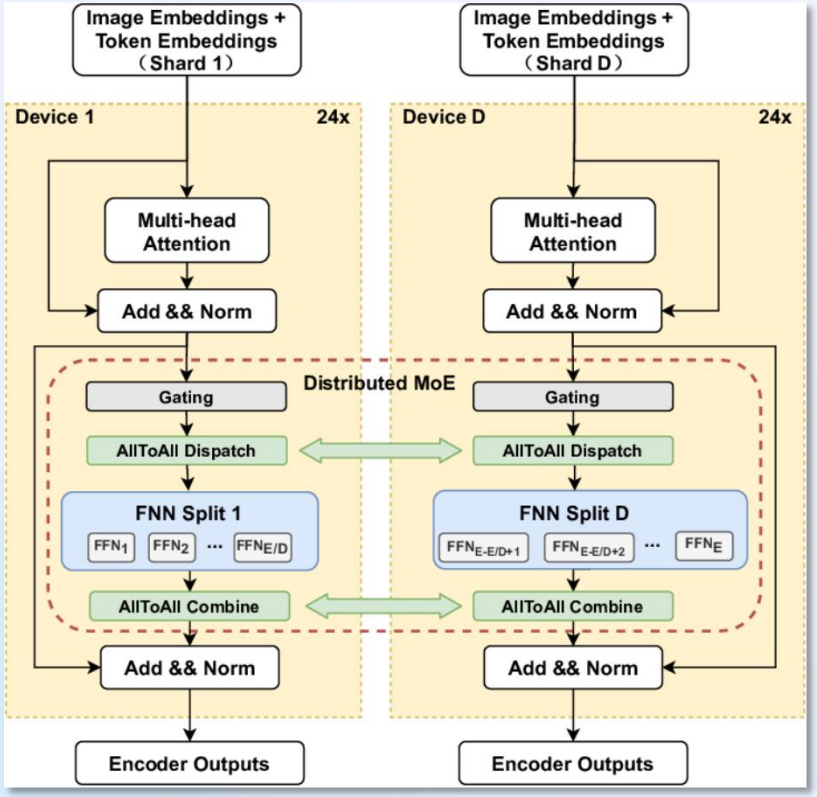
- **Attention层（FSDP2）**：节点内 AllGather/ReduceScatter 权重，节点间 AllReduce 同步梯度
- **MoE层（EP+FSDP）**：节点内 AllToAll路由token，节点间AllGather/ReduceScatter专家权重
- **数据层（DP+Ulysses SP）**：节点内 AllToAll做张量转置，节点间AllReduce 同步梯度



MoE的极致协同

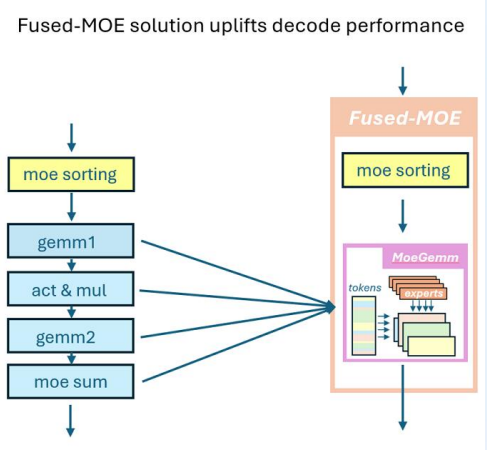
Token 经 Gate 路由后，通过两次 All-to-All 完成跨卡分发与回收，中间执行本地 Expert GEMM：

All-to-All (Dispatch) → Expert GEMM → All-to-All (Combine)



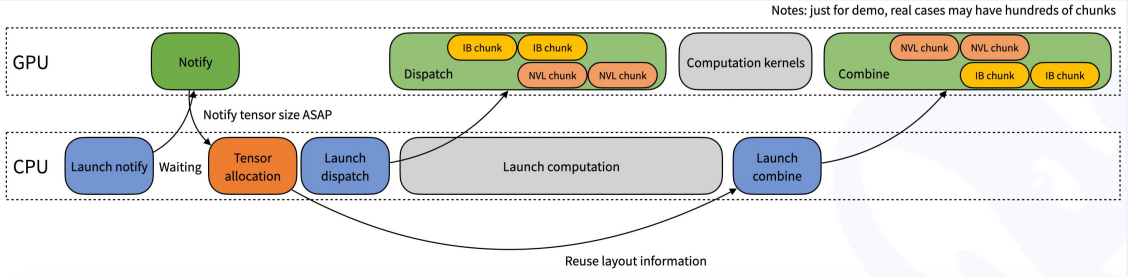
➢ 专家计算优化：Fused-MoE

将所有 expert 的gemm计算+SwiGLU激活融合为单一fused kernel，消除逐 expert 循环带来的 kernel launch overhead与显存读写往返。



➢ 通信优化：DeepEP

针对EP场景提供了高度优化的 All-to-All 通信原语，在节点内（利用 NVLink）和节点间（利用 RDMA）均能实现极高的带宽利用率。

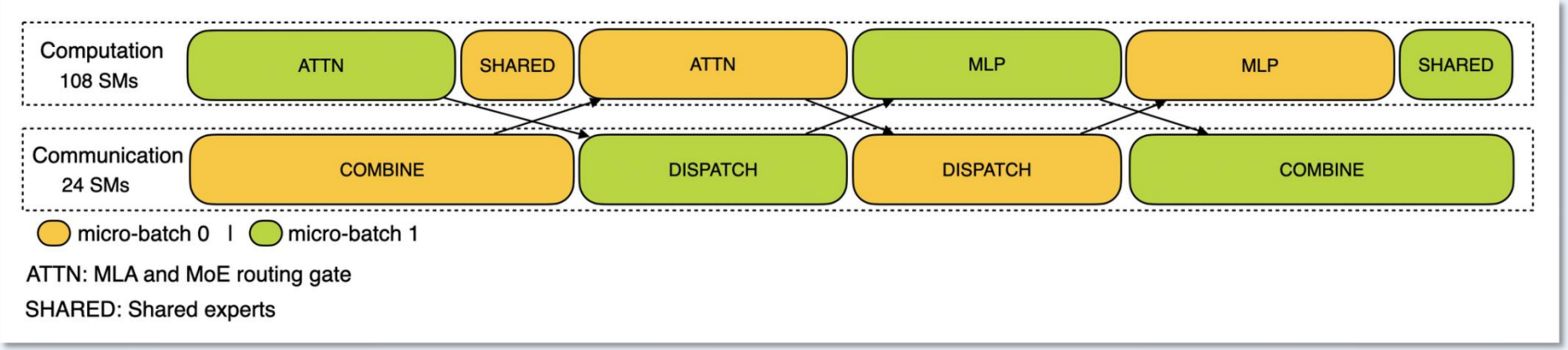


MoE的极致协同

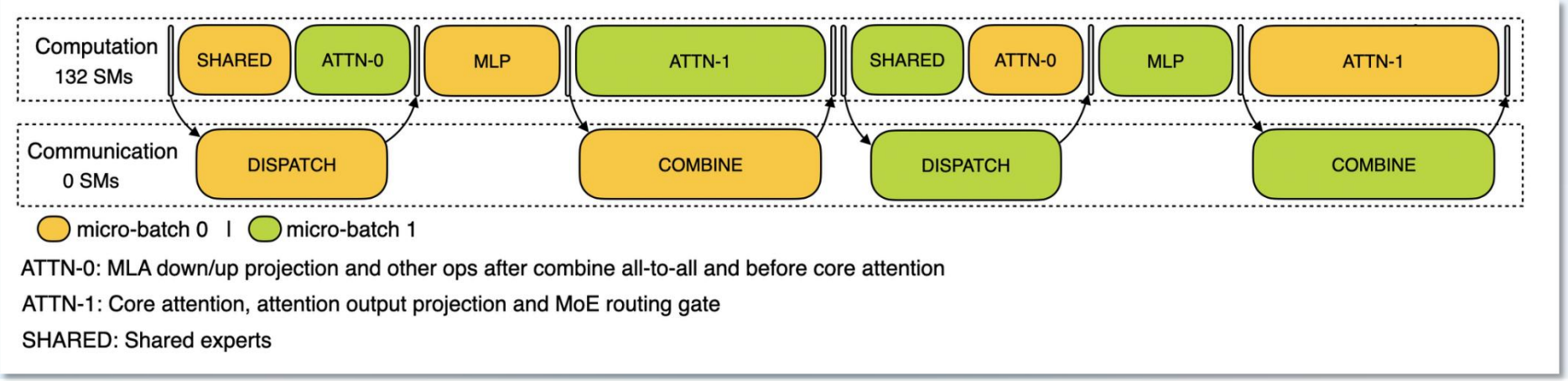
➢ 通信计算并行重叠策略

引入两个 micro-batch 交替执行来实现**通信-计算重叠**：在 MoE 层中，一个 micro-batch 进行 dispatch/combine 通信时，另一个 micro-batch 的 Expert GEMM 计算同步在独立的 CUDA Stream 上执行，将通信延迟隐藏在计算时间内。

prefill

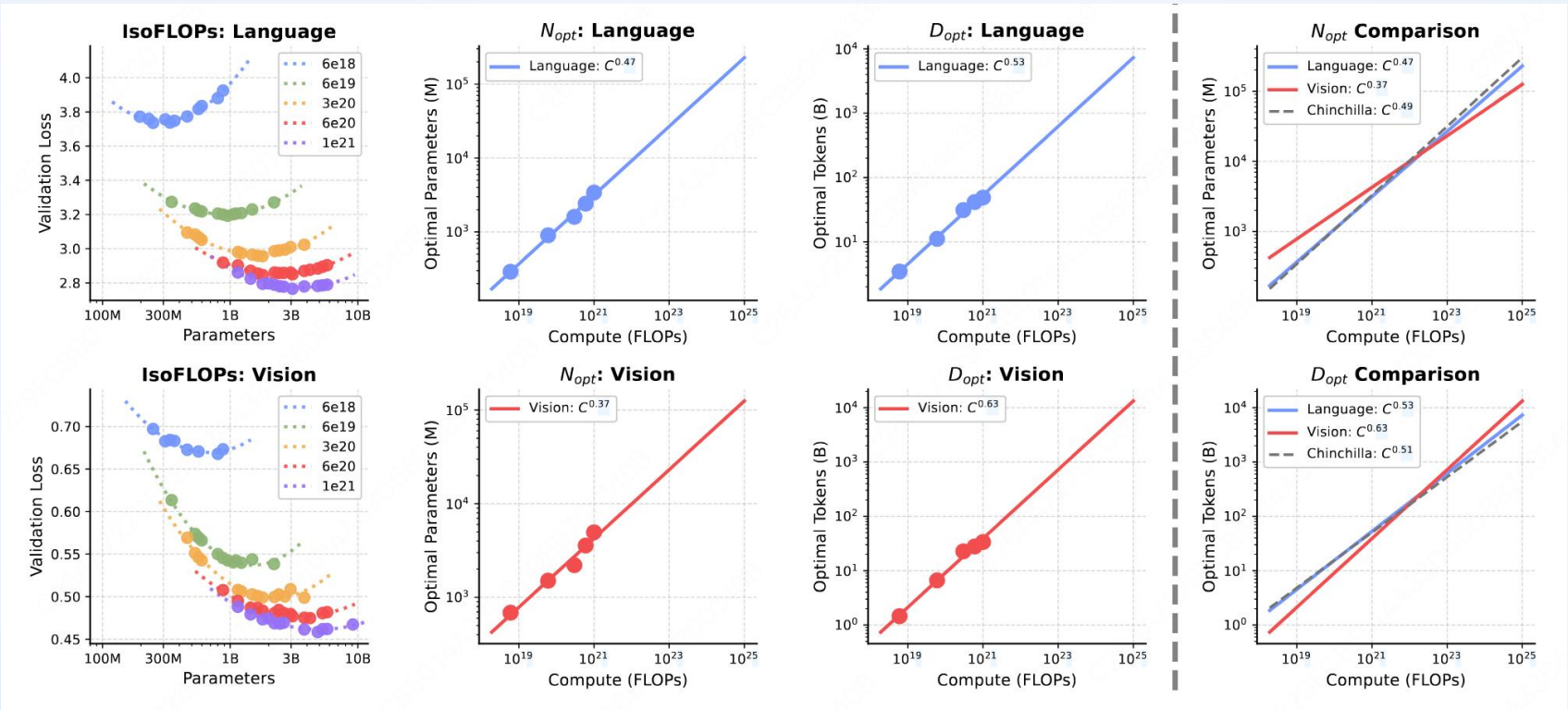


decode



数据科学与配比

统一模态训练中，生成与理解遵循不同的 scaling 趋势，视觉生成任务相比语言任务对参数量的要求相对低，对数据量的需求更高。



2026 Meta Beyond Language Modeling：An Exploration of Multimodal Pretraining

数据科学与配比

OneCat 三阶段训练数据配比策略：

- Stage1 - alignment: 视觉理解:视觉生成 = 8:1
- Stage2 - unified mid-training: 纯文本:视觉理解:视觉生成 = 1:2:6
- Stage3 - SFT: 纯文本:视觉理解:视觉生成 = 1:5:6

Token Ratio	Understanding								Generation	
(T:U:G)	MMB	MME-S	MMVet	SEED	AI2D	ChartQA	TextVQA	**Avg.**	GenEval	DPG
5B : 10B : 0B	61.6	1556	28.9	67.2	60.1	56.1	56.2	55.1	-	-
5B : 10B : 10B	62.7	1547	29.7	66.7	60.3	59.8	55.4	55.6	80.7	74.6
5B : 10B : 30B	62.0	1602	29.4	66.7	59.8	61.6	55.5	56.0	81.2	74.9
5B : 10B : 45B	62.1	1544	30.1	66.8	58.4	58.6	55.0	55.2	81.4	75.6

增加生成数据不会损害理解能力，能大幅提升生成能力

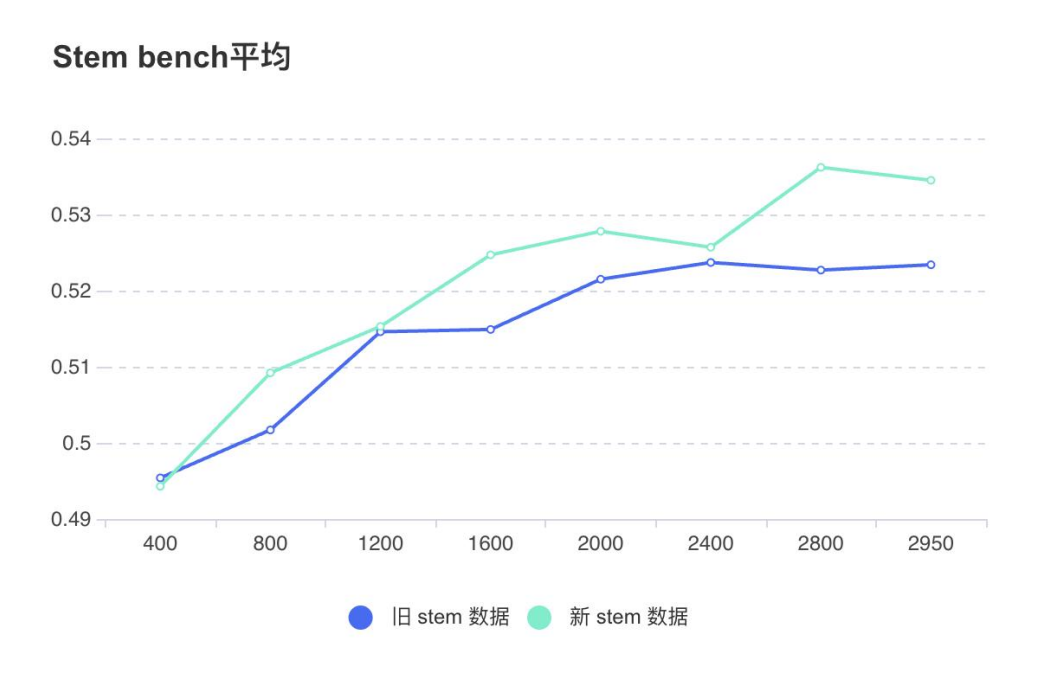
Stage	Task Type	Data Sources
Stage-1	Multimodal Understanding	Recap-DataComp-1B (Li et al., 2024a), Capsfusion (Yu et al., 2024), Detailed-Caption (Li et al., 2024b), SA1B-Dense-Caption (Data, 2024), Moondream2-COYO-5M-Captions (isidentical, 2024), COYO700M (Byeon et al., 2022), CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018), LAION-400M (Schuhmann et al., 2021), and Zeor250M (Xie et al., 2023).
	Visual Generation	ImageNet-1k (Deng et al., 2009), COYO700M (Byeon et al., 2022), LAION-400M (Schuhmann et al., 2021), CC12M (Changpinyo et al., 2021), and additional synthetic images generated by FLUX.
Stage-2	Multimodal Understanding & Text-only Instruction	Detailed-Caption (Li et al., 2024b), ALLaVA (Chen et al., 2024a), ShareGPT4V (Chen et al., 2024b), SA1B-Dense-Caption (Data, 2024), WIT (Srinivasan et al., 2021), pdfa-eng-wds (pdf), UReader (Ye et al., 2023), DVQA (Mathew et al., 2021b), OCR-VQA (Mishra et al., 2019), WebSRC (Chen et al., 2021), GQA (Hudson & Manning, 2019b), visual-genome (Krishna et al., 2017), GRIT (Peng et al., 2023), and additional in-house visual and text-only instruction samples.
	Visual Generation	Visual generation data of Stage-1, AnyEdit (Yu et al., 2025), UltraEdit (Zhao et al., 2024b), HQ-Edit (Hui et al., 2024), and OmniEdit (Wei et al., 2024).
Stage-3	Multimodal Understanding & Text-only Instruction	MAmmoTH-VL (Guo et al., 2024), AI2D (Kembhavi et al., 2016b), OKVQA (Marino et al., 2019), VQAv2 (Goyal et al., 2017), ART500K (Mao et al., 2017), ScienceQA (Saikh et al., 2022), GQA (Hudson & Manning, 2019b), CLEVR-Math (Lindström & Abraham, 2022), COCO-ReM (Singh et al., 2024), TallyQA (Acharya et al., 2019), Docmatix (Laurençon et al., 2024), DVQA (Mathew et al., 2021b), DreamSim (Fu et al., 2023), and ShareGPT4o (Chen et al., 2024d).
	Visual Generation	UniWorld (Lin et al., 2025b), BLIP3o-60k (Chen et al., 2025a), ShareGPT-4o-Image (Chen et al., 2025b), and additional synthetic data generated by GPT-4o and FLUX using the partial prompts from JourneyDB (Sun et al., 2023).

数据科学与配比

Mid-Training 阶段进行模型整体训练和学习率退火，对模型能力起决定性作用。 重点在 Mid-Training 阶段进行数据质量和模型能力提升。

视觉理解能力优化

- 视觉跨 domain 配比：将数据分为 通用VQA, STEM, OCR, Grounding, Agent, Video 等多个 domain，进行综合配比。基于浓度需求进行合理上采。
- 推理能力前置优化：在 mid-train 阶段引入充分浓度的 thinking 数据，提升推理长度和知识能力。
- 更广的数据规模：总训练数据扩量到 T 级别。



视觉生成能力优化

- 模型能力扩展：支持长 interleave 和多图交互
- 数据质量提升：基于质量打分的生图数据清洗优化

06

总结与未来展望

Decoder Only原生多模态大模型性能

Model	# Params LLM	# Params Vis. Enc.	MME-P↑	MME-S↑	MMBench↑	MMMU↑	MM-Vet↑	MathVista↑	SEED↑
Encoder-based Understanding Only Models									
InternVL2 (Chen et al., 2024c)	1.8B	0.3B	1440	1877	73.2	34.3	44.6	46.4	71.6
InternVL2.5 (Chen et al., 2024b)	1.8B	0.3B	-	2138	74.7	43.6	60.8	51.3	-
Qwen2-VL (Wang et al., 2024a)	1.5B	0.6B	-	1872	74.9	41.1	49.5	43.0	-
Qwen2.5-VL (Bai et al., 2025)	3B	0.6B	-	2157	79.1	53.1	61.8	62.3	-
Encoder-free Understanding Only Models									
Mono-InternVL (Luo et al., 2025)	1.8B	/	-	1875	65.5	33.7	40.1	45.7	67.4
EvE (Diao et al., 2024)	7B	/	-	1628	52.3	32.6	25.7	-	64.6
EvEv2.0 (Diao et al., 2025)	7B	/	-	1709	66.3	39.3	45.0	-	71.4
HoVLE (Tao et al., 2025)	2.6B	/	-	1864	71.9	33.7	44.3	46.2	70.7
Unified Models									
Chameleon (Team, 2024)	7B	-	-	-	35.7	28.4	8.3	-	30.6
Emu3 (Wang et al., 2024b)	8B	0.3B	-	-	58.5	31.6	37.2	-	68.2
Harmon (Wu et al., 2025c)	1.5B	0.9B	1155	1476	65.5	38.9	-	-	67.1
Show-o2 (Xie et al., 2025)	1.5B	0.1B	1450	-	67.4	37.1	-	-	-
Janus-Pro (Chen et al., 2025d)	1.5B	0.3B	1444	-	75.5	36.3	39.8	-	-
ILLUME+ (Huang et al., 2025)	3B	-	1414	-	**80.8**	**44.3**	40.3	-	**73.3**
VILA-U (Wu et al., 2024)	7B	0.4B	1401	-	-	-	33.5	-	59.0
Janus-Pro (Chen et al., 2025d)	7B	0.3B	1567	-	79.2	41.0	50.0	-	-
Tar (Han et al., 2025)	7B	0.4B	1571	1926	74.4	39.0	-	-	-
Show-o2 (Xie et al., 2025)	7B	0.1B	1620	-	79.3	48.9	-	-	69.8
Ours	3B	/	**1630**	**2051**	78.8	41.9	**52.2**	**61.7**	72.5

Model	# Params A-LLM	# Params Vis. Enc.	TextVQA↑	ChartQA↑	InfoVQA↑	DocVQA↑	OCRBench↑	GQA↑	AI2D↑
Encoder-based Understanding Only Models									
InternVL2-2B (Chen et al., 2024c)	1.8B	0.3B	73.4	76.2	58.9	86.9	-	784	74.1
InternVL2.5-2B (Chen et al., 2024b)	1.8B	0.3B	74.3	79.2	60.9	88.7	-	804	74.9
Qwen2-VL-3B (Bai et al., 2025)	1.5B	0.6B	79.7	73.5	65.5	90.1	809	-	74.7
Qwen2.5-VL-3B (Bai et al., 2025)	3B	0.6B	79.3	84.0	77.1	93.9	797	-	81.6
Encoder-free Understanding Only Models									
Mono-InternVL (Luo et al., 2025)	1.8B	/	72.6	73.7	-	-	767	59.5	68.6
EvE (Diao et al., 2024)	7B	/	56.8	59.1	-	-	398	62.6	61.0
EvEv2 (Diao et al., 2025)	7B	/	71.1	73.9	-	-	702	62.9	74.8
HoVLE (Tao et al., 2025)	2.6B	/	70.9	78.6	55.7	86.1	740	64.9	73.0
Unified Models									
Chameleon (Team, 2024)	7B	?	4.8	2.9	5.0	1.5	7	-	46.0
Emu3 (Wang et al., 2024b)	8B	0.3B	64.7	68.6	43.8	76.3	687	60.3	70.0
Harmon-1.5B (Wu et al., 2025c)	1.5B	0.9B	-	-	-	-	-	58.9	-
Show-o2-1.5B (Xie et al., 2025)	1.5B	0.1B	-	-	-	-	-	60.0	69.0
Janus-Pro-1.5B (Chen et al., 2025d)	1.5B	0.3B	-	-	-	-	-	59.3	-
ILLUME+ (Huang et al., 2025)	3B	0.6B+?B	-	69.9	44.1	80.8	672	-	74.2
VILA-U (Wu et al., 2024)	7B	0.4B	60.8	-	-	-	-	60.8	-
Janus-Pro-7B (Chen et al., 2025d)	7B	0.3B	-	-	-	-	-	62.0	-
Tar-7B (Han et al., 2025)	7B	0.4B	-	-	-	-	-	61.3	-
Show-o2-7B (Xie et al., 2025)	7B	0.1B	-	-	-	-	-	63.1	**78.6**
Ours	3B	/	**73.9**	**81.2**	**64.8**	**91.2**	**768**	**63.1**	77.8

Table 5: Performance comparison on the GenEval [42] benchmark. The dagger (†) indicates methods that employ an LLM for prompt rewriting. Top-1 accuracy is reported (Best in **bold**, second best is underlined).

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Generation-only Models							
SDXL (Podell et al., 2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3 (Shi et al., 2020)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev† (Labs, 2024)	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified Models							
Chameleon-7B (Team, 2024)	-	-	-	-	-	-	0.39
Transfusion-7B (Zhou et al., 2024)	-	-	-	-	-	-	0.63
Emu3-8B† (Wang et al., 2024b)	0.99	0.81	0.42	0.80	0.49	0.45	0.66
ILLUME+ 3B (Huang et al., 2025)	0.99	0.88	0.62	0.84	0.42	0.53	0.72
Harmon-1.5B (Wu et al., 2025c)	0.99	0.86	0.66	0.85	0.74	0.48	0.76
Show-o2-7B (Xie et al., 2025)	1.00	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro-7B (Chen et al., 2025d)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Mogao-7B† (Liao et al., 2025)	**1.00**	**0.97**	0.83	0.93	**0.84**	**0.80**	0.89
BLIP3-o-8B† (Chen et al., 2025a)	-	-	-	-	-	-	0.84
Tar-7B (Han et al., 2025)	0.99	0.92	0.83	0.85	0.80	0.65	0.84
UniWorld-V1-20B (Lin et al., 2025a)	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1-20B† (Lin et al., 2025a)	0.98	0.93	0.81	0.89	0.74	0.71	0.84
BAGEL-7B (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL-7B† (Deng et al., 2025)	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
OneCat-3B	0.99	0.94	0.83	0.93	**0.84**	0.79	**0.89**

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
Editing-only Models										
MagicBrush [87]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [5]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [32]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [89]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
Step1X-Edit [47]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
ICEdit [88]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Unified Models										
OmniGen [75]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
OmniGen2 [72]	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	**3.44**
BAGEL-7B [16]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1-20B [44]	**3.82**	3.64	2.27	3.47	**3.24**	2.99	4.21	**2.96**	2.74	3.26
OneCAT-3B	3.65	**3.70**	**2.42**	**3.92**	3.00	**3.79**	4.61	2.23	3.53	3.43

Table 7 Comprehensive comparison on ImgEdit-Bench [80] showing performance across nine editing categories. Higher scores are better for all metrics. Best in **bold**, second best is underlined.

- 文生图Benchmark上，超越现有统一模型，展现出较强的指令遵循能力。

Decoder Only原生多模态大模型性能

➢ 多模态生成和理解任务推理效率得到显著提升

Model	Resolution of Input Image	#Tokens Visual	TTFT(s)
Qwen2.5-VL-3B	768 × 768	731	0.135
OneCAT-3B	768 × 768	731 +256*	0.067 (50%↓)
Qwen2.5-VL-3B	1792 × 1792	4098	0.583
OneCAT-3B	1792 × 1792	4098 +256*	0.225 (61%↓)

Model	Resolution of Output Image	T2I Infer. Time (s)	Edit Infer. Time (s)
BAGEL-7B	512 × 512	8.76	13.45
OneCAT-3B	512 × 512	1.40 (84%↓)	2.03 (84%↓)
BAGEL-7B	1024 × 1024	26.29	46.44
OneCAT-3B	1024 × 1024	2.85 (89%↓)	4.61 (90%↓)

Table 8 Efficiency comparison for understanding (**Left**) and generation (**Right**), tested on one NVIDIA H800. **Left:** We report the time to first token (TTFT). The number of input text tokens are fixed to 24. 256* is the number of visual tokens of thumbnail. **Right:** We report total inference time for Text-to-Image (T2I) and Image-Editing.

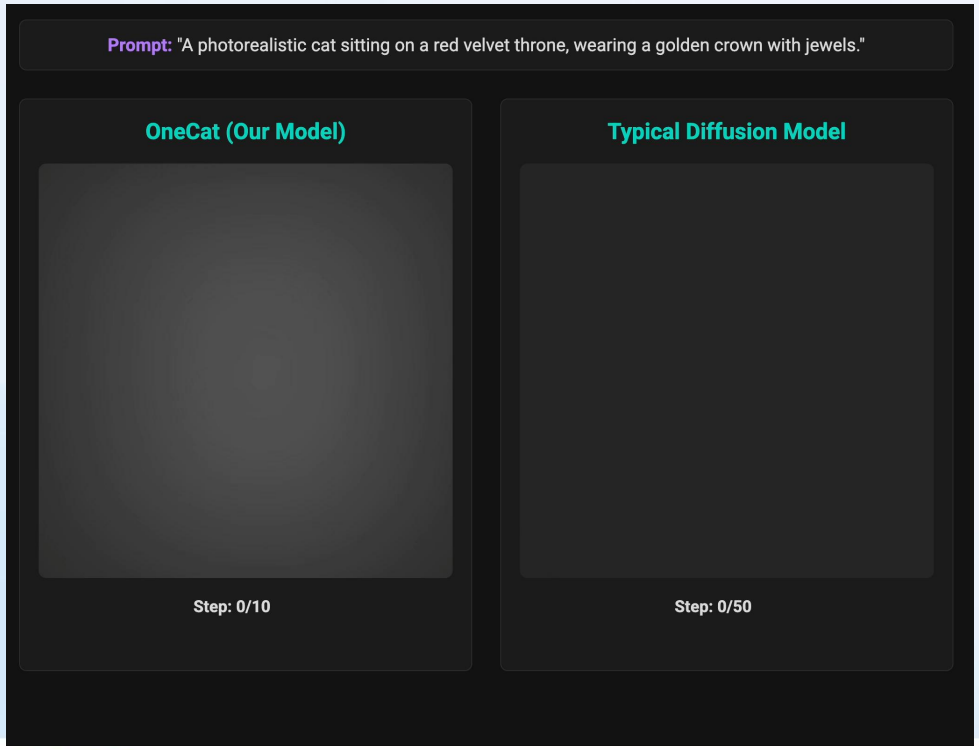
Decoder Only原生多模态大模型性能

支持不同分辨率、不同比例、不同风格的图像生成

Decoder Only原生多模态大模型性能

超大规模模型统一推理加速

- 理解单次推理提速 2 倍以上，高分辨率输入提升越多
- 生成提速8-10倍，高分辨率编辑提速超过20倍



强世界认知的生成效果-世界模型的关键

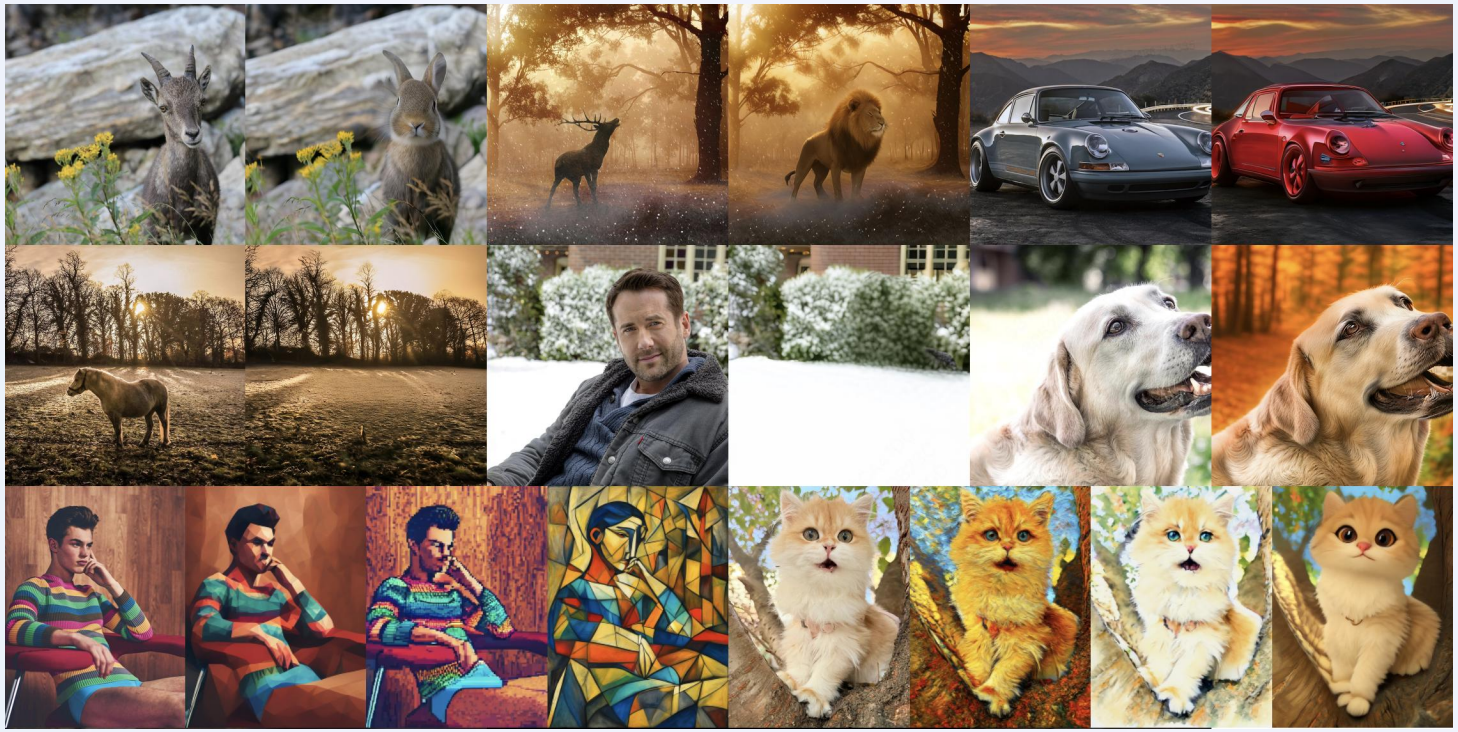
- 语义理解一致性上达到OpenAI模型相当水平

QCon
全球软件开发大会

Decoder Only原生多模态大模型性能

图像编辑

风格迁移



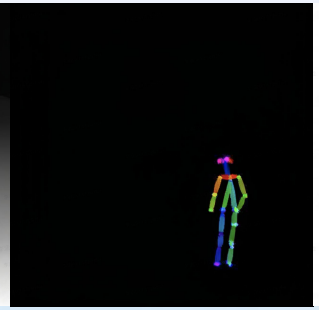
基于编辑的
图像感知任务

原图

Img-> Canny

Img-> Depth

Img-> Pose

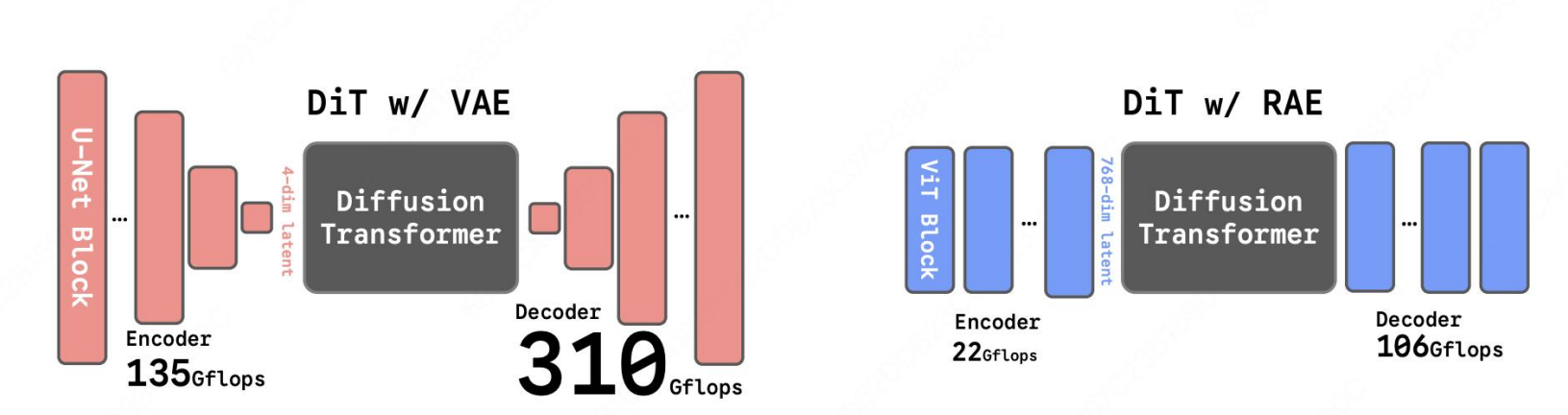


Img-> Sketech

Img-> Seg

理解生成一体化未来发展

2025.10 NYU *RAE: Diffusion Transformers with Representation Autoencoders*



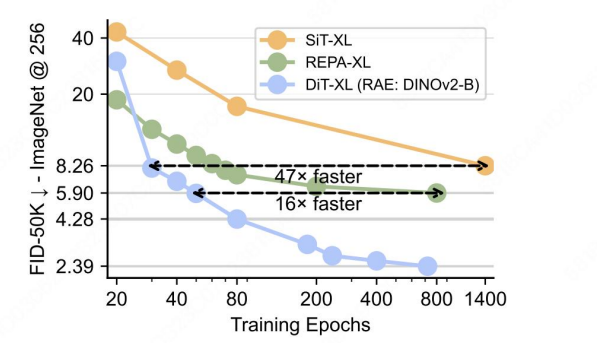
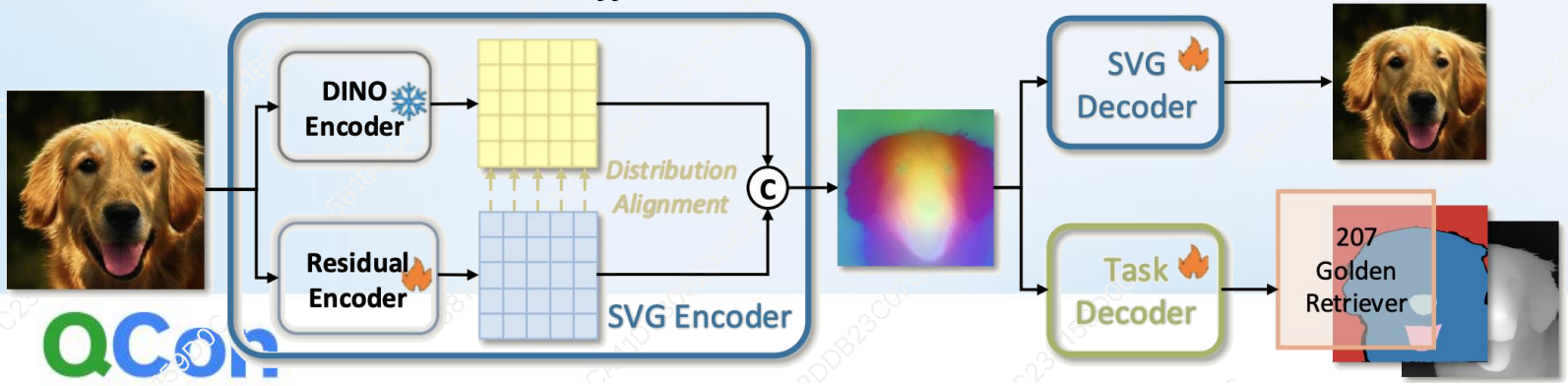


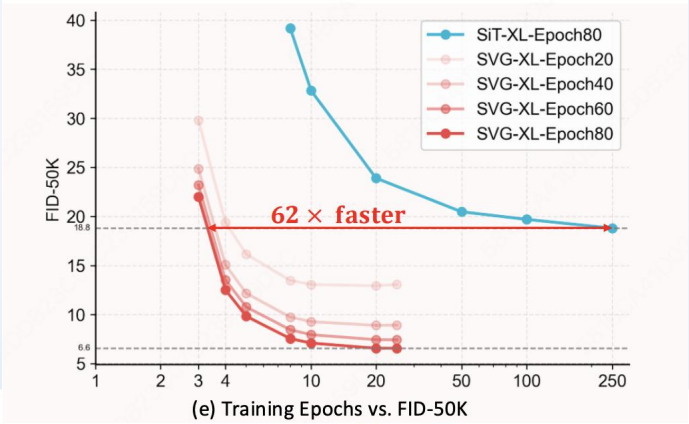
Figure 4: **DiT w/ RAE reaches much faster convergence and better FID than SiT or REPA.**

U-Net VAE has high computational cost
Noise-to-pixel latent mapping via U-Net is hard to learn

VAE encoder → pretrained ViT (CLIP / DINOv3 / SigLIP)
VAE decoder → learnable Decoder

2025.10 北大&快手 *Latent Diffusion Model without Variational Autoencoder*

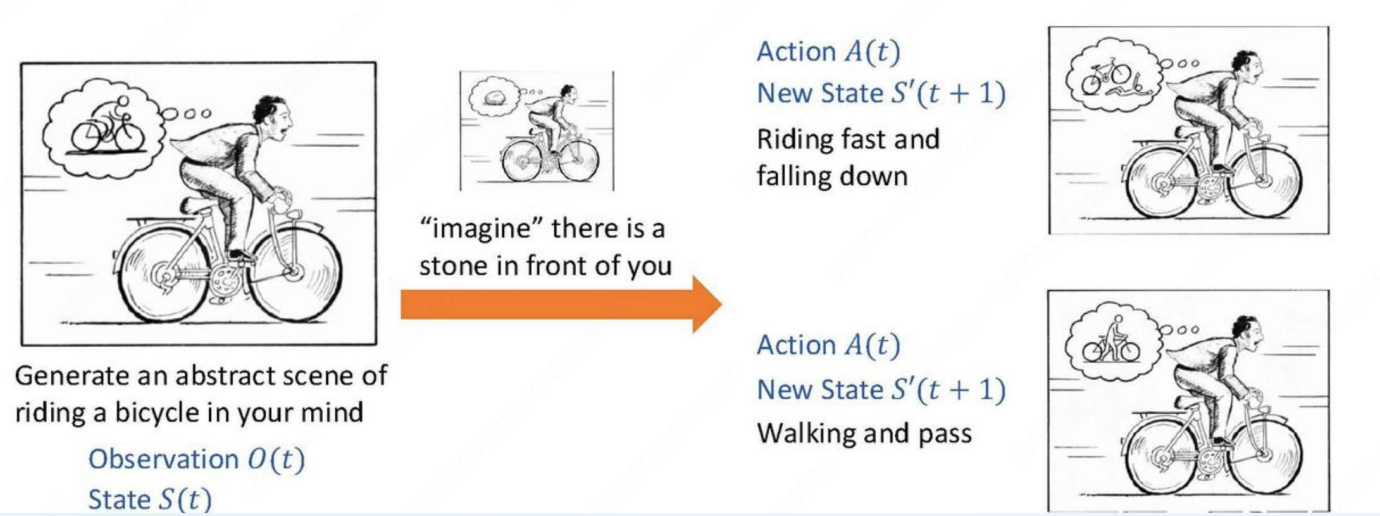




QCon 全球软件开发大会
InfoQ 极客传媒

理解生成一体化未来发展

1.图文交错生成推理 ->理解生成统一的RL Post-Training -> 理解生成在thinking阶段相互促进->**以视觉为中心的多模态推理/VLA**

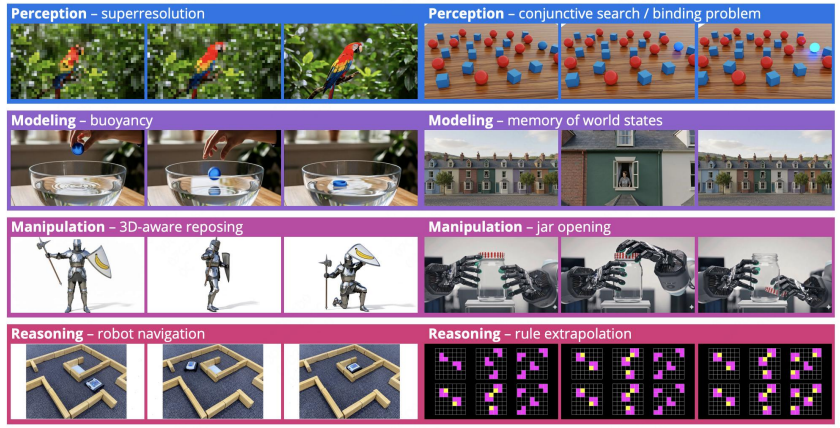


图像+文本输入–> 文本思维链 –> 文本/动作输出

图像+文本输入–> 文本+图像思维链 –> 文本/图像/动作输出

2.视频理解生成编辑一体化 ：

- 理解促进生成：依赖视频理解能力辅助**长时间一致性、物理规律一致性**的视频生成
- 生成促进理解：视频生成本身就能产生智能，能否作为因果推理范式，提供理解能力(veo3)
- 最终收敛为： 可交互世界模型/具身智能VLA模型



3.未来的多模态预训练新范式？

文本自回归建模本质上是学习token的相关性，并不是真正的因果相关的世界知识

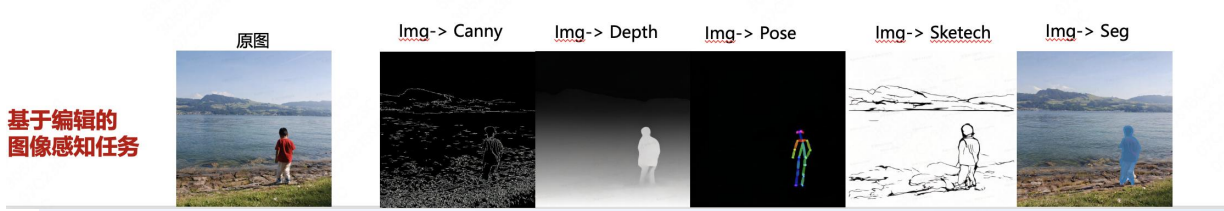
如何更加高效利用**海量视频数据**，帮助MLLM更原生地学习/压缩世界知识，构建世界模型

理解生成一体化未来发展

1.图文交错生成推理 ->理解生成统一的RL Post-Training -> 理解生成在thinking阶段相互促进->**以视觉为中心的多模态推理/VLA**

Table 2: **Evaluation results on RealUnify. WR**: World Knowledge; **CR**: Commonsense Reasoning; **MR-I**: Mathematical Reasoning; **LR**: Logical Reasoning; **SR**: Scientific Reasoning; **C2T**: Code-to-Image; **MR-II**: Mental Reconstruction; **MT**: Mental Tracking; **AF**: Attentional Focusing; **CN**: Cognitive Navigation. For each task, we present both direct and stepwise evaluation results, reported in the format direct/step. The best performance on each task is in blue.

Model	Understanding Enhances Generation							Generation Enhances Understanding					Total
	WK	CR	MR-I	LR	SR	C2I	Avg	MR-II	MT	AF	CN	Avg	
Proprietary Models													
Nano Banana	89 / -	86 / -	34 / -	65 / -	48 / -	56 / -	63.0 / -	34 / -	27 / -	36 / -	30 / -	31.8 / -	50.5 / -
Open-Source Unified Models													
MIO	24 / 35	26 / 33	18 / 13	9 / 10	10 / 11	0 / 8	14.5 / 18.3	26 / 23	19 / 18	35 / 19	23 / 21	25.8 / 20.3	19.0 / 19.1
Janus-Pro	25 / 26	77 / 71	16 / 7	13 / 17	16 / 20	3 / 10	25.0 / 25.2	21 / -	23 / -	28 / -	29 / -	25.3 / -	25.1 / -
ILLUME+	44 / 52	62 / 62	22 / 22	23 / 25	26 / 26	1 / 7	29.7 / 32.3	27 / 27	19 / 20	35 / 38	30 / 25	27.8 / 27.5	28.9 / 30.4
Show-o2	30 / 42	56 / 50	25 / 25	21 / 21	18 / 20	18 / 19	28.0 / 29.5	36 / -	28 / -	36 / -	21 / -	30.3 / -	28.9 / -
OmniGen2	36 / 55	61 / 60	21 / 26	29 / 28	16 / 20	19 / 6	30.3 / 32.5	30 / 42	21 / 24	51 / 38	28 / 19	32.5 / 30.8	31.2 / 31.8
UniPic2	61 / 62	73 / 72	31 / 30	28 / 38	25 / 26	7 / 15	37.5 / 40.5	26 / 28	20 / 24	27 / 27	23 / 16	24.0 / 23.8	32.1 / 33.8
UniWorld-V1	51 / 56	64 / 59	26 / 26	33 / 37	21 / 24	15 / 9	35.0 / 35.2	29 / 33	19 / 25	57 / 36	24 / 20	32.3 / 28.5	33.9 / 32.5
Ovis-U1	37 / 59	72 / 71	28 / 30	23 / 34	15 / 17	12 / 25	31.2 / 39.3	32 / 38	28 / 25	60 / 31	36 / 24	39.0 / 29.5	34.3 / 35.4
BLIP3-o	57 / 62	71 / 74	21 / 24	19 / 25	28 / 22	2 / 9	33.0 / 36.0	36 / -	25 / -	57 / -	32 / -	37.5 / -	34.8 / -
OneCAT	61 / 64	70 / 65	32 / 20	29 / 27	24 / 31	9 / 27	37.5 / 39.0	26 / 29	25 / 26	43 / 26	31 / 36	31.3 / 29.3	35.0 / 35.1
BAGEL	46 / 74	70 / 80	23 / 26	29 / 37	21 / 29	7 / 40	32.7 / 47.7	37 / 38	31 / 25	50 / 52	39 / 28	39.3 / 35.8	35.3 / 42.9



快手可灵团队 2025.10

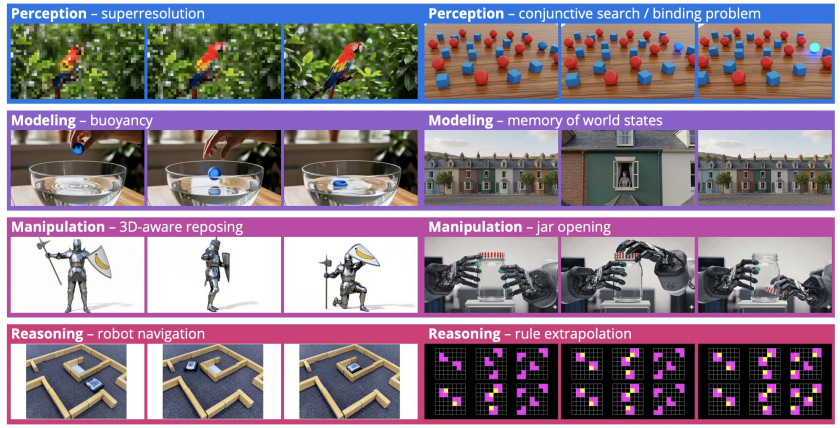
RealUnify：Do Unified Models Truly Benefit from Unification? A Comprehensive Benchmark

第三方评测显示OneCAT在开源unified模型中排名第二，仅次于字节BAGEL

在理解生成相互促进上具有较大潜力

2.视频理解生成编辑一体化：

- 理解促进生成：依赖视频理解能力辅助**长时间一致性、物理规律一致性**的视频生成
- 生成促进理解：视频生成本身就能产生智能，能否作为因果推理范式，提供理解能力(veo3)
- 最终收敛为： 可交互世界模型/具身智能VLA模型



3.未来的多模态预训练新范式？

文本自回归建模本质上是学习token的相关性，并不是真正的因果相关的世界知识

如何更加高效利用**海量视频数据**，帮助MLLM更原生地学习/压缩世界知识，构建世界模型

QCon
全球软件开发大会

InfoQ 极客传媒

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

AiCon 上海站

全球人工智能开发与应用大会

探索 AI 应用边界

Explore the limits of AI applications

2026 年 6 月 26–27 日 / 上海张江科学会堂

参会咨询

专题：世界模型与多模态智能突破

专题出品人：朱政

极佳视界 / 联合创始人 & 首席科学家

专题：Agent 系统架构与工程化实践

专题出品人：鲁浩楠

OPPO / 大模型算法负责人

专题：AI 原生数据工程

专题出品人：王彦辉

火山引擎 / 数智平台产品总监

专题：Agent 安全、评测与可信治理

专题出品人：马传雷（岳立）

支付宝 / 业务风险技术部负责人

专题：Agent 数据、记忆与运行时基础设施

专题出品人：邓亚峰

EverMind / CEO

专题：企业级研发体系的重构

专题出品人：沈浪

快手 / 研发效能负责人