rednote | hi lab

多模态生成与理解：数据与建模方法的演化

演讲者: 罗维俭，小红书人文智能(hi-lab)实验室研究员
电子邮箱: pkulwj1994@icloud.com; luoweijian@xiaohongshu.com;
个人主页: https://pkulwj1994.github.io/

注释：本slides部分内容由GPT5模型参与润色与讨论。讲稿内容仅为为个人学术观点，不构成任何商业建议。

什么是多模态模型：

- 多模态模型，指的是处理例如图像，文本，语音，遥感数据等多种不同特质而又相互关联的数据的模型；
- 人类的智能是「多模态的」：我们通过 **看图像 / 听声音 / 阅读语言 / 与环境交互** 来理解世界。
- 过去十年，AI 在 **视觉（CNN / ViT）** 和 **语言（Transformer / LLM）** 两个方向分别突飞猛进，但**仍然各自为战。**

图像-文本多模态理解与生成：

- 长期以来，图像和文本是两种最基本，也是最重要的模态；
- 按照传统的图像-文本之间的推理关系，很多研究将图像到文本的推断模型称为**多模态理解模型**，将文本到图像的推断模型称为**多模态生成模型。**

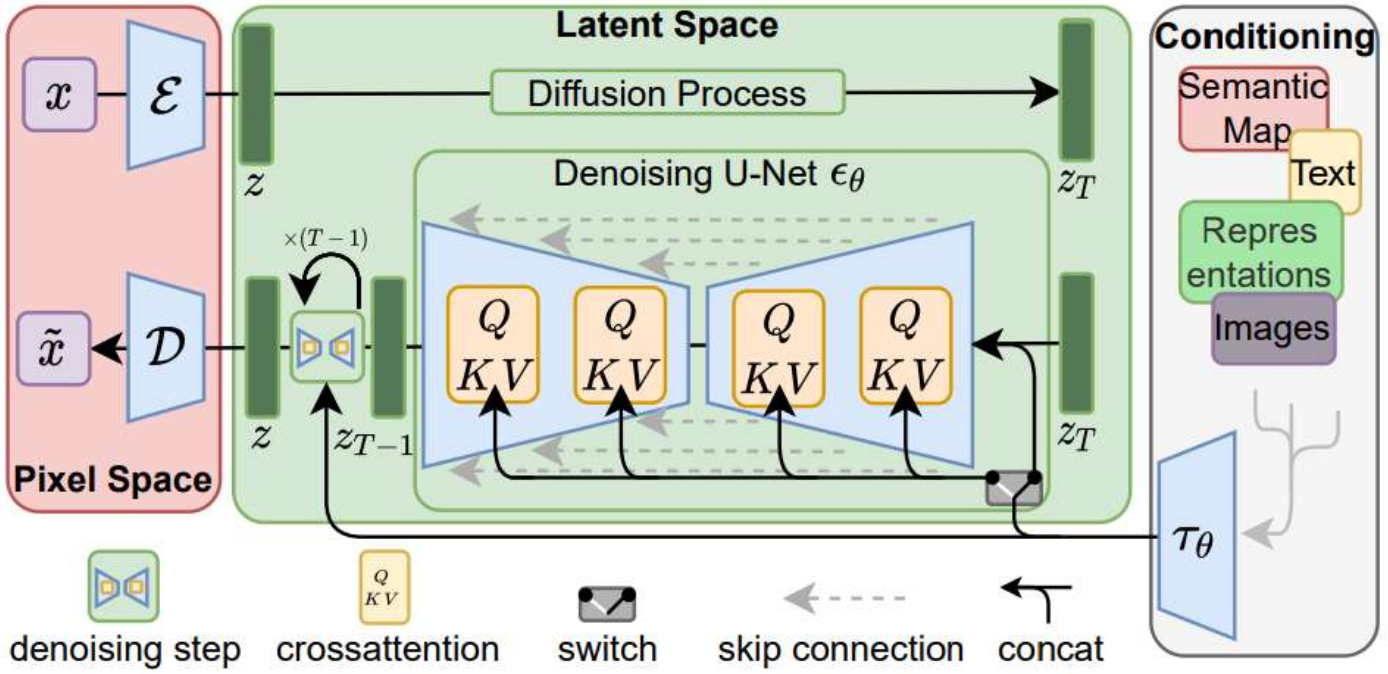
传统多模态任务和模型			
阵营	代表任务	典型模型	局限
多模态理解	VQA、Captioning、视觉推理	CLIP, BLIP, Flamingo	只能“读懂”，不会“创造”
多模态生成	Text-to-Image, Image Editing	DALL·E, Stable Diffusion	只会“画”，但不“理解语义”

Large-scale Transformer时代的多模态生成与理解：

- 以GPT为代表的大型transformer语言模型推动学术界和工业界回顾过去十年的 **数据构建方式** 与 **建模范式**，去尝试构建真正“**既能看懂又能创造**”的 AI模型？。

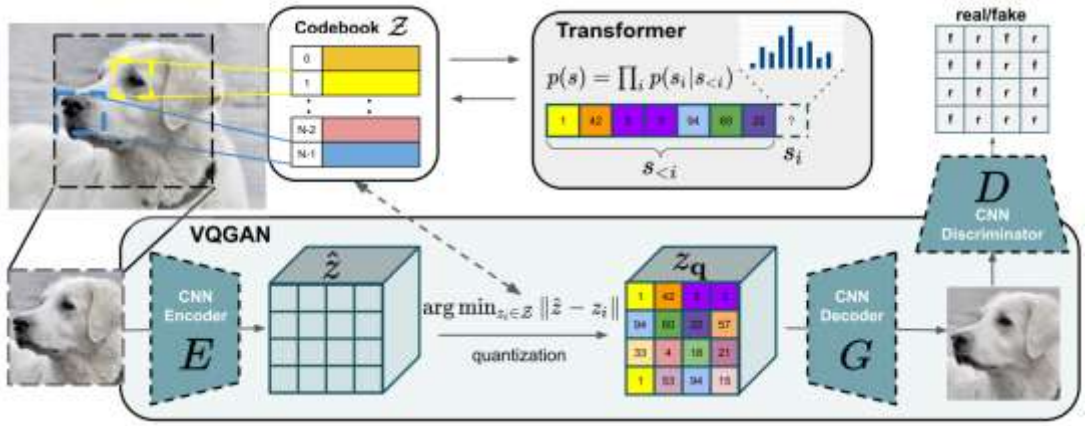
多模态生成的经典范式：条件扩散模型

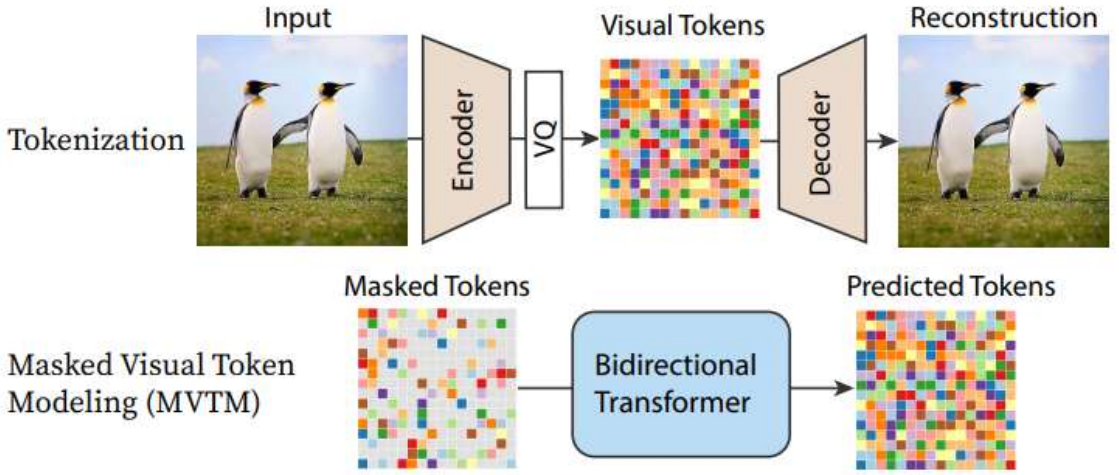
- 目前scale最好的多模态生成的经典范式是**条件扩散模型（Stable Diffusion）**：
  - 假设**x0**是一个图像数据（或者VAE编码后的图像隐变量）。**xt**是**x0**被前向扩散过程加噪声破坏的图像数据。T是扩散时刻，也可理解为噪声强度。扩散模型**d(xt, t)**是一个神经网络，他通过最小化Weighted Denoising Score Matching 的训练损失函数，可以学习到如何通过多次迭代将**xt**逐渐概率性降噪成某个干净的图像**x0**。
  - 假设 **c** 是一个文本编码，**d(xt, t, c)** 是一个条件扩散模型，那么通过使用**d(xT, t, c)** 不断迭代采样，即**x0** = Sample[**d(xt, t, c)**]可以根据输入的文本 **c** 和纯噪声 **xT** 获得生成的图像**x0**；



多模态生成的经典范式：离散图像Token生成

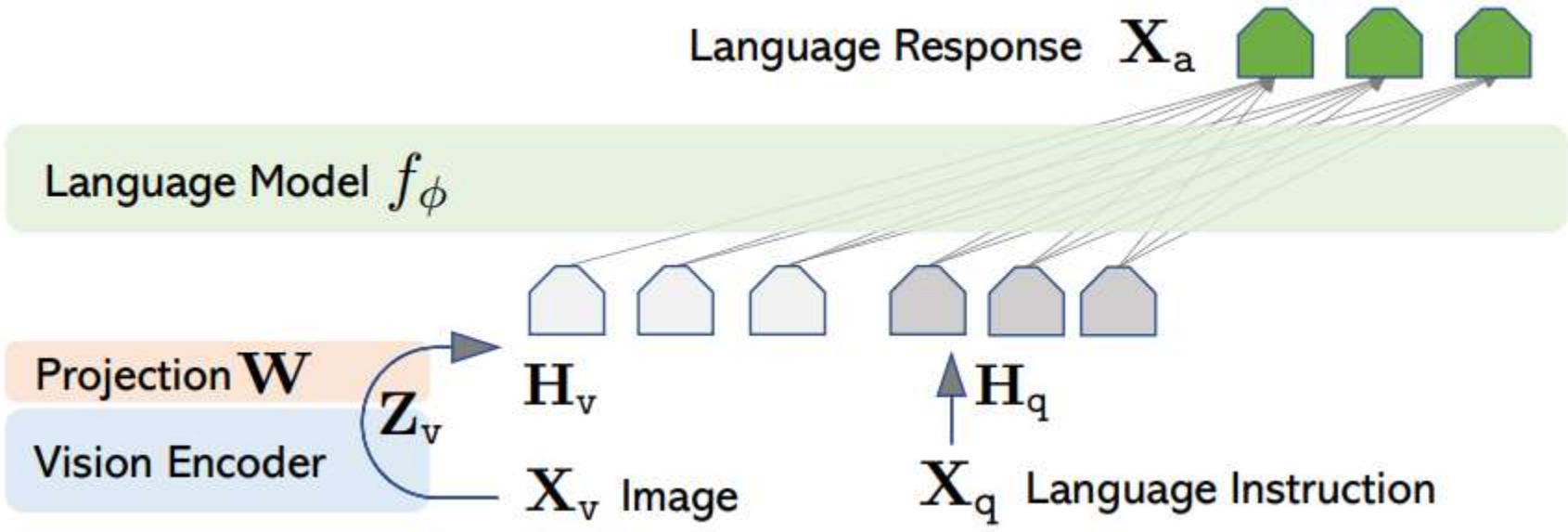
- 另一类多模态生成范式是基于**离散图像Token的图像生成（VQ-VAE，MaskGIT）**：
  - 离散图像Token先通过一个提前被训练好的图像离散Tokenizer将连续的图像处理成离散的index和对应的token embedding向量；
  - 假设输入的文本token sequence是 **c**，那么离散图像模型通过GPT类型的自回归训练，或者BERT类型的掩码训练，可以通过多次自回归或者掩码预测迭代，根据文本 **c** 生成离散的图像index，再经过查表可以得到连续的图像token embeddings。然后经过图像离散Tokenizer解码出生成的图像。





多模态理解的经典范式：视觉编码器+大语言模型的组合

- 目前scale最好的多模态理解模型的经典范式是**VE + LLM的范式（LlaVA, QwenVL, dots.vlm1）**：
  - 假设**x0**是一个图像数据。**x0**通过一个视觉编码器VE提取视觉特征，然后通过一个投影层和主干大语言模型的token embedding维度对齐；然后视觉token和语言token一起排列；
  - 训练这类模型的方法是在视觉理解语料数据上进行经典的next-token-prediction训练。
  - 理解模型的模型和训练技术在工业界和学术界都趋向于标准化（当然最近还有一些diffusion LLM的热点研究）
  - **受限于时间原因，我们本次talk重点讨论图像生成的部分。**



给定以上背景知识，现在的问题就变成了：

我们如何可以训练一个具备很强理解能力的生图模型?

大模型训练三要素：

- **模型结构**：如何同时支持 *理解* 与 *生成*？LLM？Diffusion？Hybrid？
- **训练数据**：哪些数据能培养模型的“视觉-语义对齐”能力？哪些数据能培养“生成美学 / 细节控制”能力？哪些数据让模型学会理解生图能力？
- **训练infra**：大规模多模态生成理解训练是否需要特殊并行策略？

接下来我们将交错讨论针对多模态生成大模型的开源数据的演化，以及理解生成模型的模型方案的演化；

注释：本talk提到的数据集均为开源数据集，模型方案均取材于公开发表的论文或技术报告。限于讲者本人水平有限，如有错误，烦请指正！

多模态生成理解的开源数据演化

rednote | hi lab

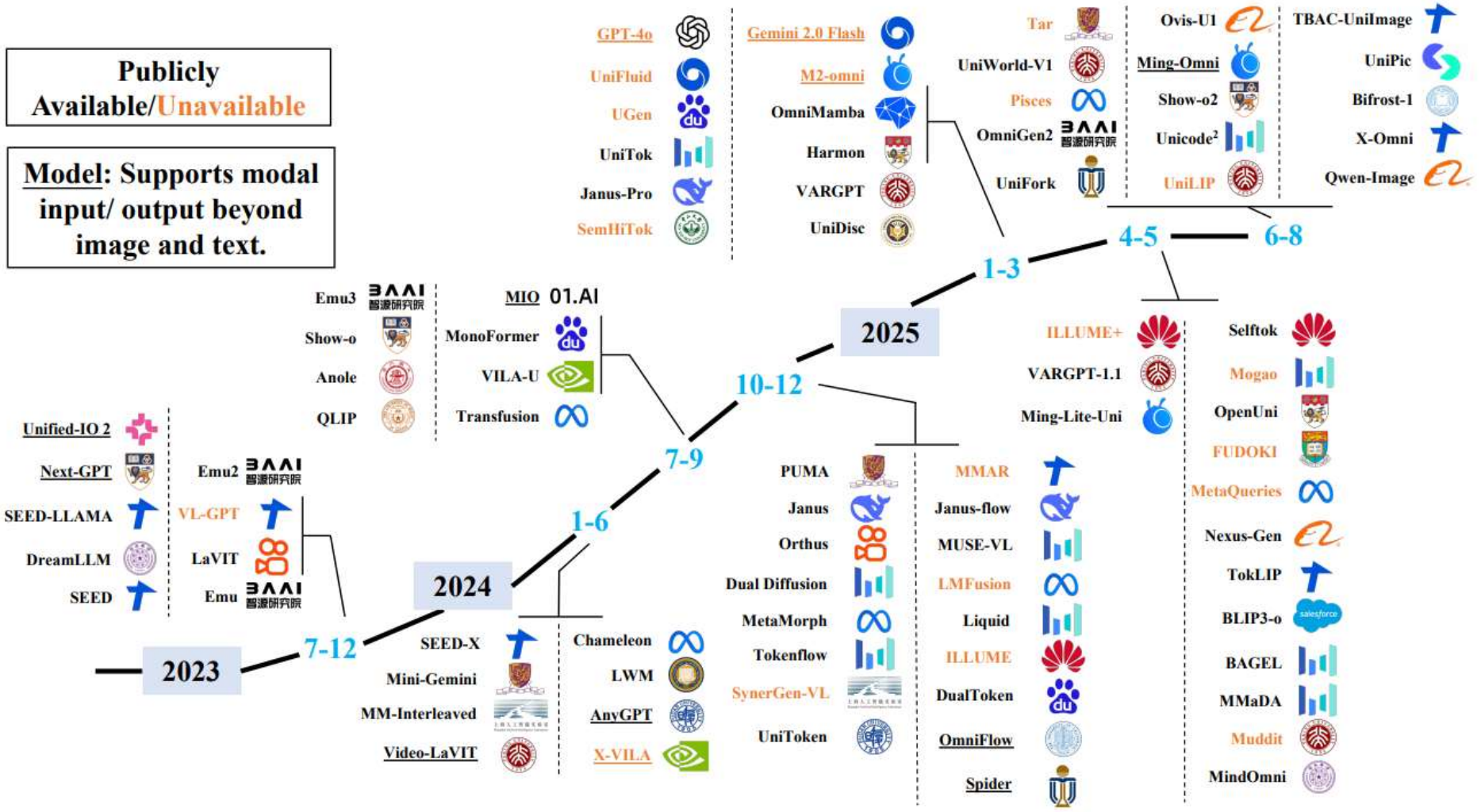
多模态生成理解数据的演化路径（总览）

阶段	数据形式	代表开源数据集	能力上限	特点
图文对 (Image-Text Pairs)	一图 + 一文本描述	Laion5B, CC12M, JourneyDB等	基础对齐	描述浅，缺乏推理
单图局部编辑数据（Local Editing）	原图+编辑指令→编辑图	InstructP2P, HQ-Edit, GPT-Edit等	图像的局部编辑	图像变化程度小
主体驱动的图文生图数据（Subject-driven）	带有人物/物体一致性约束的context到图的数据	OmniGen1&2，Virtual-Tryon等	保subject生成	数据稀缺，构造难度大
图文交错 (Interleaved Image-Text)	图/文多段交错	MMC4，OBELISC，OmniCorpus等	语境理解	粒度混乱，噪声大
多图视觉推理数据（Visual Reasoning）	具有视觉逻辑关系的多图数据	BM-6M	视觉推理能力	视觉逻辑关系普遍较弱

多模态生成理解模型的建模方式演化

rednote | hi lab

多模态生成理解模型的演化路径（总览）



图片摘自参考文献： Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities

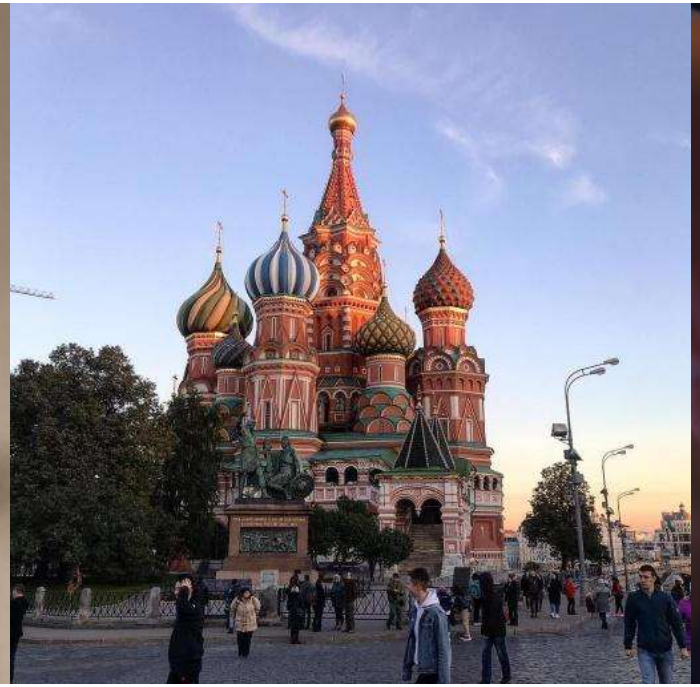
类型1：图文对数据（Image-Text Pairs）

- **典型形式：**一张图片 + 一段简短描述：
- **数据样例：**

开源数据：Laion-Aesthetic

文本：Padraig Cottage Padraig Cottage, Child / Youth Slippers

开源数据：Conceptual 12M



文本：Red Square is a city square in Moscow, Russia.

开源数据：JourneyDB

文本：A cute baby boy and a chibi kitten share a magical moment ...

- **可以说，图文对这种形式的数据容易获取，量大，噪声大，是目前占比最多的图像生成数据；这类数据作为生成模型的基础数据，经过清洗，构造后可以演化出许多不同层次的数据，有的数据可以被用于预训练，有的被用于后训练；**
- **常见的开源图文对数据有：**Laion5B以及子数据集，CC12M & 3M，SAM1B，RedCaps，DataComp，JourneyDB，DiffusionDB，DenseFusion-1M，PixelProse，PD12M，Megalith-10m，Llava-OneVision，FLUX-Reason-6M等

早期扩散生成模型 —— 图文对训练范式（以 Stable Diffusion 为例）

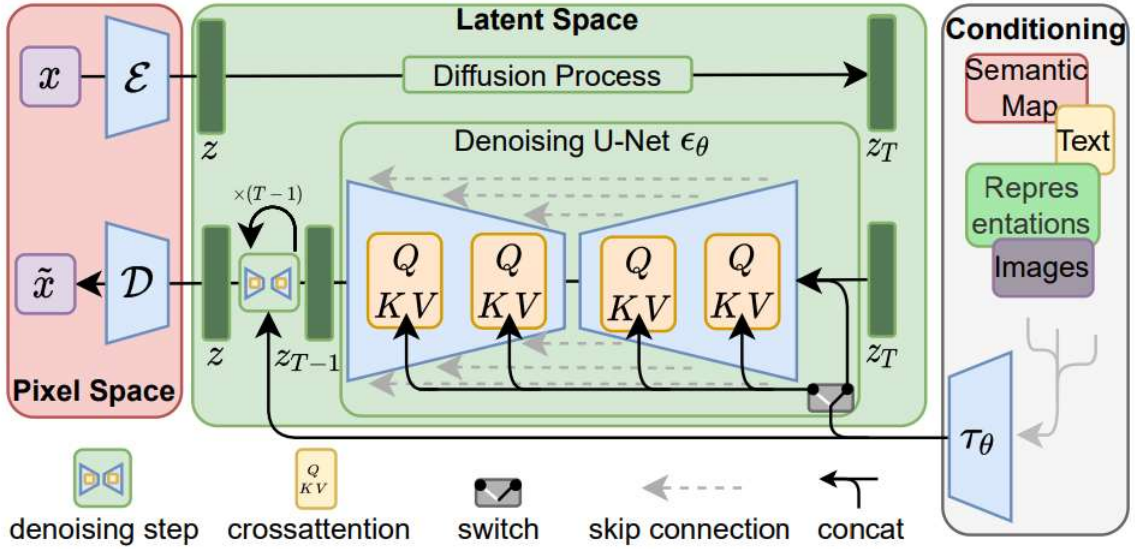
rednote | hi lab

早期多模态生成模型的训练数据来自：大量 *噪声较高* 的「图文对」（Image-Text Pairs）

- 代表数据集：LAION-400M / 5B、Conceptual Captions、COCO Captions
- 每条数据只有一句说明文字（Caption），无法形成复杂上下文或推理逻辑
- **但量大胜于质**，足以让模型学会 **图像 ↔ 文本 的粗糙对应关系（对齐 Alignment）**

Stable Diffusion =「CLIP文本编码器（定长）」+「图像压缩VAE」+「跨模态扩散UNet」

模块	具体结构	作用
文本编码器 (Text Encoder)	CLIP Text Encoder (双向注意力Transformer)	提取 Prompt 的定长语义向量，作**生成条件**
图像压缩编码器 (VAE Encoder)	Variational Autoencoder （Downsampling ConvNet）	将图像压缩到 **潜空间 (latent space)**，减少生成计算量
噪声预测 / 去噪编码器 (U-Net Diffusion Model)	U-Net + Cross-Attention	在噪声空间中逐步去噪，**融合文本信息生成图像**



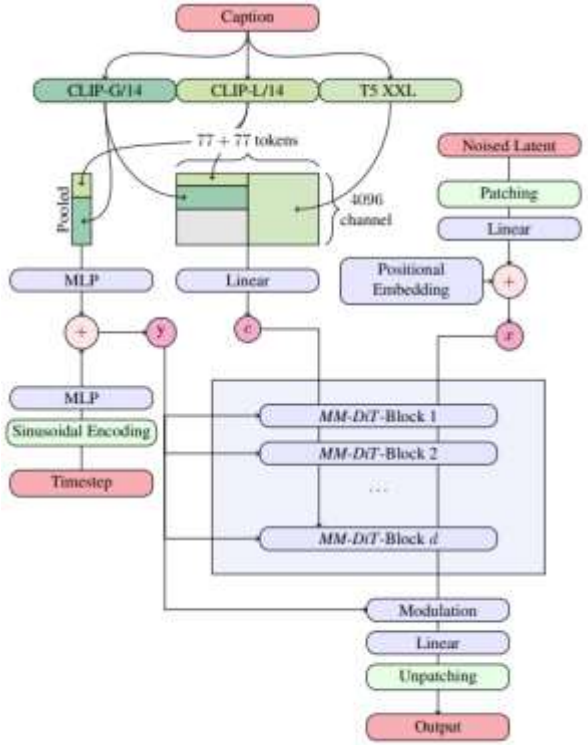
结论：在海量图文对上训练的Stable Diffusion *确实具备“理解能力”*，但非常**单向 & 浅层。**Stable Diffusion不具备图像理解能力。

经典的图像编辑模型（SD3，FLUX.1-dev 等）

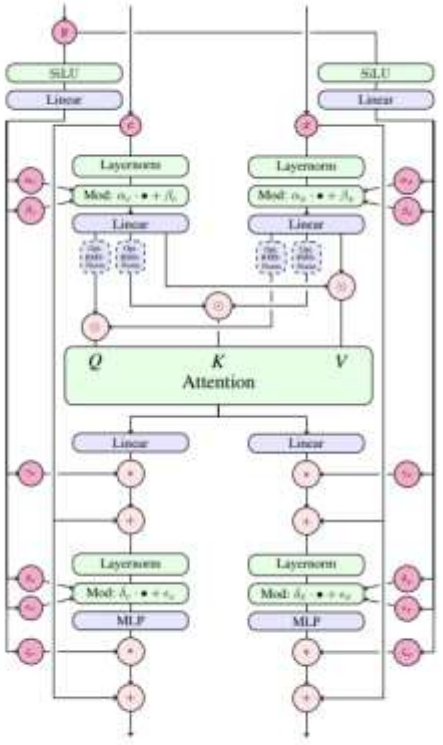
Scaling Rectified Flow Transformers for High-Resolution Image Synthesis

- 核心设计：文本离散Token和图像离散Token混合进行GPT类型的自回归训练；

模块	具体结构	作用
文本编码器 (Text Encoder)	T5-XXL + CLIP-G + CLIP-L	提取 Prompt 的定长语义向量，作**生成条件**
图像压缩编码器 (VAE Encoder)	Variational Autoencoder （Downsampling ConvNet）	将图像压缩到 **潜空间 (latent space)**，减少生成计算量
噪声预测 / 去噪编码器 (Diffusion Transformer)	MMDiT: double && single stream blocks；	通过MMDiT，让图像和语义向量双向关注，建模条件图像概率



(a) Overview of all components.



(b) One *MM-DiT* block

- 相比于SD的CLIP-Text Encoder+UNet的设计，FLUX.1-dev用了T5-TextEncoder的语义感知能力更强；MMDiT的scale性能更好；
- 然而，受限于T5-TextEncoder的定长限制，FLUX.1-dev对复杂指令的理解仍然较弱；而且FLUX.1-dev无法通过TextEncoder融入多模态图像信息；

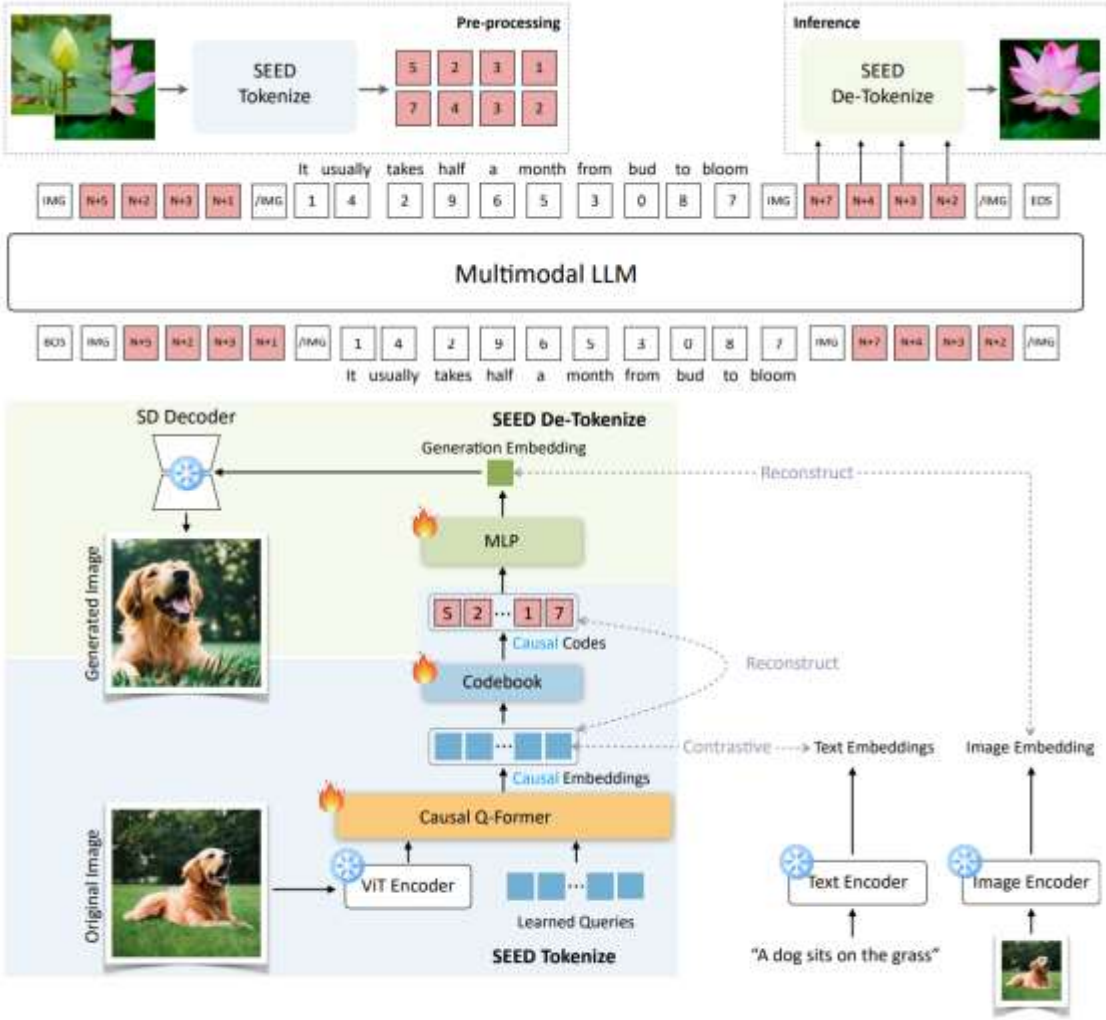
文本-图像均离散自回归的Transformer生图模型（SEED-Llama等）

rednote | hi lab

Making LLaMA SEE and Draw with SEED Tokenizer

- 主要训练数据：CC3M，Unsplash, LAION-COCO，MS-COCO；WebVid-10M，MMC4，OBELISC；
- 核心设计：文本离散Token和图像离散Token混合进行GPT类型的自回归训练；

模块	具体结构	作用
文本编码器 (Text Encoder)	LLaMA (单向注意力 Transformer LLM)	文本 → 变长的Text Token Embeddings
图像离散编码器	ViT+Q-former+Codebook+MLP	将图像压缩到 **潜空间 (latent space)**，减少生成计算量
图像生成模块	Freezed Stable Diffusion	在噪声空间中逐步去噪，**融合文本信息生成图像**



- 由于自回归语言模型的成功，纯自回归训练非常容易scale；
- 经过混合自回归训练，LLM可以同时具备理解和生成的能力；
- 其他类似思想的模型还有DreamLLM，CM3Leon，Chameleon，Next-GPT，EMU3（EMU1和EMU2不太一样，回归了CLIP-embeding）

其他纯自回归生成模型（以 Janus，EMU3为例）

Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling

Emu3: Next-Token Prediction is all You Need

- 纯自回归的路线的关键一直在于设计更好的数据+更好的Tokenizer；还有一些其他典型的纯自回归模型，例如Janus，EMU3等，均在数据和模型scale上做了更多的工作。
- 其中Janus-Pro对理解和生成采取了“双视觉编码器（VE）”的做法，即理解用VLM常用的Semantic VE，生成用Vector-Quantized（VQ）的离散VE。双VE的做法通过工程复杂性，避免了强行用一个VE的缺点，也引发了广泛讨论。
- EMU3训练了更好的VQ-Tokenizer，也做了更好的后训练。

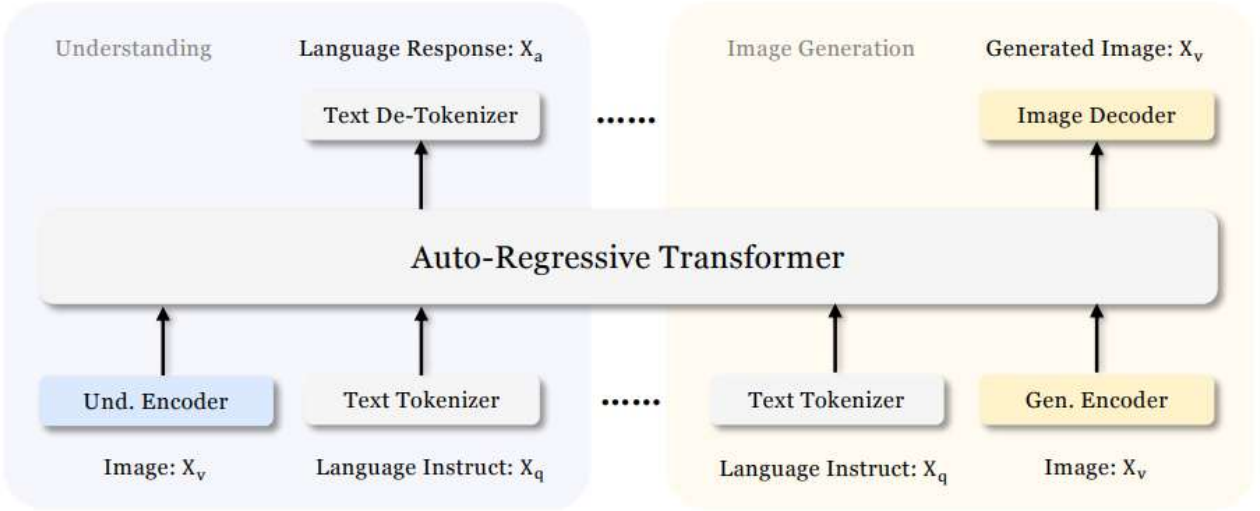
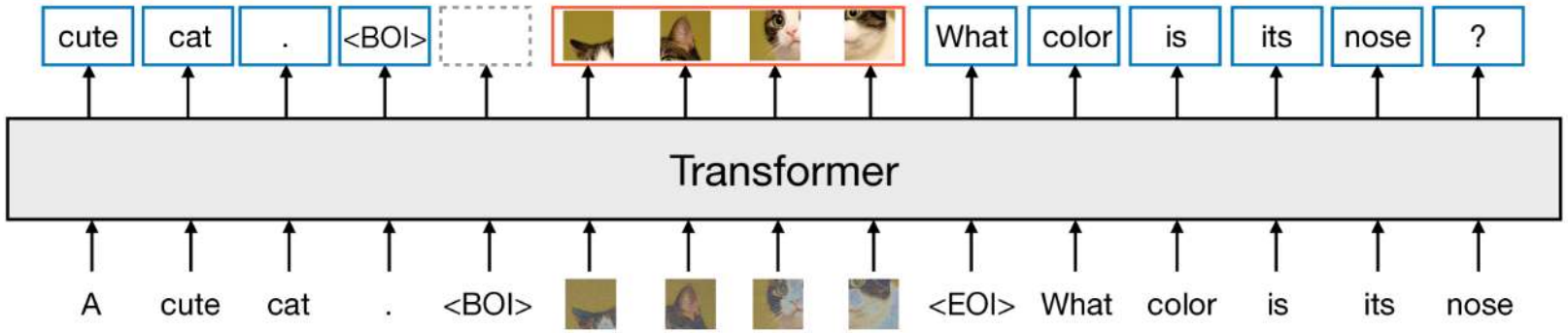


Figure 3 | **Architecture of our Janus-Pro.** We decouple visual encoding for multimodal understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed on screen.

语言离散自回+图像连续Diffusion混合一体Transformer模型（Transfusion等）

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

- 主要训练数据：1T Text Tokens from Llama2 Training Corpus；692M Images（Shutterstock，CC12M）
- 核心设计：一个Transformer模型，两个loss加和：
  - 对于理解任务，输入文本（或图像）上文，做文本next-token-prediction；
  - 对于生图任务，输入文本（或图像）上文，以及带噪声的图像noise，输出降噪图像计算diffusion loss。



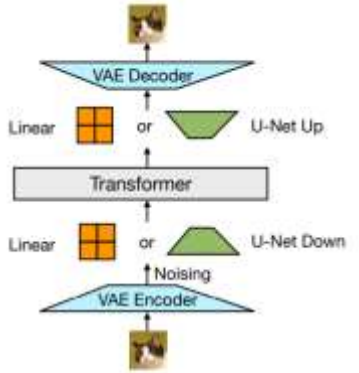


Figure 3: We convert images to and from latent representations using a pretrained VAE, and then into patch representations with either a simple linear layer or U-Net down blocks.

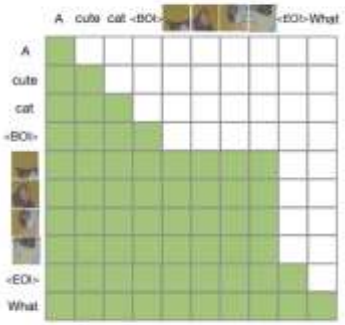


Figure 4: Expanding on the causal mask, Transfusion allows patches of the same image to condition on each other.

- 通过将NTP loss和diffusion loss加权求和，Transfusion可以让一个Transformer同时学会生成和理解两个能力；
- 通过设计attention mask，Transfusion可以通过padding或者packing的形式高效训练；

类型2：单图局部编辑数据（Local Editing）

- **典型形式：**一张图片 + 一段图像编辑指令 → 编辑好的图像：
- **数据样例：**

开源数据：InstructP2P-1M

编辑指令：Turn it into a photo
【左边原图，右边编辑图】

开源数据：X2Edit-4M

编辑指令：Can you add a small decorative pillow next to the couch?

开源数据：Nano-150K

编辑指令：Change the hair style to blond and straight.

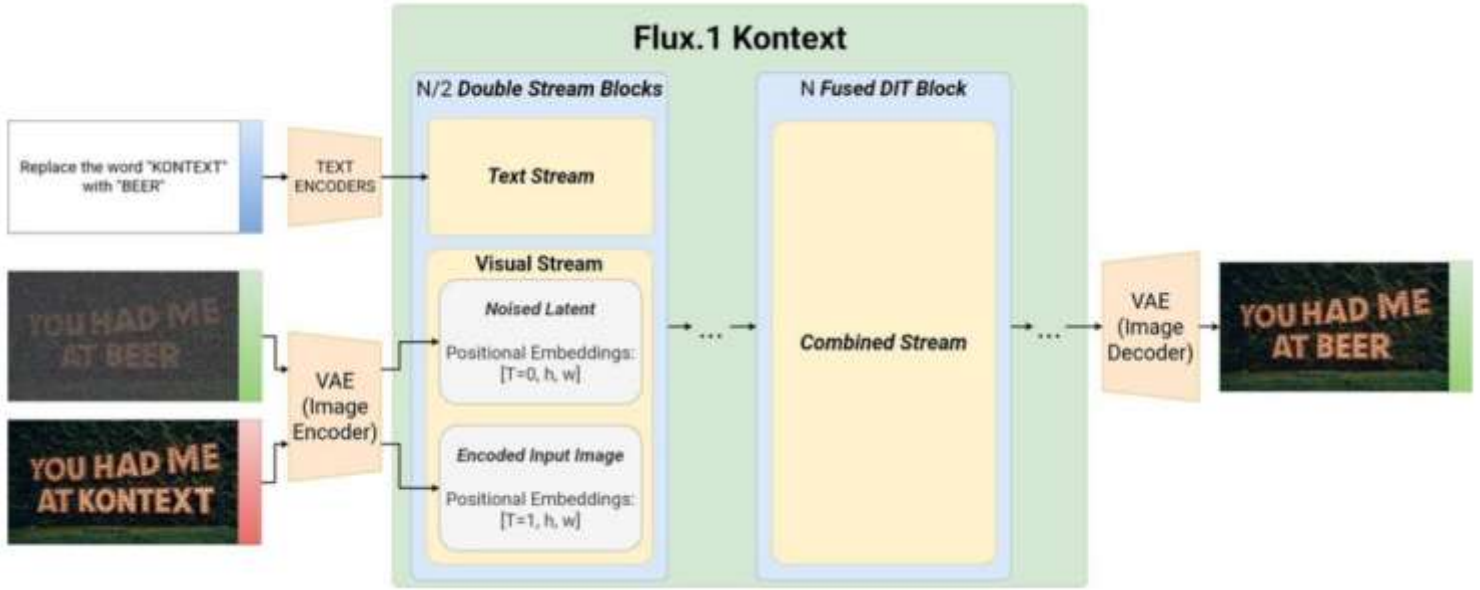
- **在很长一段时间里，图像的局部编辑更多的被看作是一个生成模型的应用，而不是生成理解的主线，直到2024年底的一些工作才有所改观；过去若干年，在图像编辑这个应用场景中，涌现出了许多质量不错的开源数据。**
- **常见的开源局部编辑数据有：**InstructP2P，OmniEdit，SEED-Data-Edit，OmniGen1&2，GPT-Image-Edit，X2Edit，Echo-4o，Share-GPT-4o-Image，HQ-Edit，Ultra-Edit，PromptFix等；
- 近期的许多工作开始探索了，较大幅度的图像编辑，例如Flux-Kontext，BAGEL，Qwen-Image，Nano-Banana等；不过相应的开源数据集很少，并且体量不大。

经典的图像编辑模型（InstructPix2Pix，FLUX-Kontext）

InstructPix2Pix: Learning to Follow Image Editing Instructions

FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space

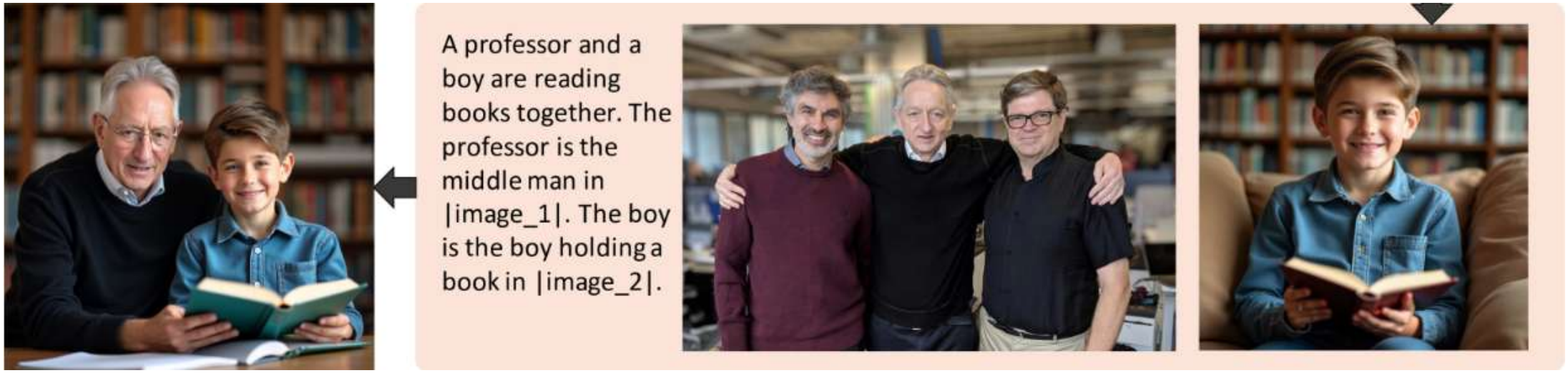
- InstructPix2Pix是最早训练Diffusion Model做图像编辑的工作；他发现，只要将条件图的VAE latents在Channel维度直接拼到Noisy Latent上，然后小改Stable Diffusion的UNet加以训练，模型就可以取得非常好的编辑效果；
- FLUX-Kontext（以及一些其他基于Diffusion Transformer的工作）发现，对于DiT，将参考图直接拼接到noisy Latent序列的后边，就可以让模型学会如何“看”参考图。
- 经过数据质量的提升，和大规模的训练，FLUX-Kontext可以做到非常好的局部编辑效果；



类型3：Subject驱动的ID融合数据

典型形式：一张或者多张不同主体的图片 + 一段语义较为复杂的主体融合指令 → 将多主体融合好的图像：

- Subject-driven数据形式让模型学习到生图的过程中该如何从输入的多张图中去寻找知识；



Subject驱动的开源数据样例：OmniGen1开源数据

OmniGen: Unified Image Generation

- **OmniGen1的数据构造是一个比较典型的subject-driven数据的例子。** 该团队针对开源的人物图像，调用web-search获取到丰富的identity参考图作为输入。然后用视觉大语言模型构造生图指令。整个数据构造的过程对后续许多工作有较大的借鉴意义。
- 类似的OmniGen1也引入了不少经典CV的数据，例如深度感知，语义分割，姿势检测等。这样可以让生成模型真正尝试习得“通用生成能力”，即将传统CV任务在图像生成这个范式下进行一个统一。

数据加持的通用生成Transformer；

OmniGen: Unified Image Generation

- 主要训练数据：
  - 图文对：RecapDataComp【56M】，SAM-LLaVA，ShareGPT4V，LAION-Aesthetic【4M】，DenseFusion等；
  - 图像编辑：InstructPix2Pix，SEED-edit，human motion，virtual try-on，StyleBooth，CV data等；
  - 其他数据类型：Subject驱动多图数据，subject驱动单图数据，传统CV数据等；
- 核心设计：一个Transformer模型，只训练Transfusion的Diffusion Loss：

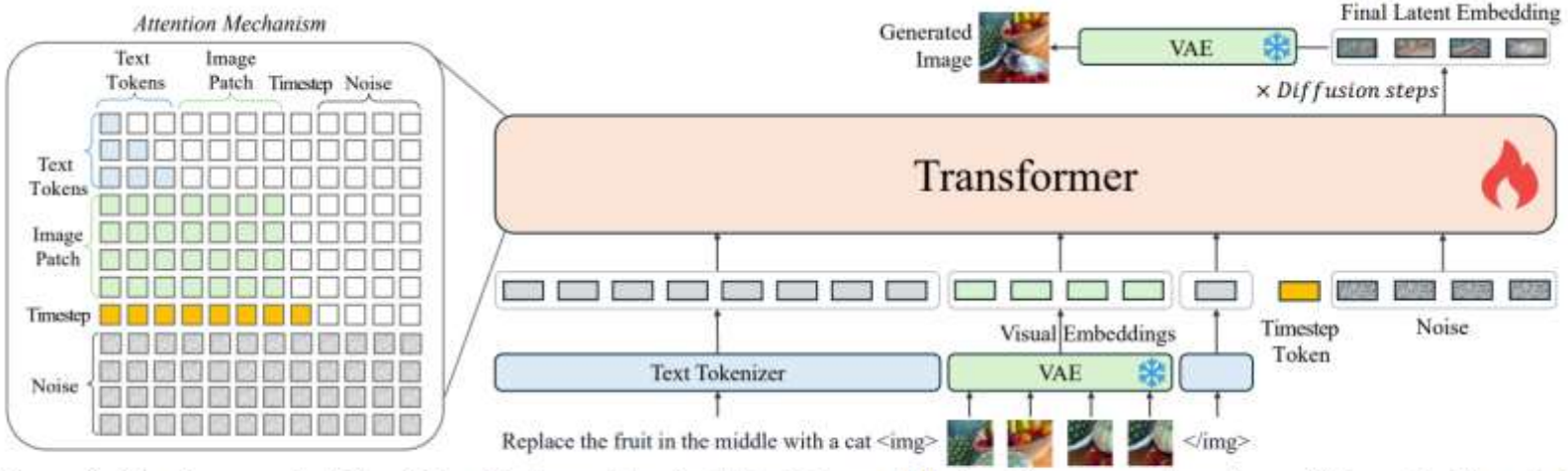
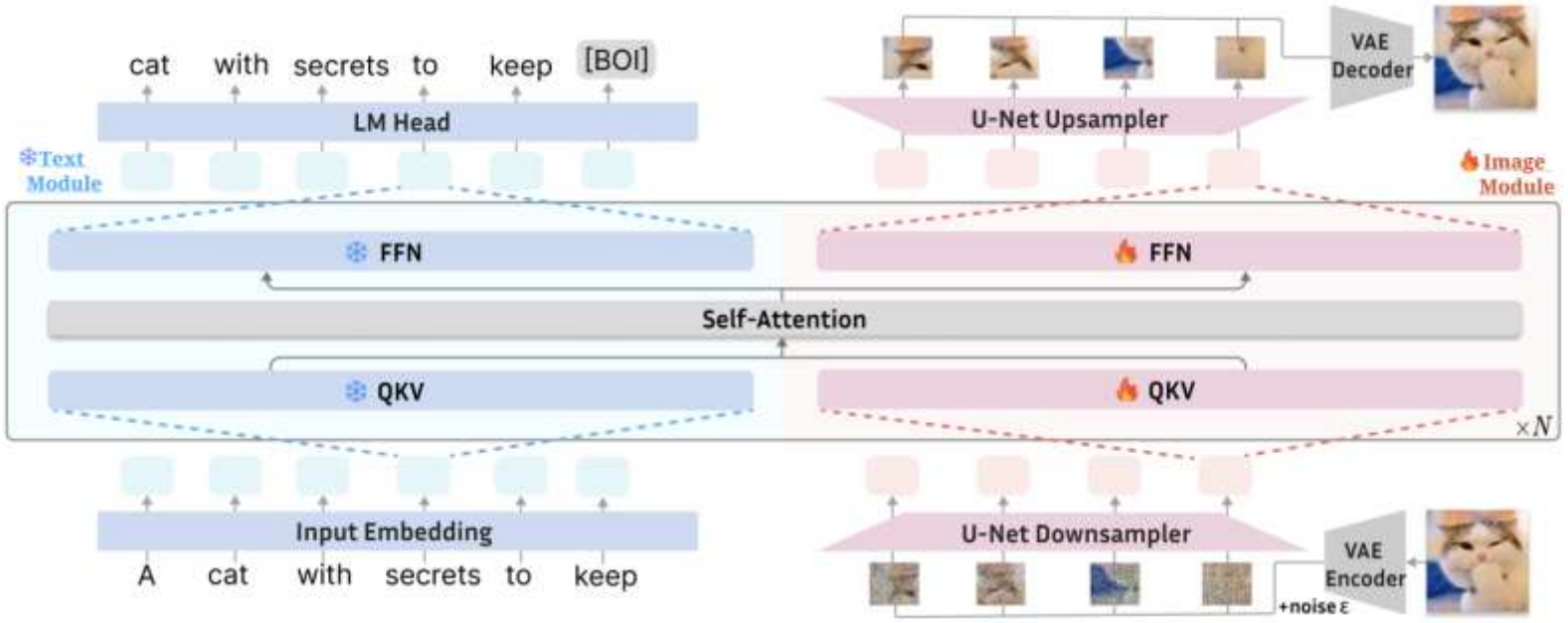


Figure 2. The framework of OmniGen. Texts are tokenized into tokens, while input images are transformed into embedding via VAE. OmniGen can accept free-form multi-modal prompts and generate images through the rectified flow approach.

两个并行的Transformer模型（LM-Fusion）

LMFusion: Adapting Pretrained Language Models for Multimodal Generation

- Transfusion采用单一Transformer模型同时学习理解和生成；后续有团队研究了采用两个并行的Transformer进行生成理解同时学习。LM-Fusion是一个典型的例子；
- 主要训练数据： Transfusion语料；380M Shutterstock 图文对；
- 核心设计：LM-Fusion对一个Transformer内的生成和理解Token分别用了不同的QKV权重和FFN权重，但是语言和视觉Token共享一个Self-Attention；这种设计相比于Transfusion将文本和图像的计算进行了解耦；



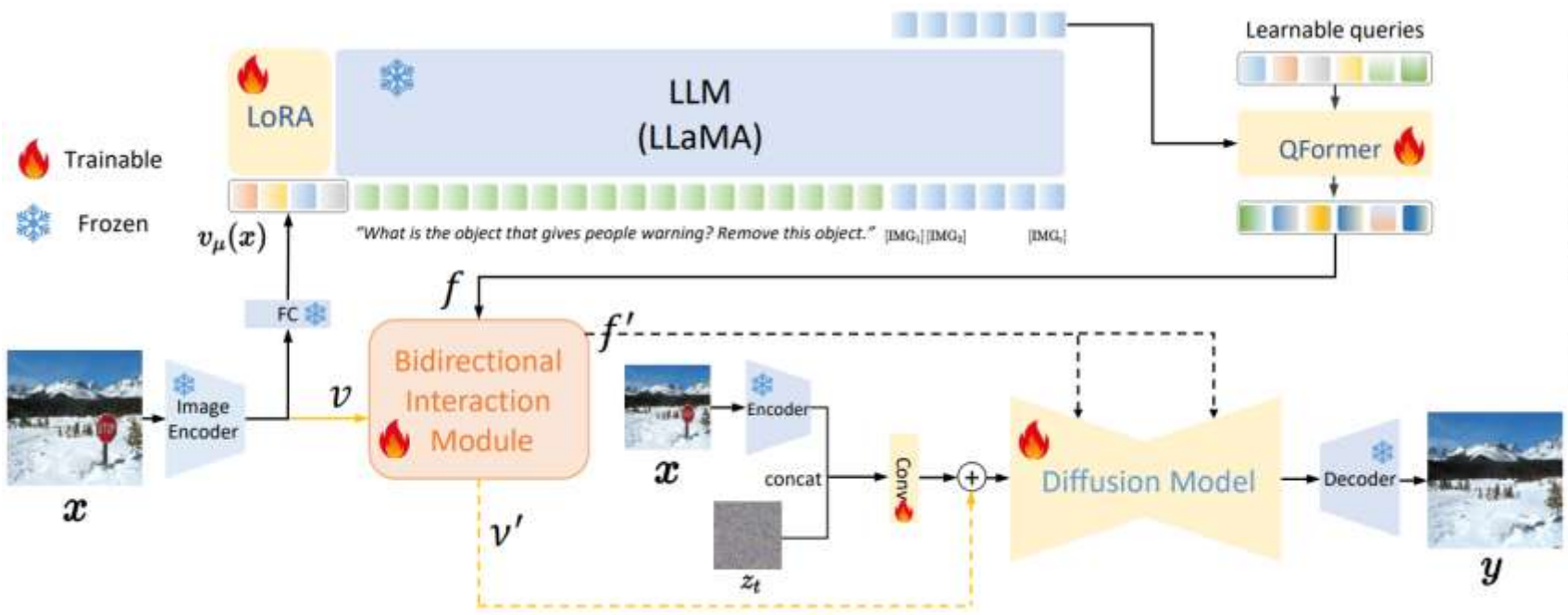
两个串行的Transformer模型（LLM+Diffusion Decoder：SmartEdit）

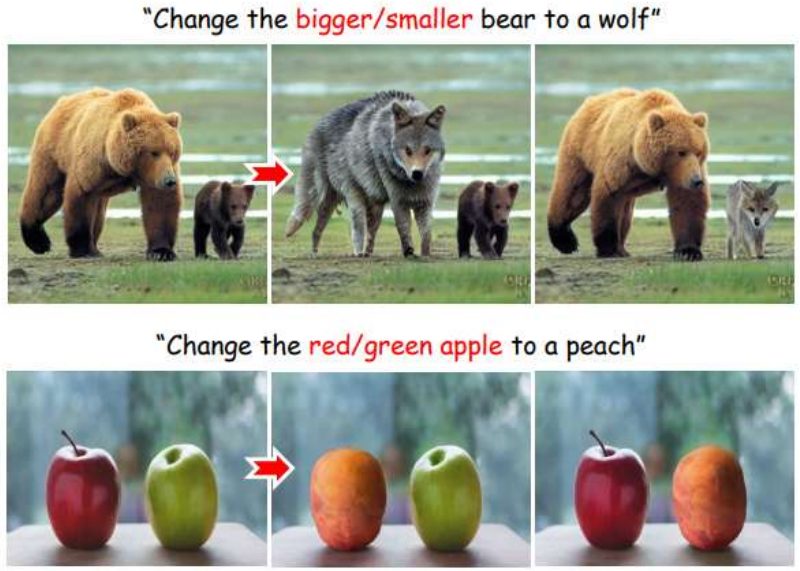
rednote | hi lab

SmartEdit: Exploring Complex Instruction-based Image Editing with Multimodal Large Language Models

SmartEdit 是一个比较早的图像编辑模型，比较早采用了LLM+Diffusion Decoder的思路；

- 主要训练数据： 25M图文对； 2.4M 图图对（从MMC4洗出来）
- 核心设计：大语言模型生成语义向量（结合Qformer变成定长）；然后通过特征分发的方式传给UNet形式的Diffusion Decoder；另一支路将VE的特征向量给到Diffusion Decoder；
- 这种设计的好处在于，大语言模型可以充分编码指令的语义，实现“智能指令”；坏处是纯粹的VLM提取的特征偏向视觉语义，较难完成精确重建；另外大规模训练Infra也有一些挑战；



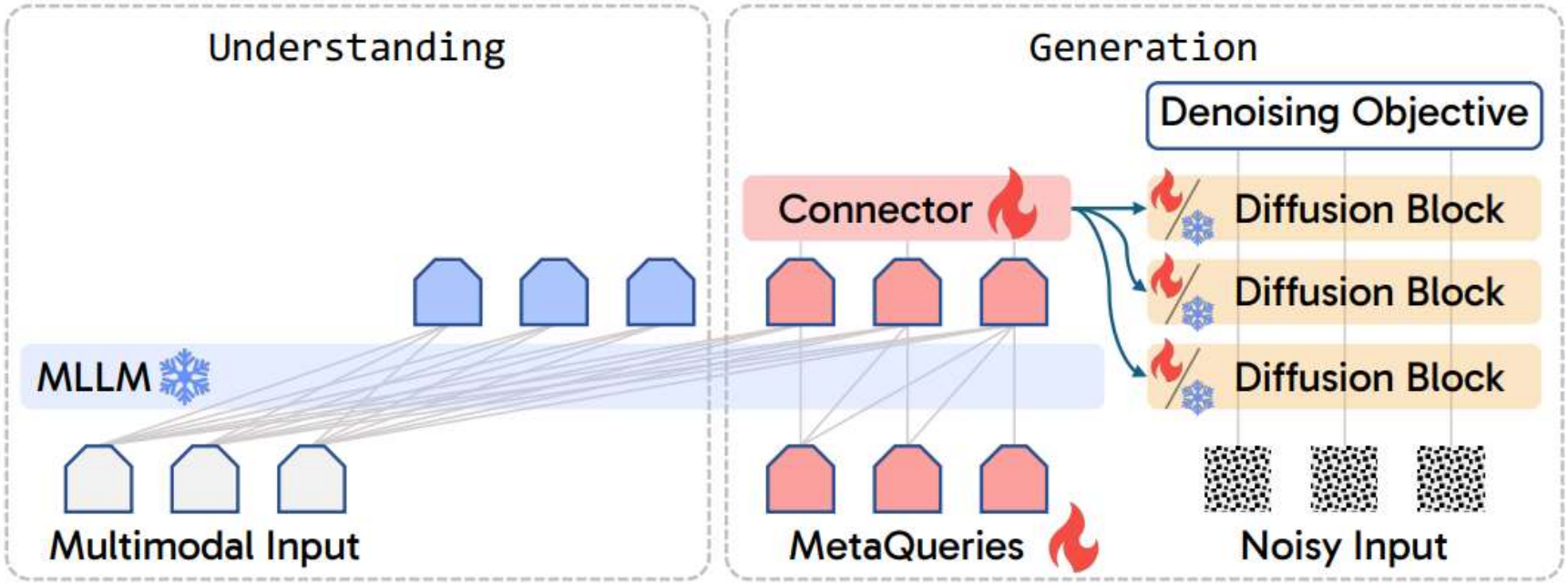


两个串行的Transformer模型（LLM+DiffusionDecoder：Metaquery）

Transfer between Modalities with MetaQueries

不同于Transfusion这类单Transformer模型；类似于Metaquery这类模型采用一个LLM+DiffusionDecoder的思路，也可以实现相对解耦的理解和生成能力；

- 主要训练数据： 25M图文对； 2.4M 图图对（从MMC4洗出来）
- 核心设计：大语言模型生成语义向量（结合Metaquery变成定长）；然后通过特征分发的方式传给Diffusion Decoder；
- 这种设计的好处在于，可以充分利用预训练好的视觉大语言模型和Diffusion Model；坏处是数据packing比较复杂，对大规模训练的infra要求较高；



类型4：图文交错数据

典型形式：图-文混合排列数据，常见于互联网页数据，以及公开的多种文档数据；

"The rectangular boxes shown below make a beautiful addition to any room in your home <image_0>. Their uses are just about limitless and they make terrific gifts for groomsmen, a graduate, or for you! They are available in oak, cherry, or walnut and are 8 inches long, 5 inches wide, and 3 inches deep. <image_1> The bottom of the interior is lined with a deep red velveteen for a more elegant look, and felt dots on the bottom prevent scratching of any surface.", "A thoughtful, beautiful gift for just about any occasion!"

<image_0> <image_1>

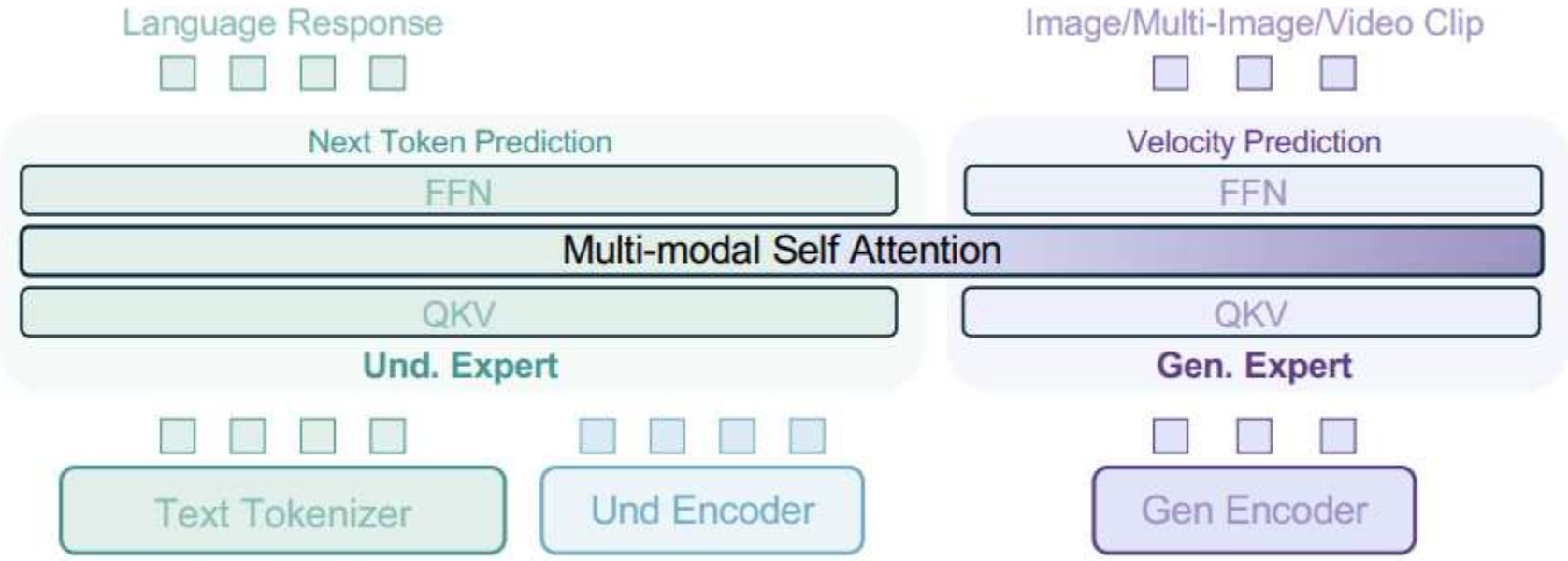
Interleaved开源数据样例：MMC4开源数据

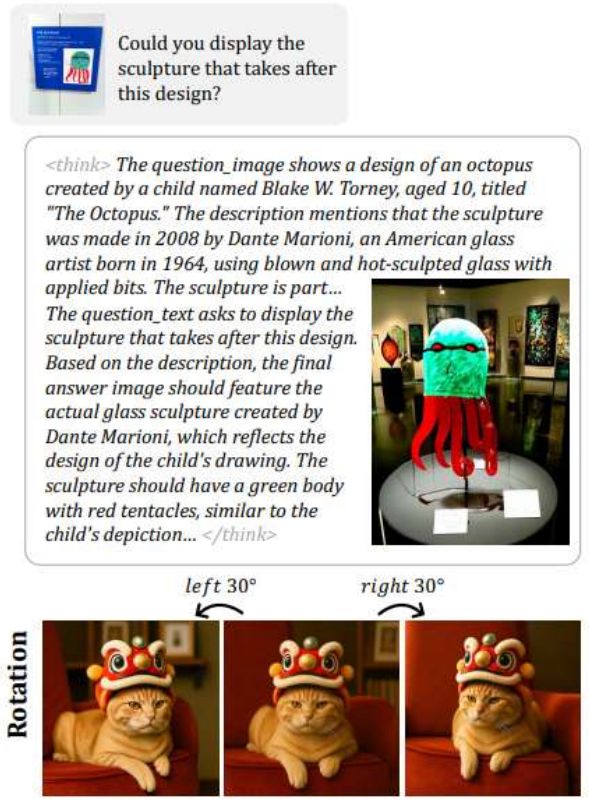
两个并行的Transformer模型（BAGEL）

Emerging Properties in Unified Multimodal Pretraining

类似于LM-Fusion，BAGEL也采用两个并行的Transformer进行生成理解同时学习；不过，BAGEL的训练数据集更大；

- 主要训练数据： 图文对，图像编辑；图文交错数据；带推理思维链的文生图数据；
- 核心设计：Transformer的核心结构和LM-Fusion一样，对文本和视觉采用了不同的QKV-FFN。不过BAGEL在输入端额外输入了参考图视觉VAE latents的patch sequence。数据方面，引入了通过Interleaved数据和视频数据构造图文图组。
- 通过提升数据质量和数据规模，BAGEL-7B具备较好的生成-理解能力。





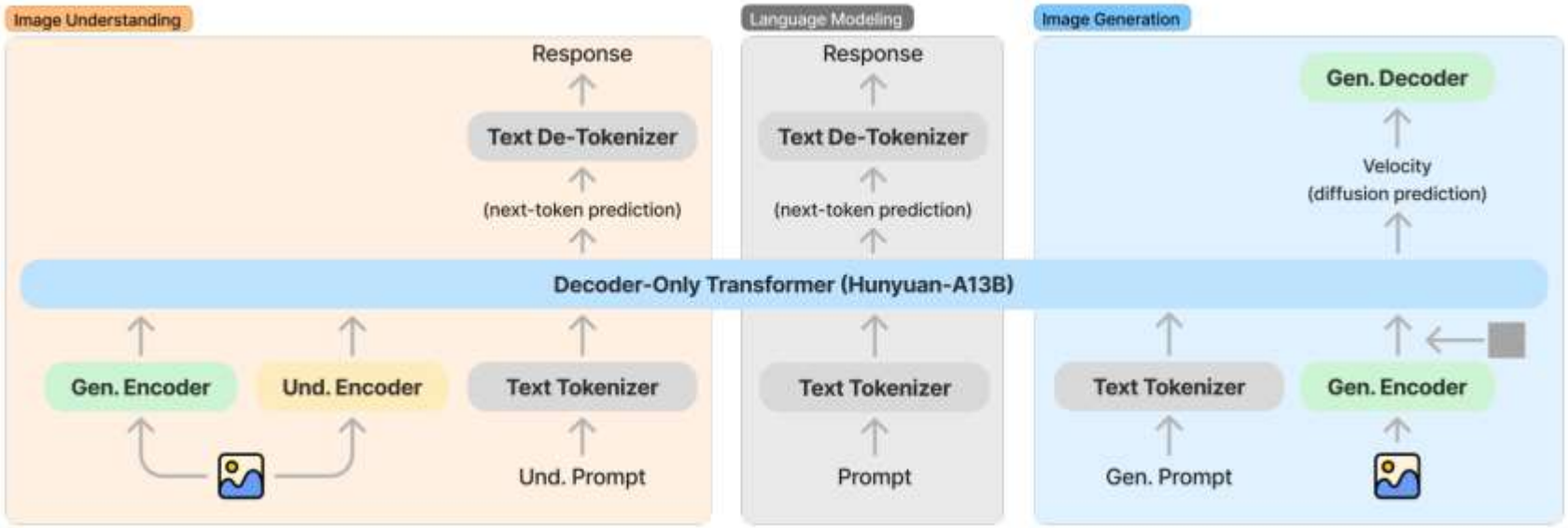
n个并行的Transformer（MoE架构的Hunyuan Image 3,0）

rednote | hi lab

HunyuanImage 3.0 Technical Report

Hunyuan-Image-3.0是一个A13B-T80B的MoE生成理解一体的Transformer；

- 主要训练数据： 自有冲标签的图文对；文本数据；带推理思维链的文生图数据；
- 核心设计：MoE Transformer，生成理解双VE。数据过滤和重构做的扎实，数据质量提升。



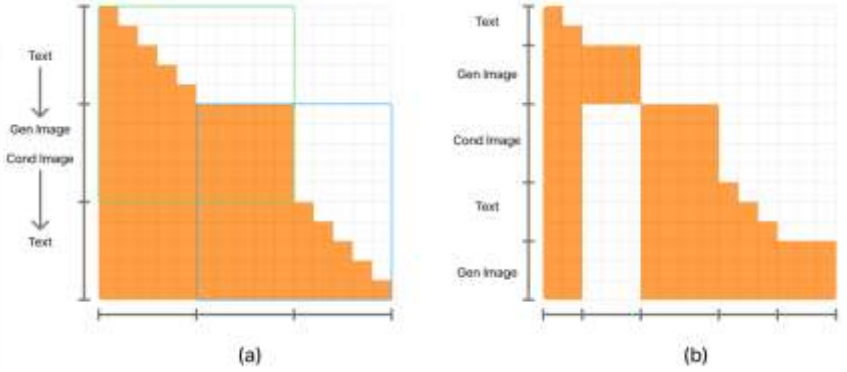
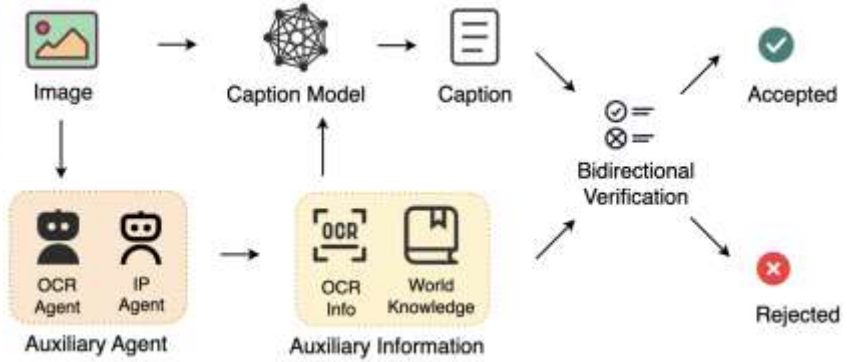


Figure 4: Two types of attention implementation.



LLM+Diffusion Decoder（Qwen-Image）

Qwen-Image Technical Report

Qwen-Image是一个7B VLM+ 20B MMDiT的串行结构；

- 主要训练数据： 比较扎实的数据管线：过滤，打标；
- 核心设计：Qwen2.5-VL的最后一层hidden-states做图文encoder，带参考图扩展的MMDiT做Diffusion Decoder。

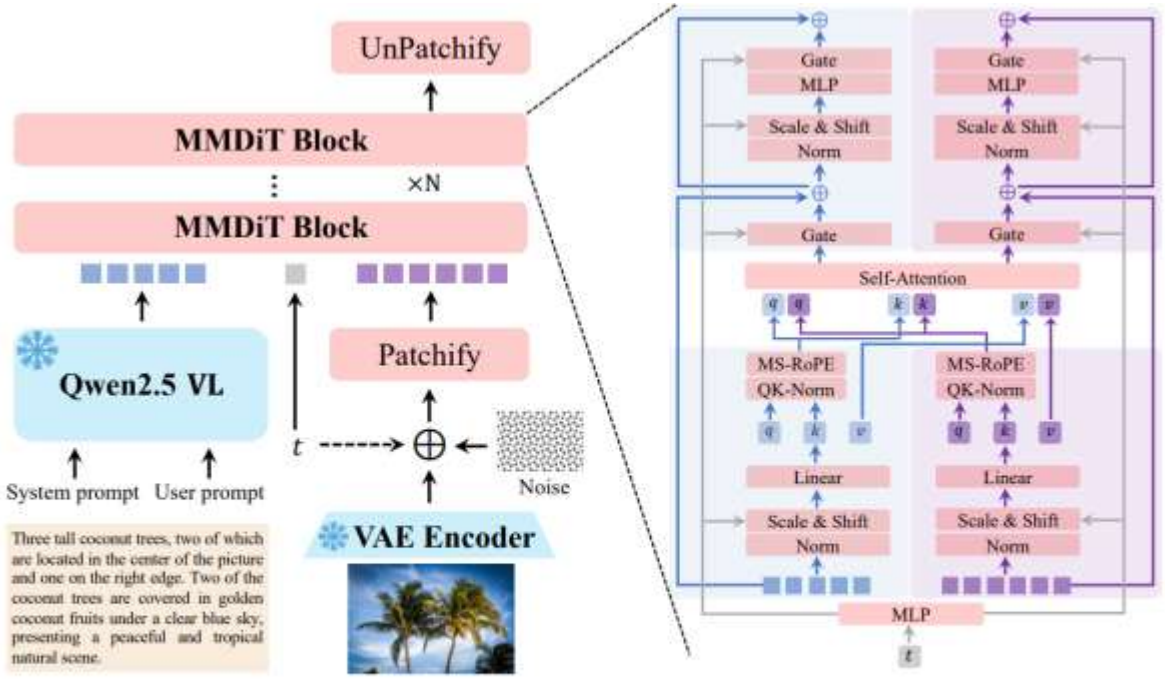
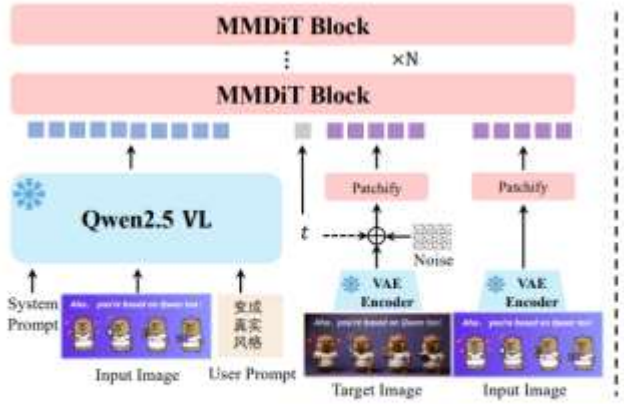


Figure 6: Overview of the Qwen-Image architecture. It adopts a standard double-stream MMDiT architecture. The input representations are provided by a frozen Qwen2.5-VL and a VAE encoder. The model employs RMSNorm (Zhang &Sennrich, 2019) for QK-Norm, while all other normalization layers use LayerNorm. Additionally, we design a new positional encoding scheme, MSRoPE (Multimodal Scalable RoPE), to jointly encode positional information for both image and text modalities.



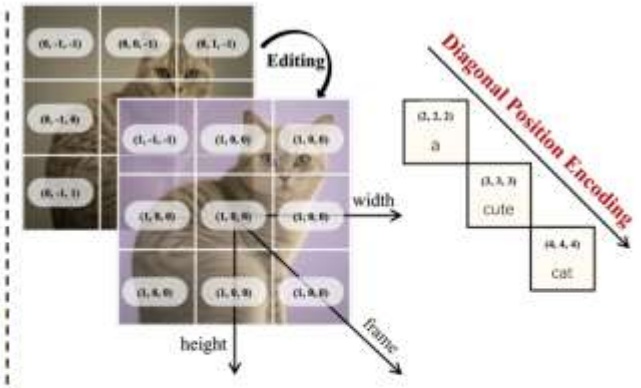
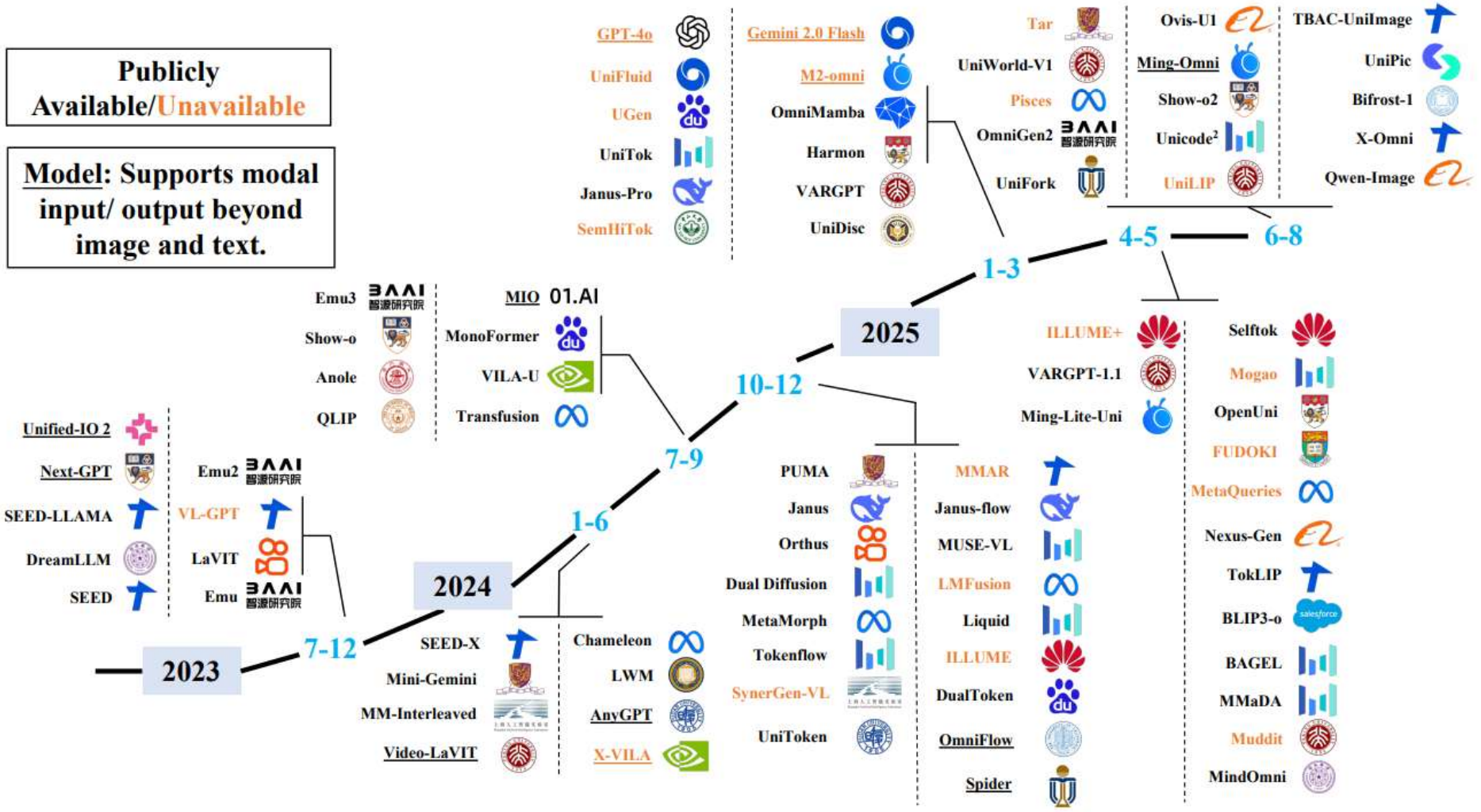


Figure 14: Overview of the Image Editing (TI2I) task. Left: Illustration of the TI2I task training input format. The user prompt is "Turn into realistic style" in English. Right: The corresponding modification to MSRoPE for TI2I, where a new dimension (frame) is introduced to distinguish between the images before and after editing.

多模态生成理解模型的建模方式演化

rednote | hi lab

多模态生成理解模型的演化路径（总览）



图片摘自参考文献： Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities

小结：

- **目前生成理解模型的方案还比较发散，主要有包括：**
  - **文本自回归+图像VQ自回归的一体Transformer：**
    - SEED-Llama，Chamelon，Janus，EMU3，LaViT，MagVIT等；
  - **文本自回归+图像连续Diffusion的一体Transformer：**
    - Transfusion，OmniGen1，LM-Fusion，Janus-flow，Mogao，BAGEL，Hunyuan-Image-3.0，等；
  - **文本自回归LLM+图像连续Diffusion Decoder的分离式架构：**
    - SmartEdit，Metaquery，MetaMorph，BLIP3o，OmniGen2，UniLiP，UniWorld-V1，SeedDream2.0，Ming-Omni， Mind-Omni，Qwen-Image等；
  - **其他：**
    - Harmon（文本自回归，图像MAR），X-Omni（文本图像先离散自回归，再接Diffusion Decoder）， Show-o（文本自回归+图像离散扩散模型）， MMaDA， FUDOKI（文本离散扩散+图像离散扩散）等；
- **不同的建模方案均获得过非常亮眼的效果；**
- **截止目前看起来，数据质量以及伴随的训练量和scale的重要性大于建模路线：**
  - 目前学界和业界开源的数据质量，无法有效验证建模方案的优劣；深耕数据是接下来半年内学业界最重要的事；

谢谢观看！请批评指正！