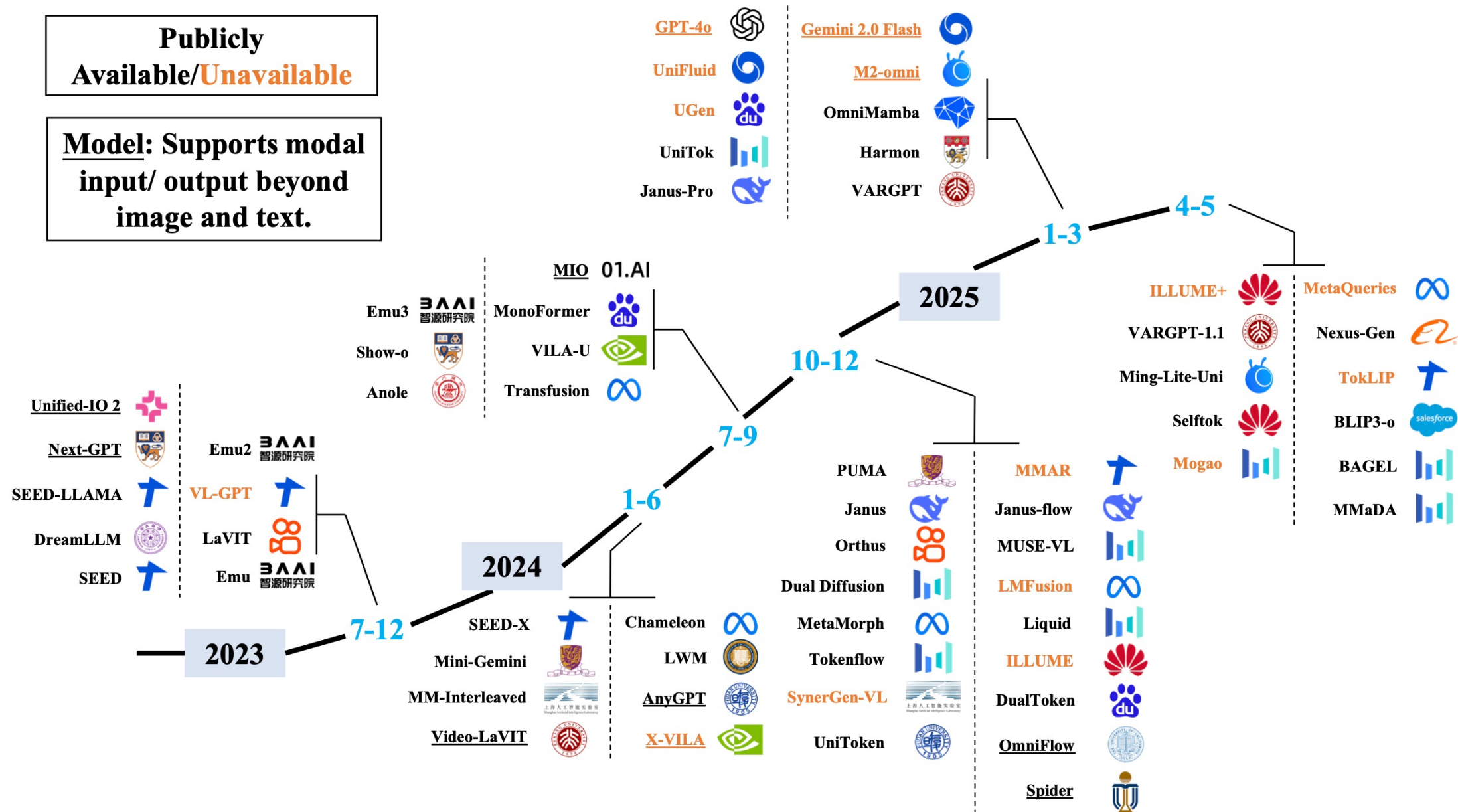




Unified Multimodal Models

<https://bagel-ai.org/>



Quantized AR – Single Encoder

这类模型使用一个单一的编码器将图像转化为离散的视觉Token

• **Representative Works:** Chameleon, Emu3

• **Pipeline:**

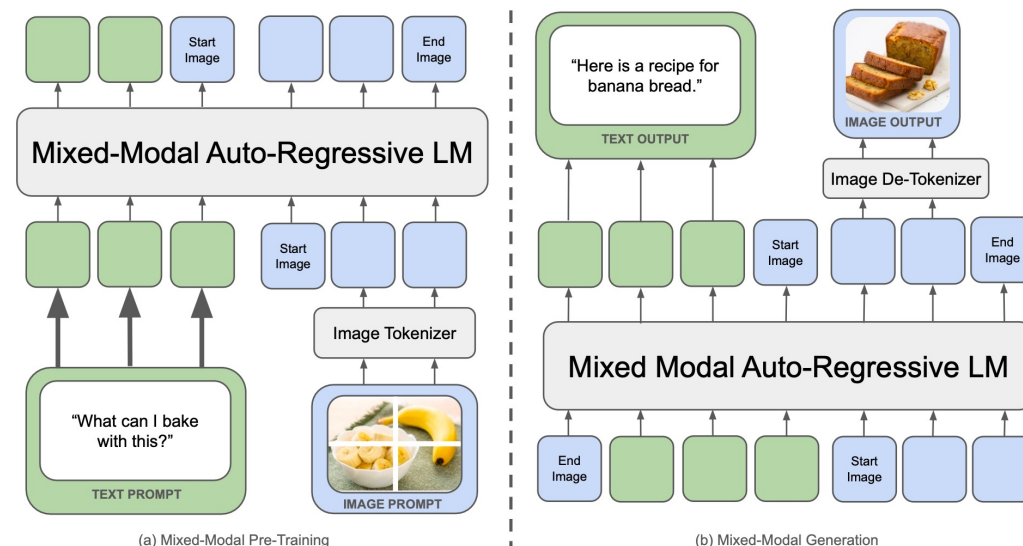
- **图像编码:** 图像通过一个预训练的VQGAN或SBER-MoVQGAN编码器, 被压缩成离散的视觉Token序列
- **Token化与融合:** 这些视觉Token与文本Token一起, 按序列排列输入到自回归LLM (如LLaMA) 中
- **自回归生成:** LLM学习预测下一个Token, 无论是文本还是视觉Token
- **图像解码:** 生成的视觉Token通过相应的VQGAN或SBER-MoVQGAN解码器重建为像素图像

• **Strengths:**

- **统一建模:** 直接利用现有LLM基建, 将图像视为离散Token, 与文本Token的处理方式高度一致, 简化了多模态内容的统一自回归建模
- **交错生成:** 结构上支持文本和图像Token的交错生成, 理论上能够处理更复杂的交互场景

• **Weakness:**

- **计算效率:** 生成的视觉Token序列通常非常长, 导致训练和推理的计算成本和内存消耗巨大, 尤其是高分辨率图像。推理延迟高, 受自回归的顺序性限制
- **生成质量:** 视觉生成质量在经验上劣于扩散模型
- **语义限制:** 视觉Token的优化主要针对像素重建, 可能缺乏高层次的语义抽象, 使得跨模态对齐和高阶推理能力受限 (理解Benchmark效果较差)



Quantized AR – Double Encoder

这类模型采用两个独立的图像编码器，一个侧重语义信息，另一个侧重像素细节

• **Representative Works:** Janus, Janus-Pro

• **Pipeline:**

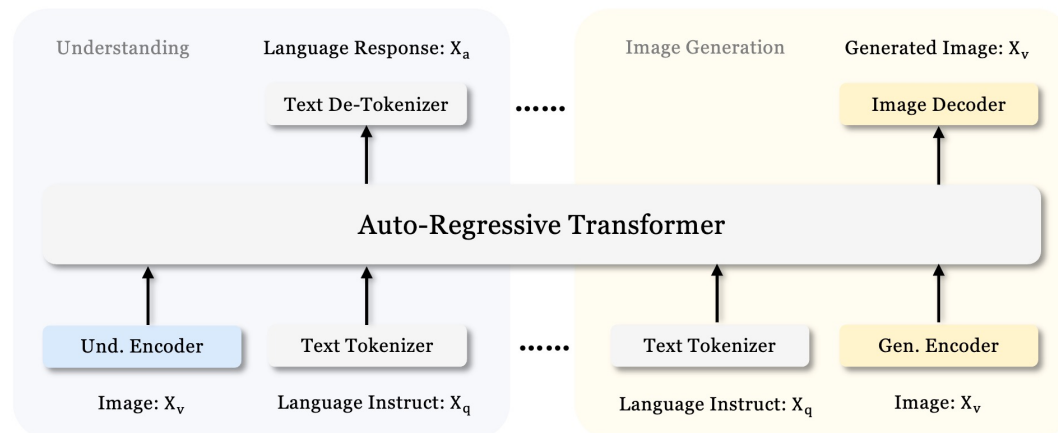
- **图像编码:** 图像并行通过两个编码器：一个语义编码器（如SigLIP）用于提取高层语义特征，一个像素编码器（如VQGAN）用于提取低层像素细节或生成量化Token
- **Token化与融合:** 两个编码器输出的特征，通过线性映射或特定处理后，输入到自回归LLM中
- **任务特定使用:** 在训练阶段，可能同时利用两种编码器的信息进行多模态理解和生成；在推理阶段，某些方法可能根据任务只激活其中一个编码器（例如理解任务用语义编码器，生成图像用像素编码器）
- **图像解码:** 通过像素解码器（如VQGAN）重建图像

• **Strengths:**

- **结合优势:** 旨在结合语义编码（高层理解）和像素编码（低层细节）的优势，以获得更丰富的视觉表示
- **训练增强:** 混合训练策略可以在一定程度上提升模型在多模态理解和生成任务上的性能
- **模块化解耦:** 允许理解和生成路径的相对解耦，简化了某些方面的训练管理

• **Weakness:**

- **信息利用不足:** 大部分方法推理时只激活一个编码器，未能充分利用两种模态的协同潜力
- **复杂性增加:** 双编码器架构会显著增加计算和内存开销
- **信号冲突:** 像素和语义Token之间的隐式对齐可能导致学习信号冲突



Quantized AR – Special Tokenizer

这类模型在AR框架下使用专门设计的视觉Tokenizer，旨在平衡语义对齐和像素保真度，或处理多粒度信息

• **Representative Works:** VILA-U, UniTok, TokenFlow, MUSE-VL

• **Pipeline:**

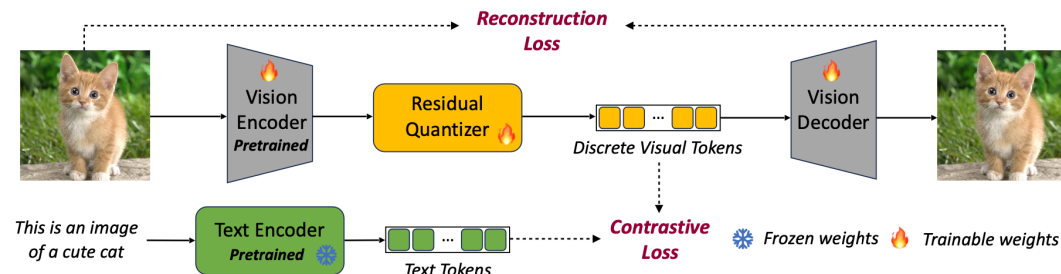
- **图像编码:** 图像通过一个精心设计的、可能结合了对比学习和重建损失的“文本对齐视觉Tokenizer”（如VILA-U和UniTok中的UNIT-like Tokenizer），或带有共享映射的双编码器（Tokenflow）等，生成视觉Token。MUSE-VL则直接拼接语义和像素特征。
- **Token化与融合:** 这些特殊视觉Token与文本Token一起输入到LLM中
- **自回归生成:** LLM进行自回归预测，生成文本和视觉Token
- **图像解码:** 生成的视觉Token通过成对训练的解码器进行解码，

• **Strengths:**

- **平衡性:** 旨在平衡高层语义理解和低层像素细节，提高文本-图像对齐和全局一致性
- **鲁棒性:** 提供更鲁棒和富有表现力的视觉Token，尤其是在处理语义压缩的视觉信息时
- **多粒度:** 能够提取多粒度的图像特征，支持丰富的多模态生成

• **Weakness:**

- **Tokenizer复杂性:** 特殊tokenizer的设计和训练可能更加复杂，需要专门的损失函数和策略
- **计算开销:** 特殊tokenizer仍面临序列长度和计算开销的问题
- **生成质量:** 最终的视觉生成质量可能仍受限于AR模型的特性，难以达到扩散模型的SOTA水平



External Diffuser

这类模型的核心是自回归LLM，负责多模态理解和文本生成；图像生成则通过LLM输出的中间表示（通常是视觉Token或嵌入）来驱动一个**独立的**（外部）扩散模型。

•**Representative Works:** Emu, Emu2, DreamLLM, SEED-X, MetaMorph, Metaqueries, BLIP3-o

•**Pipeline:**

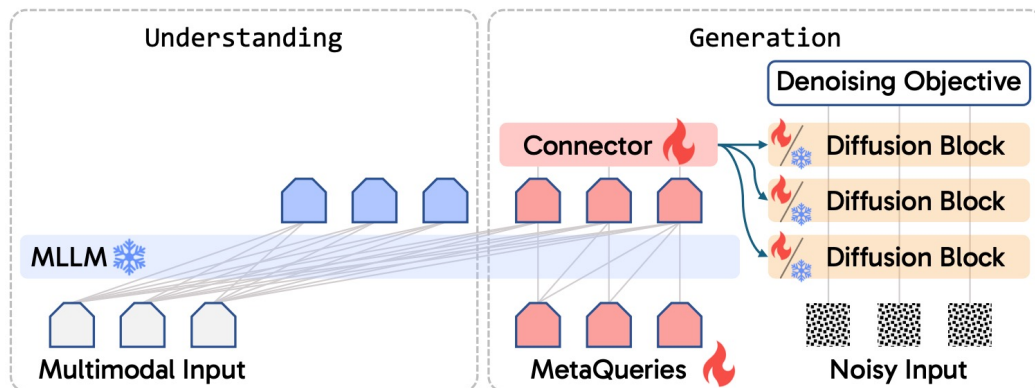
- **视觉编码:** 图像通过预训练的视觉编码器（如EVA-CLIP, OpenAI-CLIP, SigLIP）转换为视觉embedding或Token
- **LLM输入:** 视觉Token通过线性映射等与文本Token对齐，部分模型会加入可学query，作为输入送入自回归LLM
- **LLM推理/生成:** LLM进行多模态理解、推理，并生成文本回复和/或图像Token（通常是语义层面的视觉Token）
- **图像解码:** LLM生成的视觉Token或嵌入作为条件，驱动一个独立的扩散模型来合成高保真图像

•**Strengths:**

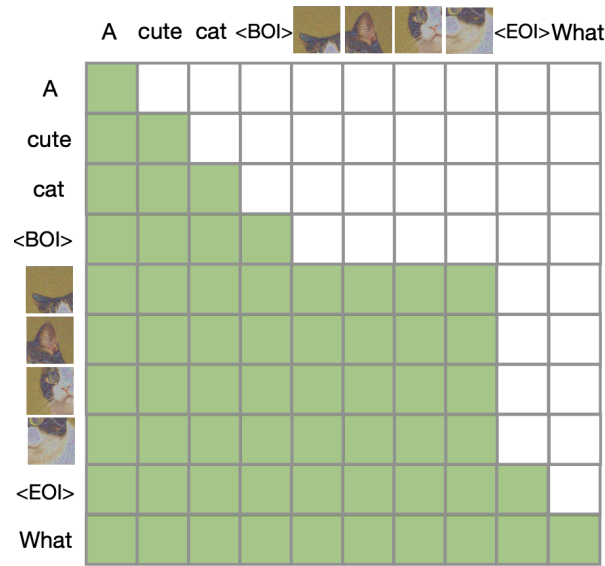
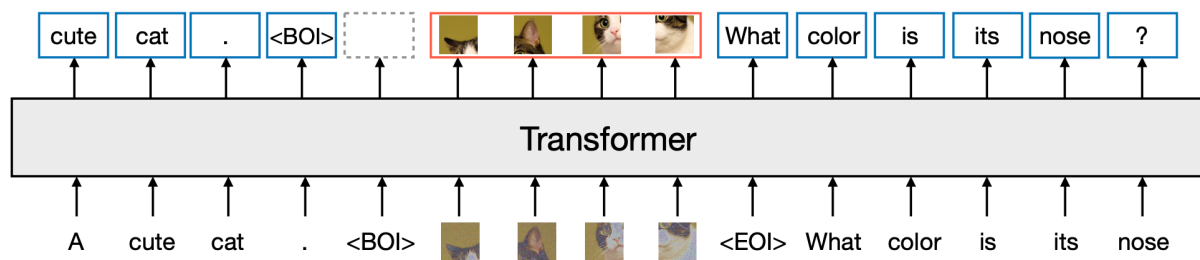
- **解耦优势:** 充分利用LLM在多模态理解、推理和文本生成方面的强大能力，同时利用扩散模型在图像合成方面的高质量 and
- **模块化与训练效率:** 允许使用冻结的LLM和视觉编码器，显著减少可训练参数，通常以最小的数据消耗实现快速收敛
- **竞争性性能:** 在已有的多模态生成和理解基准上可以取得有竞争力的性能
- **高质量生成:** 预训练扩散解码器能够将语义压缩的视觉Token逐步细化为高分辨率、逼真的图像

•**Weakness:**

- **信息瓶颈:** LLM上下文被压缩成相对少量的潜在Token，在理解和生成模块之间引入了**显式瓶颈**，可能导致大量信息丢失，尤其是在长上下文多模态推理中
- **细粒度控制受限:** LLM输出的视觉Token通常是语义层面的，缺乏像素级细节，可能导致在细粒度图像编辑或结构保持转换方面表现不佳（如GPT-4o编辑丢细节）
- **推理延迟:** 图像生成仍需扩散模型的迭代去噪过程



Integrated Transformer – Dense



这类模型在单个Transformer中，自回归和扩散的逻辑交织，共享核心参数，共同优化

•**Representative Works:** Transfusion, MonoFormer, Show-o

•**Pipeline:**

- **视觉编码:** 图像通过VAE编码为连续latent，或通过MAGVIT-v2 (Show-o) 编码为离散视觉Token。
- **统一Transformer:** 文本和视觉Token被输入到一个统一的Transformer主干网络中，为不同任务设计了特有的attention mask
- **联合训练:** 模型在训练时同时优化AR损失（文本生成）和Diffusion损失（图像重建）
- **生成过程:** 在生成阶段，对视觉latent迭代去噪，或自回归生成文本Token，整个过程通过双向注意力或跨模态上下文注入condition

•**Strengths:**

- **统一性与无瓶颈上下文:** 将AR和Diffusion的优点集成到单一Transformer架构中，关键是在所有Transformer块中保持无瓶颈上下文，从而实现生成和理解模块之间的无损交互，更易于扩展
- **平衡控制与保真度:** 有效平衡了LLM的序列生成能力和扩散模型的高视觉保真度
- **全局一致性:** 能够通过双向注意力或跨模态上下文实现更好的跨模态对齐和全局图像一致性

•**Weakness:**

- **高训练成本:** 相比External Diffuser方案，需要更高的训练计算量
- **高推理成本:** 由于图像生成涉及扩散模型的迭代去噪步骤，推理时间较长
- **训练复杂性:** 优化两种不同性质的损失（AR和Diffusion）在一个紧密耦合的架构中，可能导致训练不稳定或任务冲突

Integrated Transformer – MoT

这类模型为AR（理解/文本）和扩散（图像生成）方面使用**两套独立的参数**，但它们共享一个**共同的注意力机制**，处理来自两个路径的Token，从而实现丰富的跨模态交互和上下文共享

•**Representative Works:** LMFusion, JanusFlow, Mogao, BAGEL

•**Pipeline:**

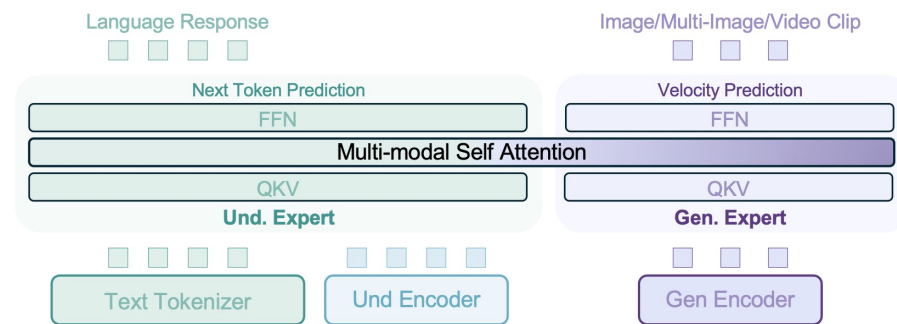
- **多模态输入与Token化:** 文本被Token化为离散Token，图像被编码为视觉Token（可能是语义latent、像素latent或混合组合）
- **共享Transformer主干与独立头部/分支:**
 - **注意力机制:** 注意力层同时处理来自AR路径和扩散路径的Token。这使得AR部分能够“看到”图像生成上下文，反之亦然
 - **参数集:** 独立的参数分别负责AR任务（文本/理解的下一个Token预测）和扩散任务（图像生成的噪声预测）
- **AR路径:** 对于理解和文本生成，Transformer执行下一个Token预测，自回归地消费文本和视觉Token
- **Diffusion路径:** 对于图像生成，Transformer在不同时间步长预测图像latent的噪声，条件注入来自AR路径的上下文
- **图像解码:** 一个独立的解码器将去噪后的latent转换为像素

•**Strengths:**

- **深度集成与丰富交互:** 通过共享注意力，实现了AR和Diffusion机制在单个Transformer中的深度融合，允许丰富的跨模态上下文共享和无瓶颈交互
- **任务多功能性:** 共享注意力使得AR部分能从视觉生成过程中获取信息，反之亦然，可能导致理解和生成任务的输出更连贯、上下文更相关
- **灵活参数化:** 独立的参数集允许AR和扩散任务的专业化，同时避免完全解耦，可能在两者上都取得更好的性能
- **统一学习:** 训练受益于AR和Diffusion损失，有助于更全面的理解和生成能力

•**Weakness:**

- **训练挑战:** 同时优化AR和Diffusion目标，可能复杂且易于出现模式崩溃或训练不稳定
- **计算成本:** 尽管共享注意力，但不共享参数和扩散的迭代性质仍导致高计算和内存需求



Others – Diffusion

这类模型基于扩散过程进行多模态生成，通过调整扩散过程的条件机制来实现理解和生成任务

• **Representative Works:** DualDiffusion, MMaDA

• **Pipeline:**

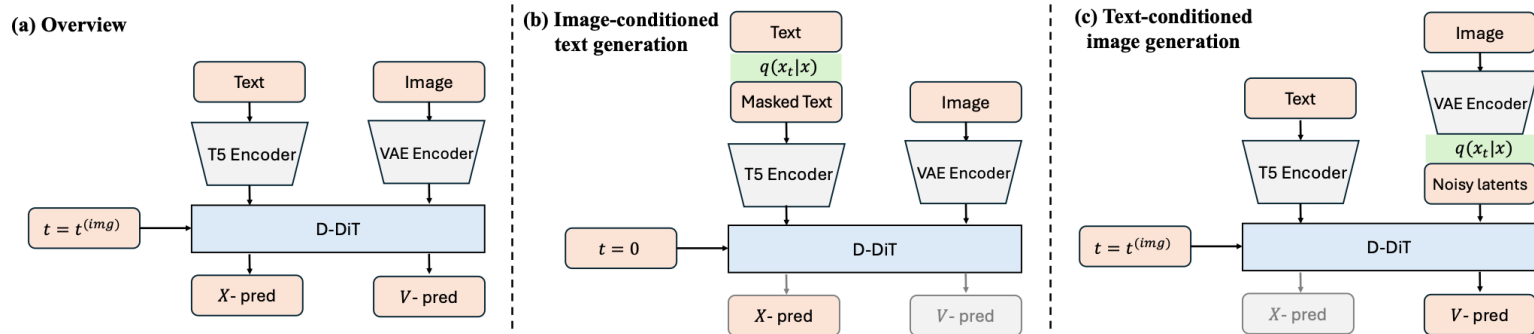
- **多模态输入:** 接受文本描述、图像或其他联合嵌入作为条件
- **独立噪声化:** 文本（离散latent）和图像（连续latent）分别通过独立的扩散前向过程被逐步噪声化
- **联合去噪:** 在反向去噪过程中，模型使用模态特定的去噪器，这些去噪器会通过跨模态条件注入（如文本latent关注图像latent，反之亦然）协同工作
- **解码重建:** 去噪后的latent通过相应的解码器（如文本解码器和VAE解码器）重建为原始模态内容

• **Strengths:**

- **高质量生成:** 在图像生成方面达到当前SOTA水平，生成图像质量高，模式覆盖范围广
- **强大的可控性:** 能够灵活地整合各种条件信号（如文本、图像、联合latent），实现精细的生成控制
- **概念学习:** 通过多模态条件注入，模型能学习跨模态关联，支持更深的语义理解

• **Weakness:**

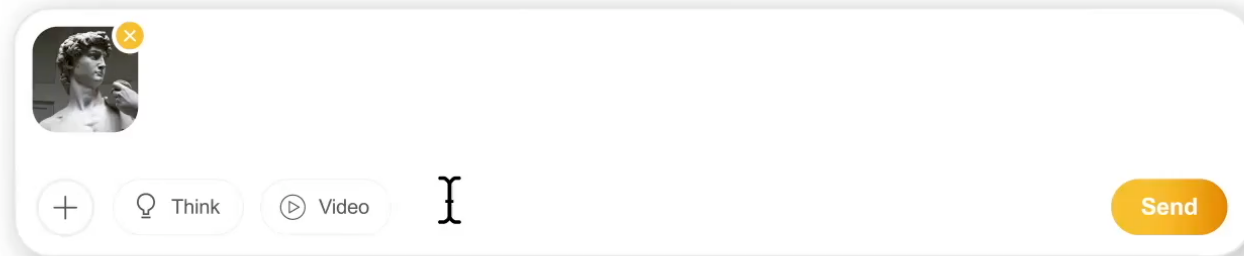
- **计算效率:** 迭代去噪过程导致生成速度慢
- **理解能力:** 理解能力相比纯理解模型仍有差距



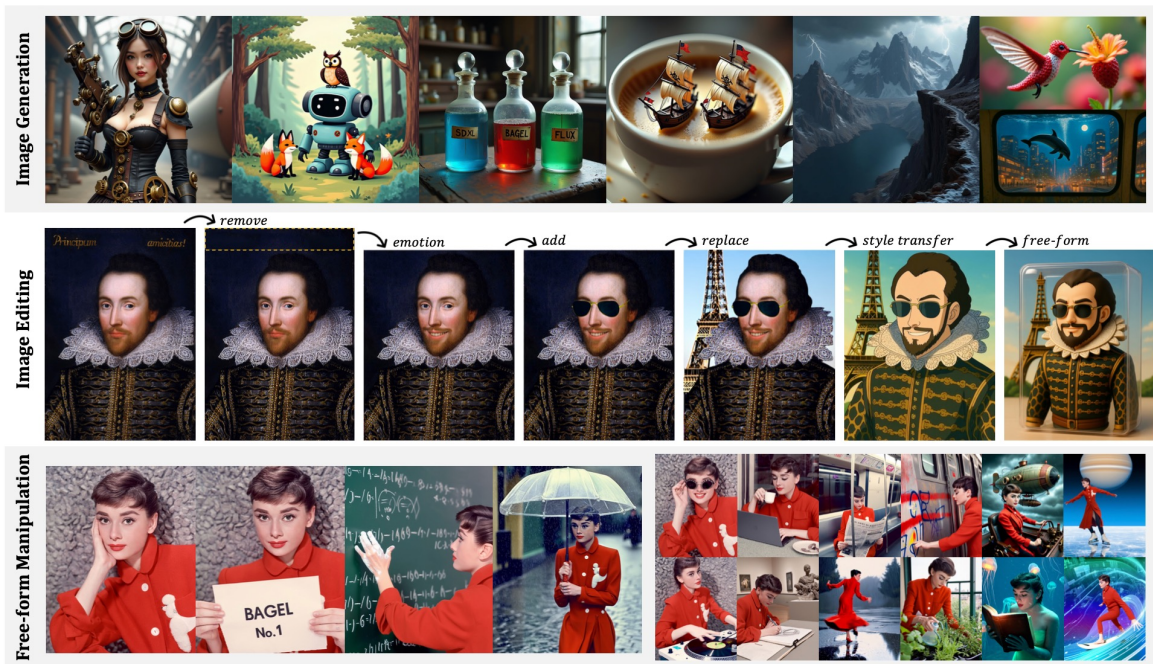
BAGEL – Model

BAGEL

The Open-Source Unified Multimodal Model



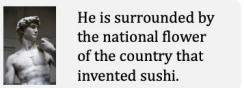
BAGEL – Model



Generation & Editing w/ Thinking

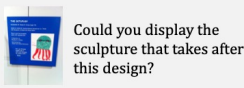
A car made of small cars.

*<think> The model should generate an image where numerous small cars are combined to form the shape and structure of a larger car.
The refined prompt is: A highly detailed and creative image showcasing a large car composed entirely of numerous small cars, meticulously arranged to replicate the shape and structure of a full-sized vehicle. The small cars are tightly interlocked, creating a seamless and realistic exterior with accurate proportions and details. The setting is neutral and clean, with soft, diffused lighting to emphasize the intricacy of the composition without any distractions...
</think>*



He is surrounded by the national flower of the country that invented sushi.

*<think> The user wants to replace the statue's surroundings with a scene full of cherry blossoms, while keeping the main body of the statue unchanged. First, I have to make sure the details of the statue, such as muscle lines, facial expressions, and clothing folds, remain clear and complete. Next, cherry blossom elements need to be added around the statue to make it conform to the setting of 'the national flower of the country that invented sushi'. The cherry blossoms...
</think>*

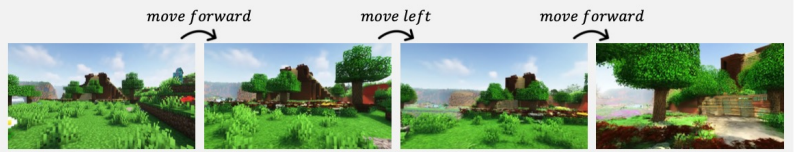


Could you display the sculpture that takes after this design?

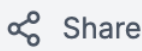
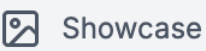
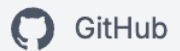
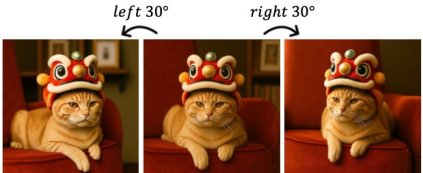
*<think> The question_image shows a design of an octopus created by a child named Blake W. Torney, aged 10, titled "The Octopus." The description mentions that the sculpture was made in 2008 by Dante Marioni, an American glass artist born in 1964, using blown and hot-sculpted glass with applied bits. The sculpture is part...
The question_text asks to display the sculpture that takes after this design. Based on the description, the final answer image should feature the actual glass sculpture created by Dante Marioni, which reflects the design of the child's drawing. The sculpture should have a green body with red tentacles, similar to the child's depiction...
</think>*



Navigation



Rotation



BAGEL – Model

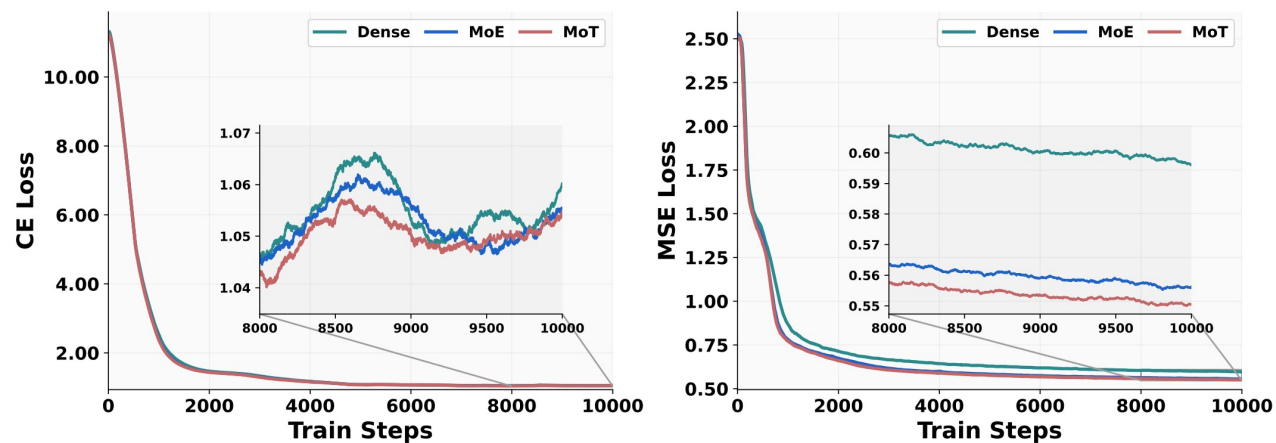
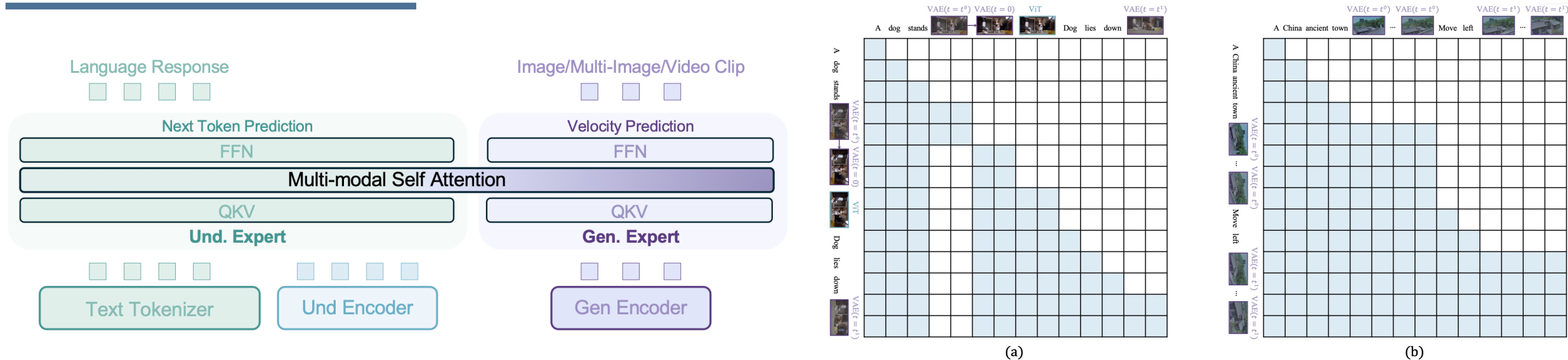
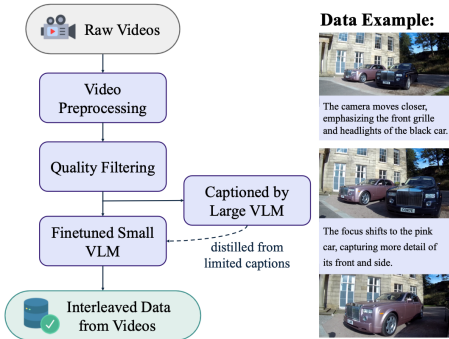


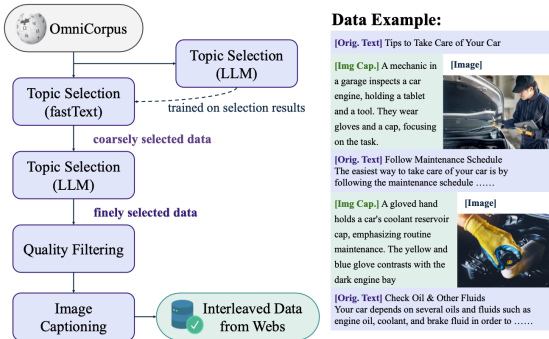
Figure 3 Loss curves of various designs. CE loss and MSE loss are computed on multimodal understanding and generation tasks, respectively. Ablation experiments are carried out on a 1.5B LLM. The sampling ratio for generation and understanding data is set at 4:1.

BAGEL – Data

Data Source	# Data (M)	# Tokens (T)
Text Data	400	0.4
Image-Text-Pair Understanding Data	500	0.5
Image-Text-Pair Generation Data	1600	2.6
Interleaved Understanding Data	100	0.5
Interleaved Generation Data: Video	45	0.7
Interleaved Generation Data: Web	20	0.4



(a) Data pipeline for interleaved data from videos.



(b) Data pipeline for interleaved data from webs.

- 视频-文本交错数据**：来自公开视频库（如 YouTube 教育视频、科普短片）+ 开源数据集 Koala36M（含交互场景）、MVImgNet2.0（多视角物体数据）
- 网页 - 文本交错数据**：来自OmniCorpus 网页数据集（含教程、百科、设计文档）+ 结构化编辑数据集（如 OmniEdit、UltraEdit）
- 以及推理以增强数据**：包括文生图、自由图像操控和智能编辑，使用开源VLM/LLM辅助构建推理过程

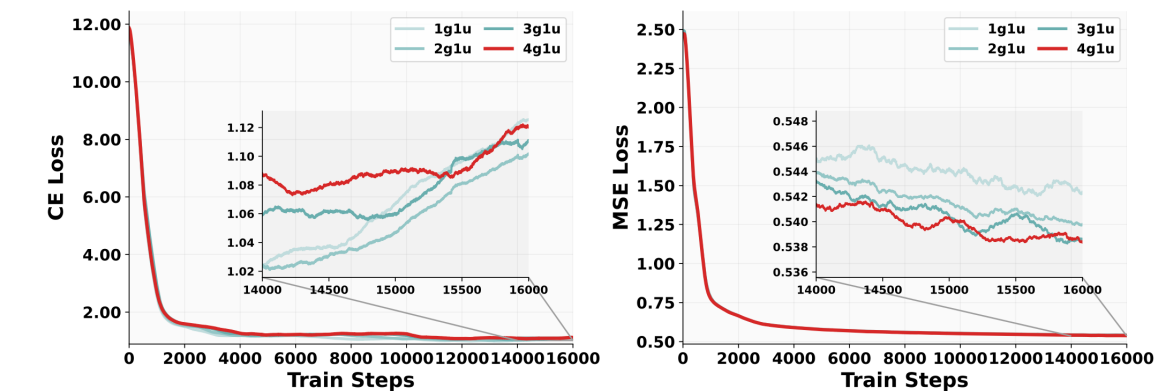


Figure 5 Loss curves of different data ratios. Ablation experiments are carried out on a 1.5B LLM. "1g1u" means that the sampling ratio for generation and understanding data is set at 1:1.

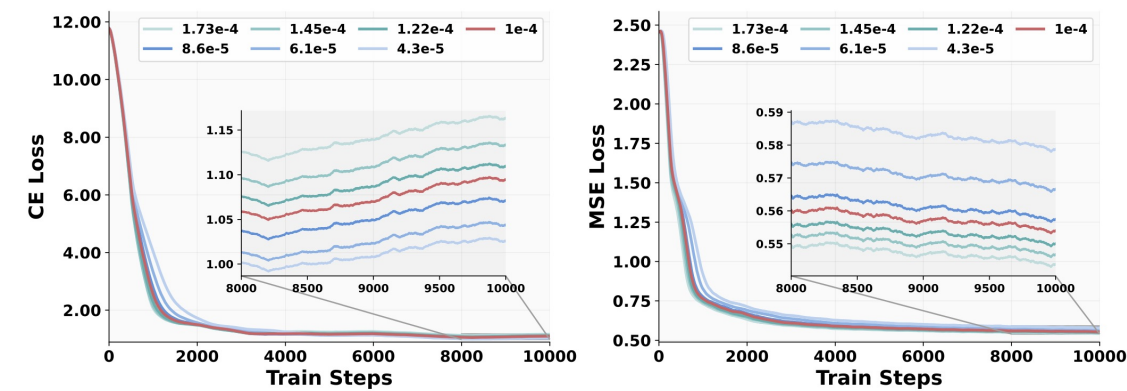


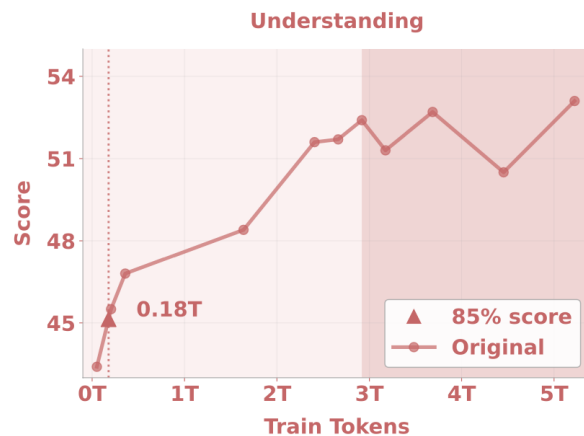
Figure 6 Loss curves of different learning rates. Ablation experiments are carried out on a 1.5B LLM. The sampling ratio for generation and understanding data is set at 1:1.

BAGEL – Training

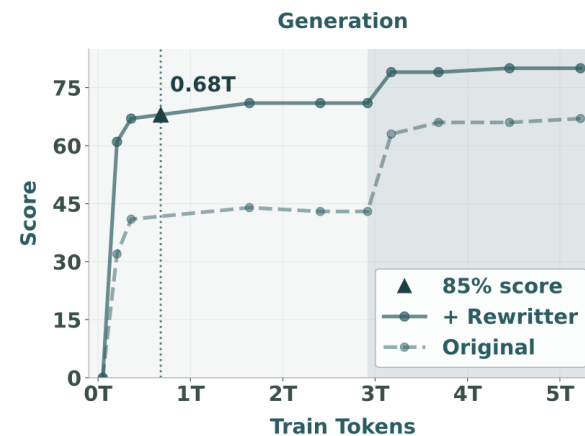
	Alignment	PT	CT	SFT
Hyperparameters				
Learning rate	1×10^{-3}	1.0×10^{-4}	1.0×10^{-4}	2.5×10^{-5}
LR scheduler	Cosine	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1.0 \times 10^{-15}$)			
Loss weight (CE : MSE)	-	0.25 : 1	0.25 : 1	0.25 : 1
Warm-up steps	250	2500	2500	500
Training steps	5K	200K	100k	15K
EMA ratio	-	0.9999	0.9999	0.995
Sequence length per rank (min, max)	(32K, 36K)	(32K, 36K)	(40K, 45K)	(40K, 45K)
# Training seen tokens	4.9B	2.5T	2.6T	72.7B
Max context window	16K	16k	40k	40k
Gen resolution (min short side, max long side)	-	(256, 512)	(512, 1024)	(512, 1024)
Und resolution (min short side, max long side)	(378, 378)	(224, 980)	(378, 980)	(378, 980)
Diffusion timestep shift	-	1.0	4.0	4.0
Data sampling ratio				
Text	0.0	0.05	0.05	0.05
Image-Text pair (T2I)	0.0	0.6	0.4	0.3
Image-Text pair (I2T)	1.0	0.1	0.1	0.05
Interleaved understanding	0.0	0.1	0.15	0.2
Interleaved generation: video	0.0	0.1	0.15	0.2
Interleaved generation: web	0.0	0.05	0.15	0.2

Table 3 Training recipe of BAGEL. Multimodal interleaved data is highlight in gray .

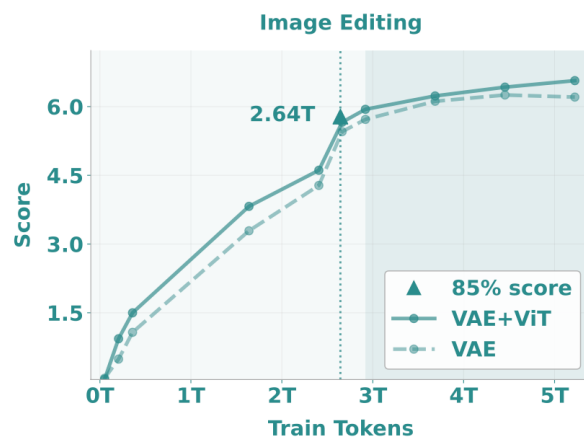
BAGEL – Emerging



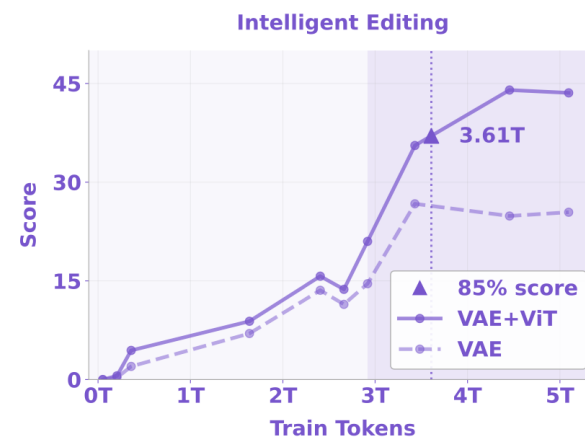
(a) Average score on Image Understanding tasks.



(b) GenEval score on Image Generation task.



(c) GEdit Overall Score on classical Image Editing task.



(d) IntelligentBench Score on Intelligent Editing task.

BAGEL – Emerging

Prompts

A clear image of a blackboard with a clean, dark green surface and the word 'Hello BAGEL' written precisely and legibly in the center with bold, white chalk letters.

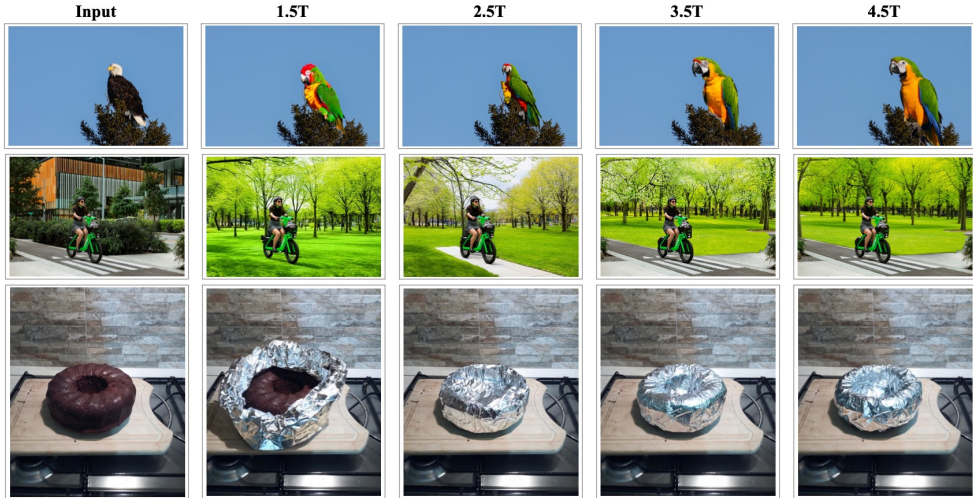
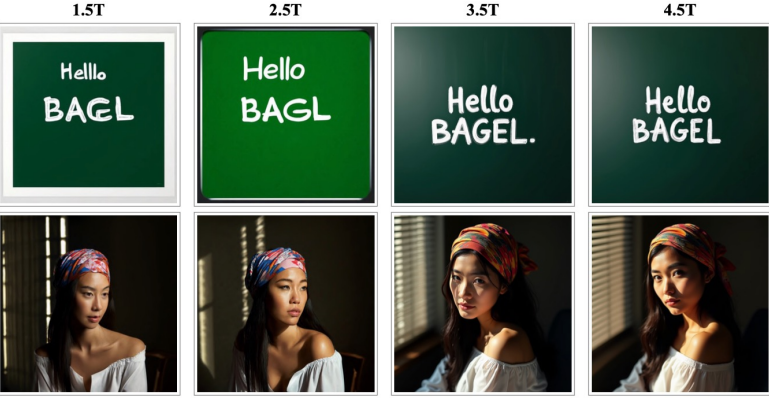
Portrait photo portrait of an Asian woman with long dark hair wearing a colorful headscarf tied around her forehead, seated against a dimly lit backdrop casting dramatic shadows across her face from window blinds or slatted light source behind her left shoulder. She wears off-the-shoulder white blouse and gazes intently towards camera creating strong eye contact. The lighting is high contrast highlighting facial features while leaving parts shadowed adding depth and moodiness. Photorealistic quality captures fine textures like fabric folds and skin tones vivid yet natural color palette emphasizes warm earthy hues contrasting deep shadows.

Prompts

Replace the eagle with a parrot.

Change the background to a spring park while keeping the person unchanged.

Transform the donut's material into aluminum foil.



Questions

Could you display what this knitting project looks like completed?

What is the appearance of the location under night lighting?

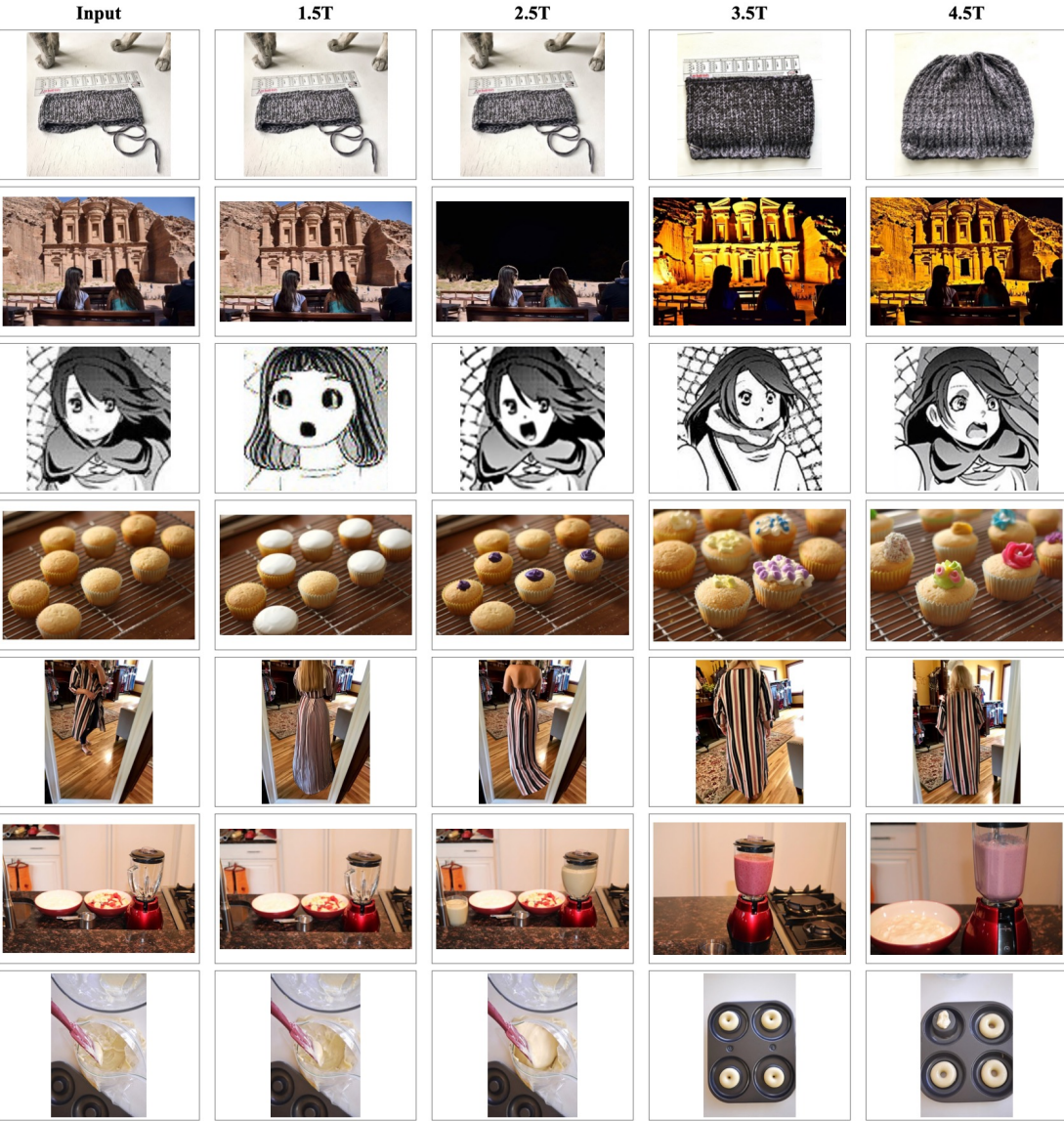
Can you share an image of this character looking surprised?

Could you put some toppings on these cupcakes for me?

Could you display the rear of this gown?

Could you display the smoothie once it's blended?

What method helps in adding this batter to donut molds?



BAGEL – Benchmark

Type	Model	# LLM Params	MME-P [†]	MME-S [†]	MMBench [†]	MMMU [†]	MM-Vet [†]	MathVista [†]	MMVP [†]
<i>Und. Only</i>	InternVL2 [13]	1.8B	1440	1877	73.2	34.3	44.6	46.4	35.3
	InternVL2.5 [12]	1.8B	-	2138	74.7	43.6	60.8	51.3	-
	Qwen2-VL[76]	1.5B	-	1872	74.9	41.1	49.5	43.0	-
	Qwen2.5-VL[4]	3B	-	2157	79.1	53.1	61.8	62.3	-
	BLIP-3 [89]	4B	-	-	76.8	41.1	-	39.6	-
	LLava-OV [37]	7B	1580	-	80.8	48.8	57.5	63.2	-
	InternVL2 [13]	7B	1648	2210	81.7	49.3	54.2	58.3	51.3
	InternVL2.5 [12]	7B	-	2344	<u>84.6</u>	56.0	62.8	64.4	-
	Qwen2-VL [76]	7B	-	2327	83.0	54.1	62.0	58.2	-
	Qwen2.5-VL[4]	7B	-	<u>2347</u>	83.5	58.6	<u>67.1</u>	68.2	-
	Emu3-Chat** [78]	8B	1244	-	58.5	31.6	37.2	-	36.6
	Kimi-VL [70]	2.8B/16B	-	-	-	<u>57.0</u>	66.7	<u>68.7</u>	-
	DeepSeek-VL2 [85]	4.1B/28B	-	-	-	51.1	60.0	62.8	-
<i>Unified</i>	Show-o ₅₁₂ [87]	1.3B	1097	-	-	26.7	-	-	-
	Janus [82]	1.5B	1338	-	69.4	30.5	34.3	-	-
	Janus-Pro [11]	1.5B	1444	-	75.5	36.3	39.8	-	-
	BAGEL	1.5B MoT	1610	2183	79.2	43.2	48.2	63.4	<u>54.7</u>
	ILLUME [75]	7B	1445	-	75.1	38.2	37.0	-	-
	VILA-U ₂₅₆ ** [84]	7B	1336	-	66.6	32.2	27.7	-	22.0
	Chameleon** [69]	7B	-	-	35.7	28.4	8.3	-	0.0
	Janus-Pro [11]	7B	1567	-	79.2	41.0	50.0	-	-
	MetaQuery-XL [†] [56]	7B	<u>1685</u>	-	83.5	58.6	66.6	-	-
	LlamaFusion** [65]	8B	1604	-	72.1	41.7	-	-	-
	MetaMorph [72]	8B	-	-	75.2	41.8	-	-	48.3
	SEED-X [23]	13B	1457	-	70.1	35.6	43.0	-	-
	TokenFlow-XL [58]	13B	1546	-	68.9	38.7	40.7	-	-
	MUSE-VL [88]	32B	-	-	81.8	50.1	-	55.9	-
	BAGEL	7B MoT	1687	2388	85.0	55.3	67.2	73.1	69.3

Table 4 Comparison with state-of-the-arts on visual understanding benchmarks. MME-S refers to the summarization of MME-P and MME-C. For MoE models, we report their activate params / total params. [†]: MetaQuery [56] adopts pre-trained model from Qwen2.5-VL [4] and freezes it during training. **: Partial results are from by MetaMorph [72] or MetaQuery [56].







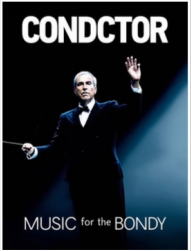





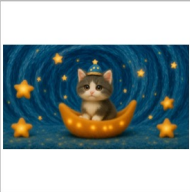

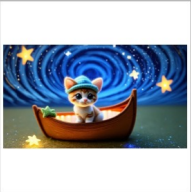









BAGEL – Benchmark

Type	Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	PixArt- α [9]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [60]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [59]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [78]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [57]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [5]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [19]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	FLUX.1-dev† [35]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified	Chameleon [69]	-	-	-	-	-	-	0.39
	LWM [41]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X [23]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	TokenFlow-XL [58]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	ILLUME [75]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	Janus [82]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [101]	-	-	-	-	-	-	0.63
	Emu3-Gen† [78]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
	Show-o [87]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Janus-Pro-7B [11]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	MetaQuery-XL† [56]	-	-	-	-	-	-	0.80
	BAGEL	0.99	0.94	0.81	0.88	0.64	0.63	0.82
	BAGEL†	0.98	0.95	0.84	0.95	0.78	0.77	0.88

Table 5 Evaluation of text-to-image generation ability on GenEval benchmark. ‘Gen. Only’ stands for an image generation model, and ‘Unified’ denotes a model that has both understanding and generation capabilities. † refer to the methods using LLM rewriter.

Type	Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Gen. Only	SDv1.5 [60]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
	SDXL [57]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
	SD3.5-large [19]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
	PixArt-Alpha [9]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
	playground-v2.5 [38]	0.49	0.58	0.55	0.43	0.48	0.33	0.49
	FLUX.1-dev [35]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified	Janus [82]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
	VILA-U [84]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
	Show-o-512 [87]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
	Janus-Pro-7B [11]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
	Emu3 [78]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
	MetaQuery-XL [56]	0.56	0.55	0.62	0.49	0.63	0.41	0.65
	GPT-4o**	0.81	0.71	0.89	0.83	0.79	0.74	0.80
	BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
	BAGEL <i>w/</i> Self-CoT	0.76	0.69	0.75	0.65	0.75	0.58	0.70

Table 6 Comparison of world knowledge reasoning on WISE. WISE examines the complex semantic understanding and world knowledge for T2I generation. ‘Gen. Only’ stands for an image generation model, and ‘Unified’ denotes a model that has both understanding and generation capabilities. **: Results of GPT-4o are tested by [90].

Prompts	BAGEL	Janus-Pro	SD3-medium	GPT-4o
Book cover, A surreal double exposure portrait that blends a woman’s face with a beautiful seascape. The overall mood is dreamy and mystical, with rich colors and intricate details. 1:1				
A movie poster for a film titled "CONDUCTOR" The poster features a person in a dark suit, holding a conductor’s baton, with their left hand raised in a gesture that suggests they are leading or guiding. The background is dark and somewhat abstract, with a hint of a stage or performance setting. The title "CONDUCTOR" is prominently displayed at the top in bold, white capital letters. Below the title, the subtitle "Music for the body" is written in a smaller, white font. The overall design is sleek and professional, with a focus on the conductor’s role and the theme of music and performance 4:3				
Photorealistic closeup image of two pirate ships battling each other as they sail inside a cup of coffee. 1:1				
微缩景观，毛茸茸羊毛毡，超级特写，浅景深，梵高风格的星空下，一只小猫眯坐在一艘发光的小船上，船的周围漂浮着毛茸茸的羊毛毡星星，猫咪头戴戴着星星光点的小帽子，背景是旋转的漩涡星空，散发着梦幻的蓝光光芒，生物发光，细节丰富，3D立体。 9:16				
A female cosplayer portraying an ethereal fairy or elf, wearing a flowing dress made of delicate fabrics in soft, mystical colors like emerald green and silver. She has pointed ears, a gentle, enchanting expression, and her outfit is adorned with sparkling jewels and intricate patterns. The background is a magical forest with glowing plants, mystical creatures, and a serene atmosphere. 1:1				
On Mars, a rugged landscape of reddish-brown soil and jagged rocks stretches under a pale pink sky. A towering volcano looms in the distance, its peak shrouded in a faint plume of smoke. Nearby, a deep canyon with intricate, erosion-carved walls cuts through the terrain. A small robotic rover moves slowly across the surface, leaving faint tracks in the fine Martian dust. The scene captures the stark beauty and otherworldly atmosphere of the Red Planet. 1:1				




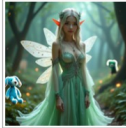
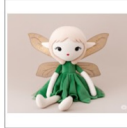


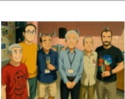
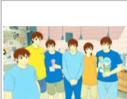


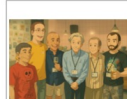










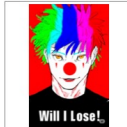
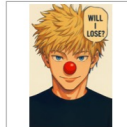
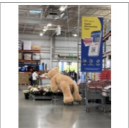
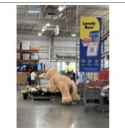
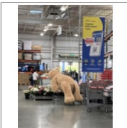

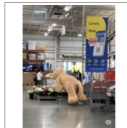
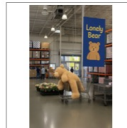


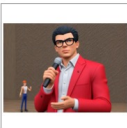
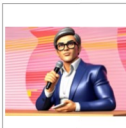




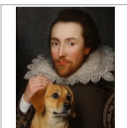



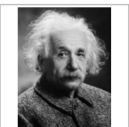
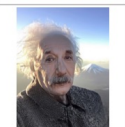
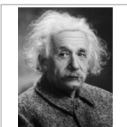

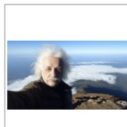
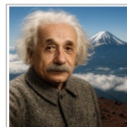




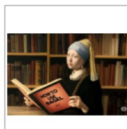
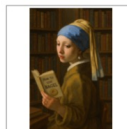
BAGEL – Benchmark

Type	Model	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
		G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Private	Gemini 2.0 [24]	6.73	6.61	6.32	5.43	6.78	5.36
	GPT-4o [54]	7.85	7.62	7.53	7.67	7.56	7.30
Open-source	Instruct-Pix2Pix [6]	3.58	5.49	3.68	-	-	-
	MagicBrush [97]	4.68	5.66	4.52	-	-	-
	AnyEdit [93]	3.18	5.82	3.21	-	-	-
	OmniGen [86]	5.96	5.89	5.06	-	-	-
	Step1X-Edit [42]	7.09	6.76	6.70	7.20	6.87	6.86
	BAGEL	<u>7.36</u>	<u>6.83</u>	6.52	<u>7.34</u>	6.85	6.50







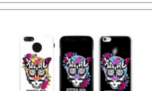


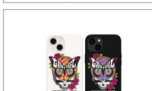

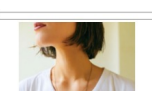

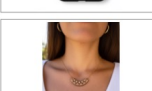
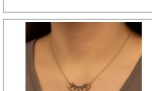
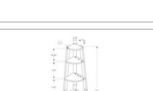
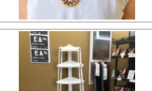
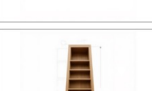

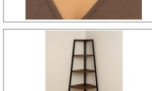

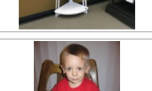
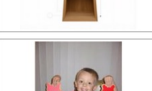
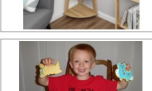
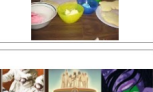

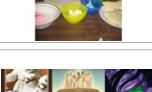
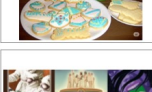

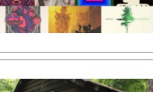

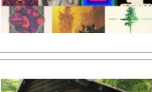
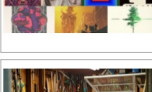
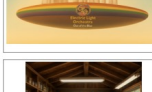

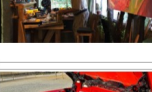


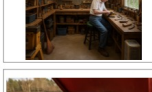

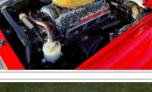
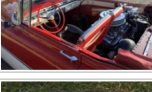


Table 7 Comparison on GEdit-Bench. All metrics are reported as higher-is-better (↑). G_SC, G_PQ, and G_O refer to the metrics evaluated by GPT-4.1.

Type	Model	Score↑
Private	GPT-4o** [54]	78.9
	Gemini 2.0** [24]	57.6
Open-source	Step1X-Edit [42]	14.9
	BAGEL	44.9
	BAGEL w/ Self-CoT	55.3

Table 8 Comparison on IntelligentBench. IntelligentBench examines complex reasoning ability in an image-editing context. **: Results are reported only on the subset of cases answered (some responses were rejected). GPT-4o answered 318 of 350 questions, while Gemini 2.0 answered 349 questions.

Questions	Input	BAGEL	IC-Edit	Step1X-Edit	Gemini 2.0	GPT-4o
Make her a Jellycat plush toy.						
Change to Ghibli style.						
Remove passers-by in the background						
Convert the image to colorful one. Give him a clown nose and change the text to English 'Will I Lose'.						
Change the text in the picture to 'Lonely Bear'.						
Generate a 3D action figure model based on this photo of a handsome man, boldly restoring his muscular physique and dashing posture, with details showcasing dynamic strength. Place the action figure on a wooden table.						
He squatted down and touched a dog's head.						
He took a selfie on the top of Mount Fuji.						
She is reading a book in the library. The cover of the book is 'How to use BAGEL'.						

BAGEL — Showcase

Questions	Input	BAGEL	Step1X-Edit	Gemini 2.0	GPT-4o
Could you provide an image of this car's interior?					
Are there phone cases available that match the shirt designs? Could you display some?					
Could you provide a visual of the necklace being worn?					
Could I see a real-life example of this corner shelf concept?					
How do the cookies appear after decoration?					Rejected
Describe the image in the top middle.					
Could you display the inside of this workshop?					
Can you show me the engine of this car?					
Could you display how this fabric appears unrolled?					

Prompts	Results	Results w/ Thinking	
A car made of small cars.		<p><think> The model should generate an image where numerous small cars are combined to form the shape and structure of a larger car. The refined prompt is: A highly detailed and creative image showcasing a large car composed entirely of numerous small cars, meticulously arranged to replicate the shape and structure of a full-sized vehicle. The small cars are tightly interlocked, creating a seamless and realistic exterior with accurate proportions and details. The setting is neutral and clean, with soft, diffused lighting to emphasize the intricacy of the composition without any distractions. The image captures the concept in a photorealistic style, with a focus on precision and craftsmanship, highlighting the innovative use of scale and form.</p> <p></think></p> 	
A cat is twice as large as the dog next to it.		<p><think> The model should generate an image where the cat appears significantly larger than the dog to emphasize the difference in size. The refined prompt is: A highly detailed and realistic image of a large, majestic cat standing next to a smaller dog, with the cat's size being visually emphasized by its towering presence compared to the dog. The cat should have sleek fur, vibrant colors, and a confident, regal posture, while the dog appears smaller, with a friendly and slightly curious expression. The setting is an outdoor environment with soft, natural lighting, warm tones, and a gentle atmosphere. The perspective is slightly low-angle, capturing the contrast in size between the two animals in a visually striking composition, with sharp focus on the cat and a subtle depth of field for the background.</p> <p></think></p> 	
A man is standing beside a female brown bear and its cub. Generate an image to show what will likely happen.		<p><think> The model should generate an image of a man in close proximity to the female brown bear and its cub, potentially showing a tense interaction or cautious behavior from the man, as bears are generally protective of their cubs. The refined prompt is: A man stands close to a female brown bear and its cub, depicted in a tense yet natural interaction, with the bear appearing protective and the cub nestled nearby. The scene captures a moment of cautious distance from the man, who exhibits concern or wariness, his posture slightly hunched and alert. The setting is a rugged, forested area with dense greenery and earthy tones, illuminated by soft, natural light filtering through the trees. The mood is tense yet realistic, emphasizing the raw connection between humans and wildlife, with detailed textures on the bear's fur, the man's clothing, and the surrounding foliage. Ultra-realistic, with a focus on lifelike anatomy and environment, capturing the emotional and environmental dynamics.</p> <p></think></p> 	
生成一幅真实水果与微型行星（土星、火星、地球）混合而成的果盘照片		<p><think> 首先，用户需要一张真实水果与微型行星（土星、火星、地球）混合而成的果盘照片。这意味着需要将三个行星的模型与常见的水果进行融合，形成一个视觉上真实但又富有创意的场景。接下来，我需要考虑如何将行星与水果相结合。土星、火星和地球三个行星都具有独特的外观特征，但它们的大小和比例与水果相差悬殊。例如，土星的体积远超过一般水果，而火星和地球相对较小。因此，需要将这些行星以比例协调的方式放置在水果盘中，使它们看起来像是水果的一部分，而不是突兀的物体。水果的颜色和形状也需要与行星的颜色相匹配。例如，和土星可以与较大的水果（如苹果、橙子）结合，以体现它们的体积。火星和地球可以与较小的水果（如草莓、蓝莓）结合，以保持比例的平衡。此外，还需要考虑光线和阴影，以使行星与水果之间的融合更加自然。光线应从上方或侧面照射，以强调水果的表面纹理和行星的细节。最后，整体构图应具有视觉吸引力，水果和行星的排列要错落有致，避免过于杂乱。最终生成的图像应该是水果与微型行星完美融合的场景，既保持了水果的真实感，又巧妙地展现了微型行星的独特魅力。</p> <p></think></p> 	
(a) Thinking Helps Generation: Text-to-Image Generation Cases			
Questions	Input	Results	Results w/ Thinking
Could you transfer the bag from the truck to the SUV?			<p><think> The question involves transferring a bag from a truck to an SUV. The question image shows a truck bed with a large, olive-green bag labeled "keeping calves bags cool with advanced cooling." The bag appears to be designed for storage or transport, likely for items that require temperature control. The question text asks for the bag to be moved to an SUV, suggesting the user wants to visualize the bag in a different context or vehicle. To address this, I would consider the size and shape of the bag, ensuring it fits well in the SUV's cargo area. The SUV in the answer image appears to be a larger vehicle with a spacious cargo area, which should accommodate the bag comfortably. The image would likely depict the bag placed neatly inside the SUV, possibly on the floor or secured in a way that prevents it from shifting during transport. In forming the answer text, I would confirm the transfer and describe the bag's placement in the SUV. The answer image would show the SUV's interior, with the bag visible and properly situated, emphasizing its functionality and utility. This approach ensures the user's request is met clearly and visually.</p> <p></think></p> 
Could you show me the cabinet with its drawers open so I can see inside?			<p><think> The question text asks for a visualization of the cabinet with its drawers open, allowing the interior to be visible. The question image shows a closed three-drawer cabinet with ornate handles and a metallic finish, placed on a wooden floor. The user likely wants to assess the cabinet's storage capacity or the design of its interior. To address this, I would ensure the answer image depicts the cabinet with all three drawers pulled out, revealing their interior. The drawers should be open at a standard angle, showing the inside surfaces, which might include lining, compartments, or empty space. The lighting should remain consistent with the original image to maintain continuity, and the cabinet should remain in the same position for clarity. The answer text would confirm that the cabinet is now displayed with its drawers open, fulfilling the user's request for a visual inspection of the inside. This approach ensures the user can evaluate the cabinet's functionality and design.</p> <p></think></p> 
Could you illustrate how to adorn this shelf with decorative pieces?			<p><think> The question involves removing the tape from the surface of a guitar body that is in the process of being crafted or repaired. The question image shows a guitar body with yellow masking tape applied along the edges, likely to protect the wood during a specific process, such as sanding or painting. The user is asking for the tape to be removed, implying they want to see the surface underneath. To address this, I would focus on ensuring the tape is completely peeled off, revealing the smooth, unobstructed surface of the guitar body. The final image would need to display the guitar body without any tape residue, highlighting its natural wood grain and finish. The workspace in the background, including tools and materials, should remain consistent to maintain context. When describing the answer image, I would note the absence of tape, the clean and polished surface of the guitar body, and any visible signs of craftsmanship, such as the wood grain pattern and the smooth curves of the body. This would fulfill the user's request for a clear view of the guitar body without the tape.</p> <p></think></p> 
(b) Thinking Helps Generation: Image Editing Cases			

BAGEL – Showcase

World Navigation

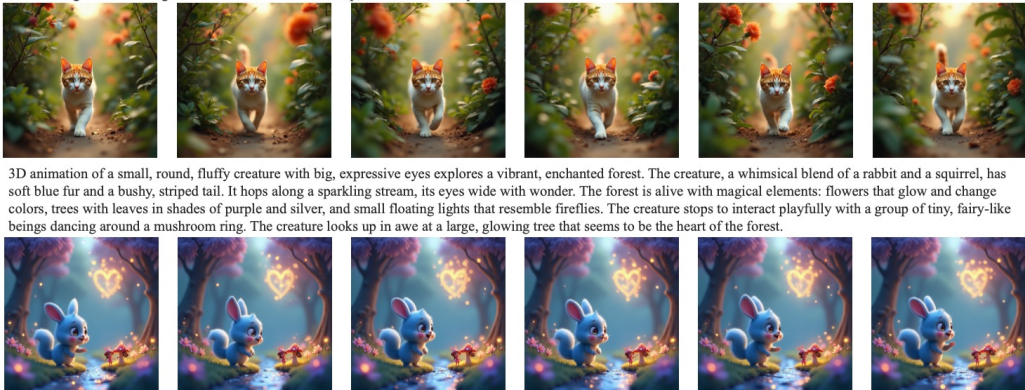


Rotation



A white and orange tabby cat is seen darting through a dense garden, as if chasing something. Its eyes are wide and happy as it jogs forward, scanning the branches, flowers, and leaves as it walks. The path is narrow as it makes its way between all the plants. The scene is captured from a ground-level angle, following the cat closely, giving a low and intimate perspective. The image is cinematic with warm tones and a grainy texture. The scattered daylight between the leaves and plants above creates a warm contrast, accentuating the cat's orange fur. The shot is clear and sharp, with a shallow depth of field.

Multi-Image Generation



3D animation of a small, round, fluffy creature with big, expressive eyes explores a vibrant, enchanted forest. The creature, a whimsical blend of a rabbit and a squirrel, has soft blue fur and a bushy, striped tail. It hops along a sparkling stream, its eyes wide with wonder. The forest is alive with magical elements: flowers that glow and change colors, trees with leaves in shades of purple and silver, and small floating lights that resemble fireflies. The creature stops to interact playfully with a group of tiny, fairy-like beings dancing around a mushroom ring. The creature looks up in awe at a large, glowing tree that seems to be the heart of the forest.

Prompts

A young woman with long, flowing hair, wearing a vintage sundress, standing in a field of wildflowers. She has a gentle smile, and the sunlight creates a soft, ethereal glow around her. The image is styled in a watercolor painting format, with delicate brushstrokes and pastel colors.

Janus-Pro-7B



BAGEL-1.5B



BAGEL-7B



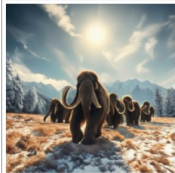
A wooden desk by the window. Its surface is smooth, with a stack of books on one side and a lamp with a soft - glowing bulb on the other. Beside the lamp, there's a half - filled coffee mug and a pen. A small potted plant adds a bit of greenery.



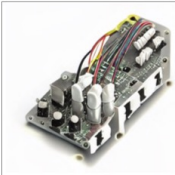
The image is a book cover. It features a woman standing in an urban setting, likely in London, given the theme of the book. The woman is wearing a red top and a green skirt, and she is holding a pink shopping bag, suggesting a relatable, modern-day scenario. In the background, there are other people, including children, which adds to the everyday city life theme. The title of the book, "Made in London," is prominently displayed in a large, elegant, cursive font. The author's name, "CLARE LYDON," is placed above the title. The overall design is vibrant and eye-catching, with a mix of warm and cool tones to create a sense of contrast and appeal.



The image is a book cover. It features a woman standing in an urban setting, likely in London, given the theme of the book. The woman is wearing a red top and a green skirt, and she is holding a pink shopping bag, suggesting a relatable, modern-day scenario. In the background, there are other people, including children, which adds to the everyday city life theme. The title of the book, "Made in London," is prominently displayed in a large, elegant, cursive font. The author's name, "CLARE LYDON," is placed above the title. The overall design is vibrant and eye-catching, with a mix of warm and cool tones to create a sense of contrast and appeal.



The image shows a compact electronic module with multiple wired connections attached via white connectors. The colored wires (red, black, yellow, and others) indicate power, signal, or data transmission. A metallic shield covers a key component, likely for RF or EMI protection. The board features multiple ports, allowing extensive wired interfacing. Two small push buttons and surface-mount components are visible. The design suggests it is a wired communication or control module, possibly for IoT, telemetry, or embedded systems, relying on physical connections for data and power transmission.



Questions

Input

Step1X-Edit

BAGEL-1.5B

BAGEL-7B

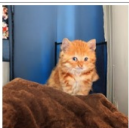
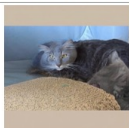
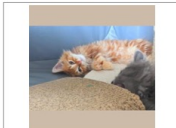
Could a heart-shaped crystal be incorporated into this bracelet design?



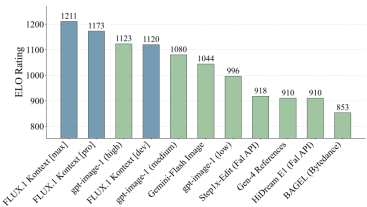
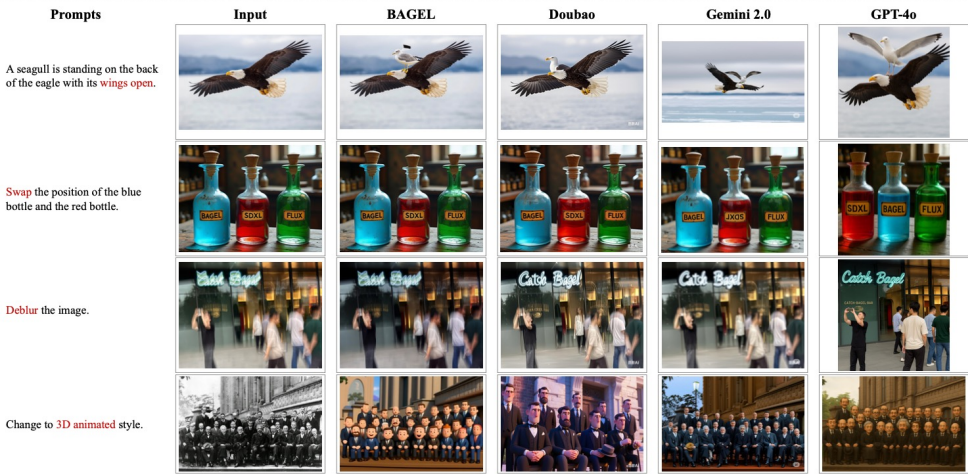
Could you provide a picture of these shoes being worn?



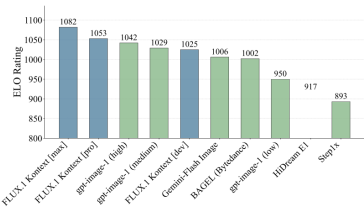
Could you show me this cat sitting upright?



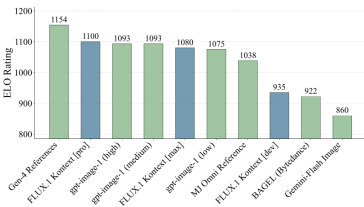
BAGEL — Showcase



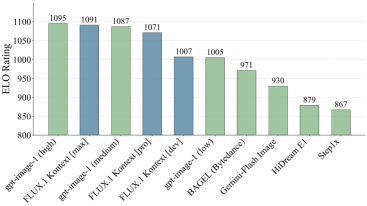
(a) Text Editing



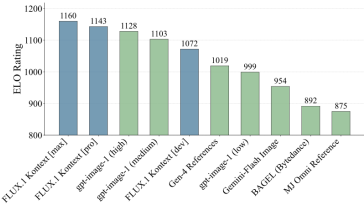
(b) Local Editing



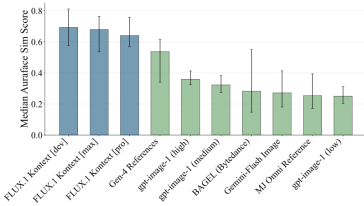
(c) Style Reference



(d) Global Editing



(e) Character Reference



(f) AuraFace Similarity

Figure 9: **Image-to-image evaluation on KontextBench.** We show evaluation results across six in-context image generation tasks. *FLUX.1 Kontext* [pro] consistently ranks among the top performers across all tasks, achieving the highest scores in Text Editing and Character Preservation.

BAGEL – Showcase

Table 2: Performance of different models across different reasoning dimensions and metrics, including Visual Consistency (VC), Visual Quality (VQ), Instruction Following (IF), and Knowledge Plausibility (KP). Scores marked with * indicate models unable to handle multi-image input tasks, with the corresponding task scores set to 0. The performance of open-source and closed-source models is separately marked with the best performance in **bold**, and the second best underlined.

	Reasoning Dimension	Metric	Closed-Source Models			Open-Source Models								
			GPT-4o	Gemini 2.0	Doubao	OmniGen	Emu2	BAGEL	BAGEL-Think	StepIX-Edit	AnyEdit	MagicBrush	InsPix2Pix	
Factual Knowledge	Attribute Perception	VC	74.50	69.50	66.75	35.75	47.75	66.75	74.75	63.00	54.75	53.50	17.50	
		VQ	94.75	81.75	89.00	49.50	75.25	67.00	75.00	70.25	67.50	76.25	55.50	
		IF	80.25	47.75	57.00	28.50	31.50	40.50	49.50	33.25	20.75	32.00	18.00	
		Avg	83.17	66.33	70.92	37.92	51.50	58.08	66.42	55.50	47.67	53.92	30.33	
	Spatial Perception	VC	69.50	60.50	67.50	24.00	41.50	53.50	77.25	64.25	55.75	38.00	13.25	
		VQ	94.50	83.25	89.00	50.00	77.75	71.25	81.25	83.00	72.00	69.25	40.25	
		IF	73.25	46.25	21.00	10.75	18.25	38.75	44.75	8.00	7.75	11.50	10.50	
		Avg	79.08	63.33	59.17	28.25	48.83	54.50	67.75	51.75	45.17	39.58	21.33	
	Temporal Prediction	VC	54.00	54.50	26.75	19.25	12.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	
		VQ	86.25	75.00	77.50	26.25	37.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	
		IF	64.50	62.25	17.50	20.00	16.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	
		Avg	68.25	63.92	40.58	21.83	22.17	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	
	Average	–	79.80	65.26	63.30	33.11	45.40	47.71	55.77	45.52	39.26	41.84	23.33	
Conceptual Knowledge	Social Science	VC	83.00	77.00	72.00	37.25	32.75	75.75	76.50	63.25	62.00	54.00	15.75	
		VQ	95.75	83.75	86.50	46.00	72.75	75.50	77.75	72.50	66.75	70.00	50.00	
		IF	84.50	59.00	54.75	22.50	22.00	34.25	46.00	25.50	15.00	27.25	14.25	
		KP	78.75	53.00	48.75	16.75	11.25	25.25	38.25	17.50	10.50	20.50	10.25	
		Avg	85.50	68.19	65.50	30.63	34.69	52.69	59.63	44.69	38.56	42.94	22.56	
	Natural Science	VC	80.00	65.00	70.25	31.00	35.00	65.75	68.00	71.25	61.75	47.00	18.75	
		VQ	96.00	83.75	87.25	47.00	75.50	76.00	80.25	78.00	77.75	72.75	58.25	
		IF	76.50	44.75	48.00	18.25	25.00	38.25	49.00	27.50	18.25	19.00	17.50	
		KP	67.75	34.25	39.25	12.50	18.25	28.00	40.25	19.50	14.00	13.50	11.75	
		Avg	80.06	56.94	61.19	27.19	38.44	52.00	59.38	49.06	42.94	38.06	26.56	
	Average	–	81.37	59.65	62.23	28.02	37.54	52.17	59.44	48.01	41.88	39.24	25.59	
	Procedural Knowledge	Logical Reasoning	VC	81.00	73.50	64.75	15.00	23.50	74.75	71.25	58.75	55.50	37.25	14.75
			VQ	95.00	84.50	85.00	26.75	66.25	84.25	83.00	72.25	72.75	75.50	58.75
IF			59.25	33.00	24.75	4.25	7.25	23.25	29.25	20.25	10.25	5.25	3.75	
KP			51.00	25.50	16.50	1.75	2.25	16.25	21.25	12.25	7.75	2.00	2.00	
Avg			71.56	54.13	47.75	11.94	24.81	49.63	51.19	40.88	36.56	30.00	19.81	
Instruction Decomposition		VC	71.00	58.25	51.50	28.75	31.00	30.75*	32.25*	25.75*	29.75*	20.75*	9.50*	
		VQ	96.25	82.50	76.75	46.50	64.75	29.00*	25.25*	26.50*	39.25*	39.25*	27.75*	
		IF	88.00	74.25	53.50	32.25	39.25	32.75*	24.50*	16.00*	11.75*	9.25*	7.00*	
		Avg	85.08	71.67	60.58	35.83	45.00	30.83*	27.33*	22.75*	26.92*	23.08*	14.75*	
Average		–	78.32	62.90	54.17	23.89	34.91	40.23	39.26	31.82	31.74	26.54	17.28	
Overall Average			80.09	62.41	60.70	28.85	39.70	47.76	53.36	43.29	38.55	37.15	22.82	

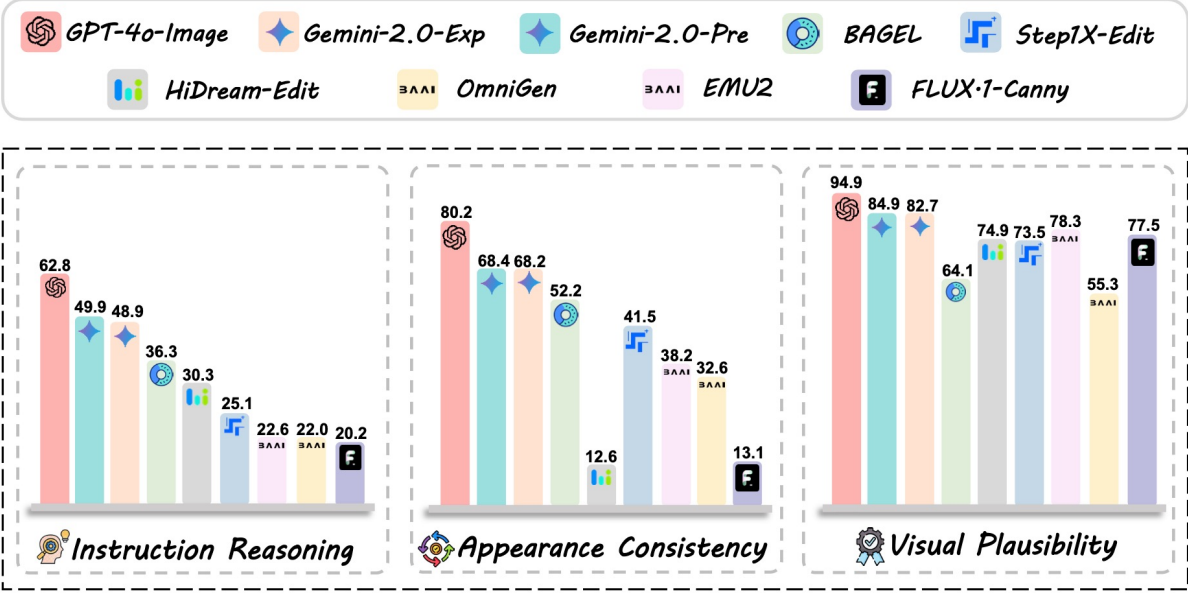


Figure 5: Comparison across models on three evaluation sub-dimensions. GPT-4o-Image demonstrates superior performance, achieving the highest scores across all three evaluation metrics. Gemini-2-Flash-Series also exhibits competitive performance on these criteria. In contrast, the performance of many other evaluated models was considerably lower, indicating significant limitations in their ability to follow instructions and maintain visual integrity.