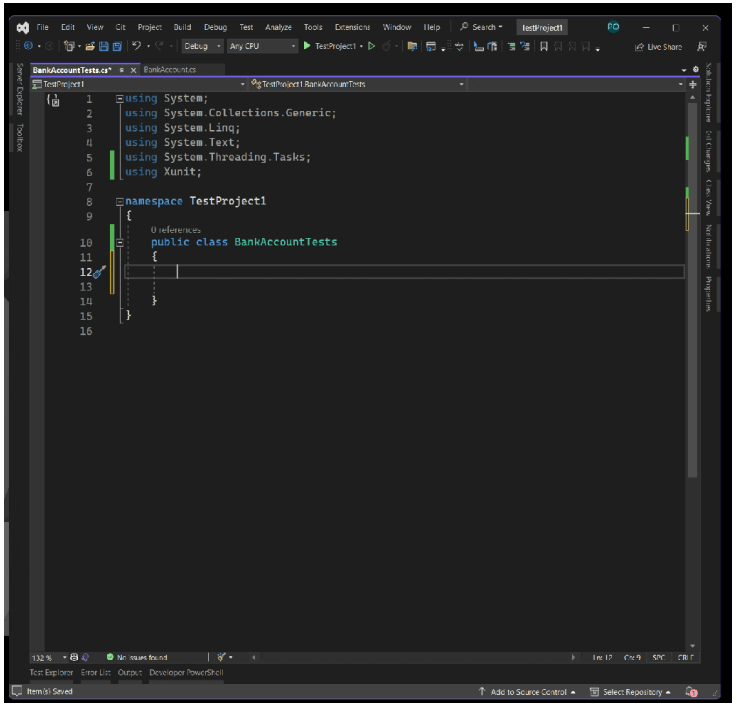
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

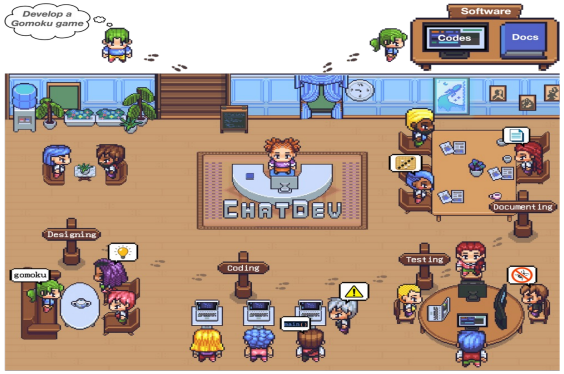
高效多模态生成：挑战、方法及应用

邓志杰
上海交通大学

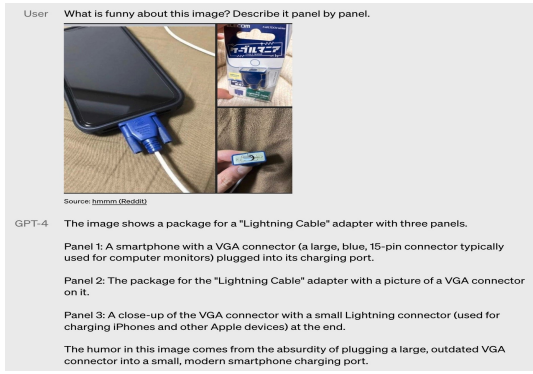
背景：语言生成已产生巨大实用价值



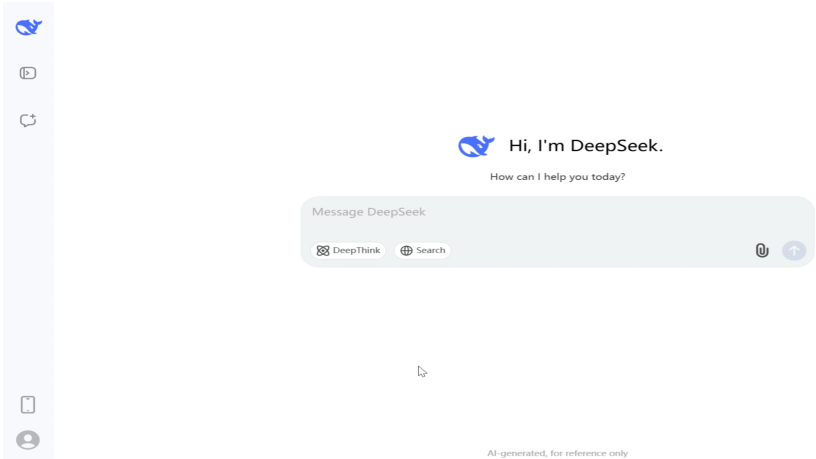
Coding assistant



Software development

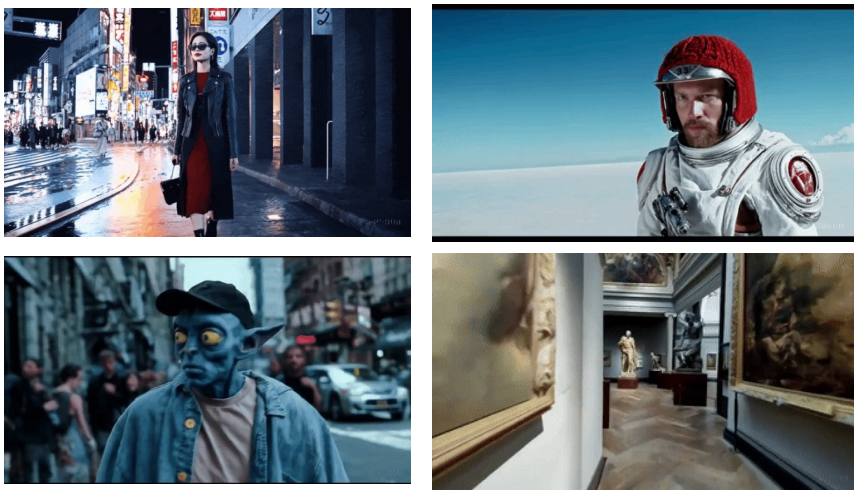


Multimodal understanding



Reasoning

背景：视觉生成构建起“世界模拟器”



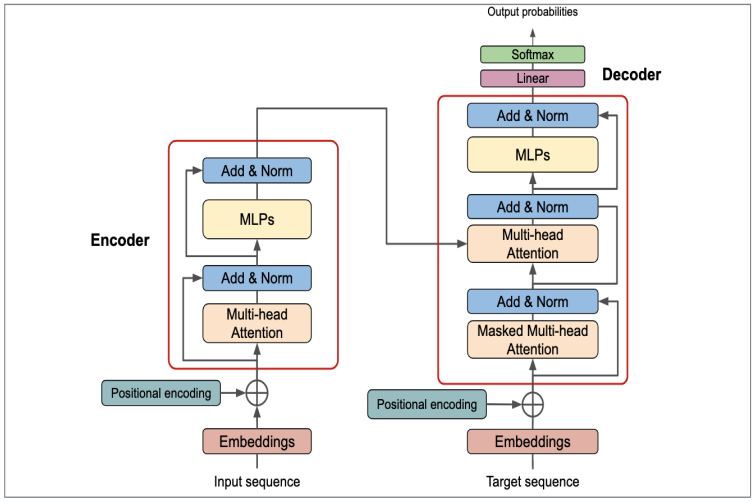
Sora by OpenAI

Vidu by ShengShu

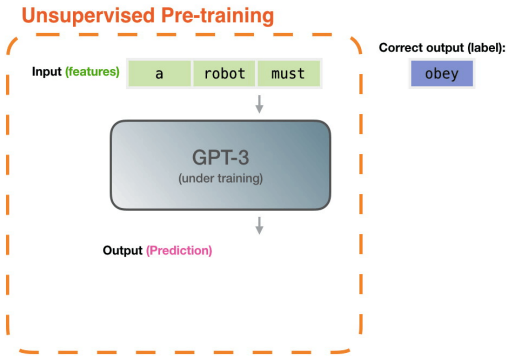
W.A.L.T

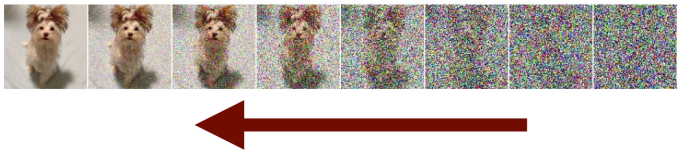
https://openai.com/index/sora/; https://www.vidu.io/text-to-video-ai; https://walt-video-diffusion.github.io/

生成式大模型的趋势

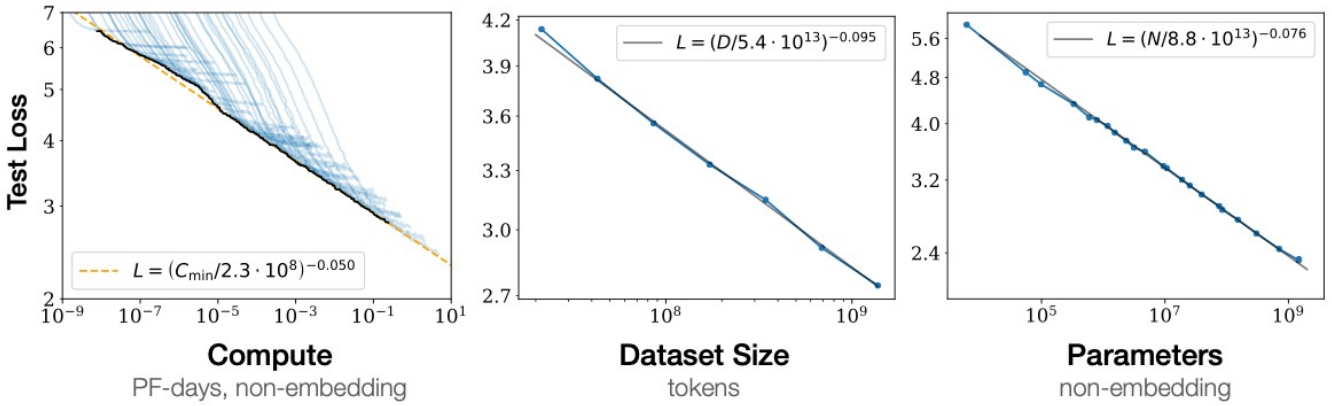


架构统一：**transformer**





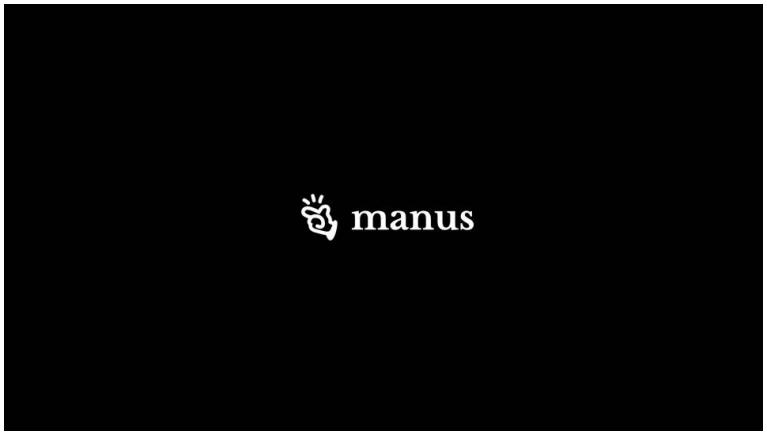
自回归和扩散建模并存（前者：长程依赖，后者：连续细节）



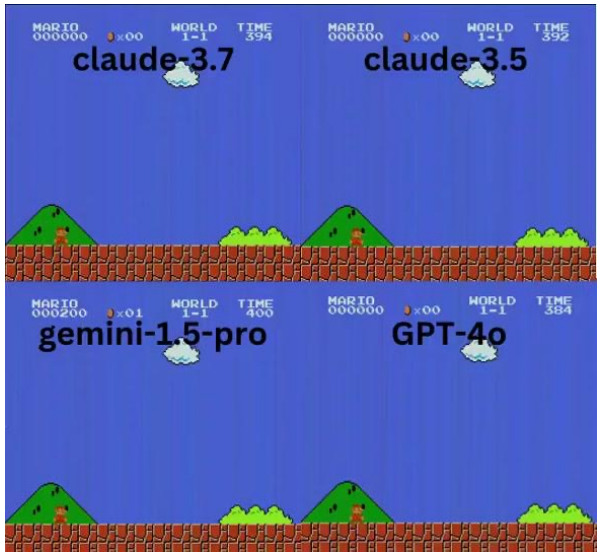
各细分领域均呈现**Scaling law**

Vaswani, Ashish, et al. Attention is all you need. NeurIPS 2017.
Kaplan, Jared, et al. Scaling laws for neural language models. 2020.

生成式大模型的最终形态: **Agent** (LLM + memory + planning skills + tool use)



Manus (a newest agent even better than OpenAI Deep Research)



Game Agent

Embodied Agent

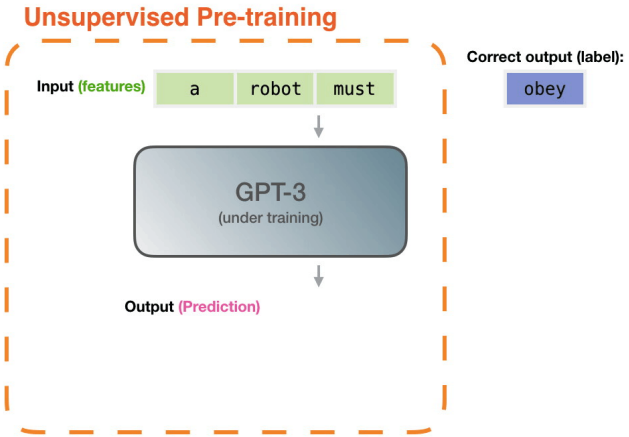
Virtual vs. reality

生成式大模型需要如何发展?

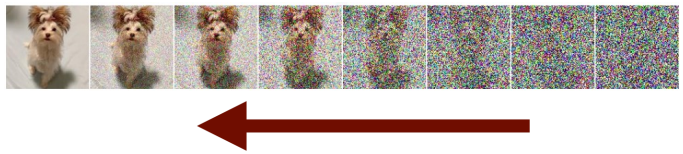
- Diverse use case -> Cross modality is needed
- Memory -> Long context -> Efficiency matters
- Planning + tool use -> Reason is important

- 我们需要： 高效多模态生成，同时兼具慢思考

挑战1：语言、视觉生成范式存在分歧



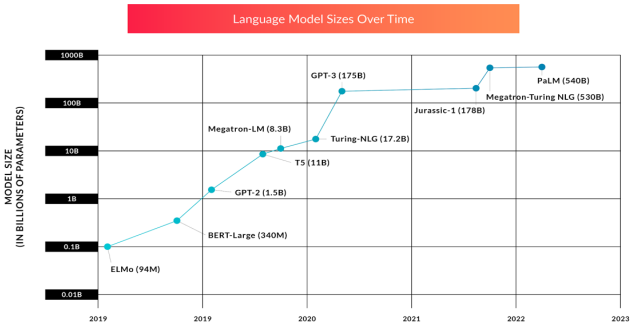
VS.



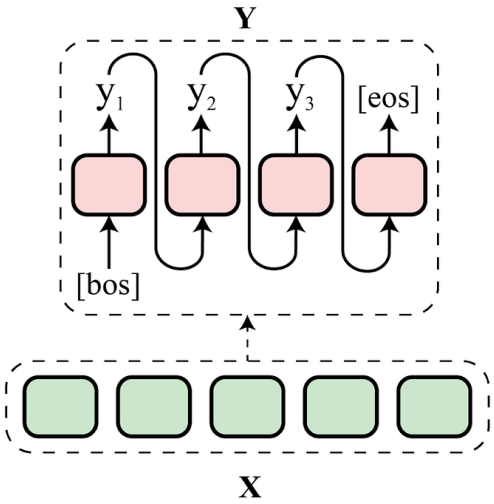
语言：自回归生成，刻画长程依赖

图像：扩散建模，准确预测连续细节

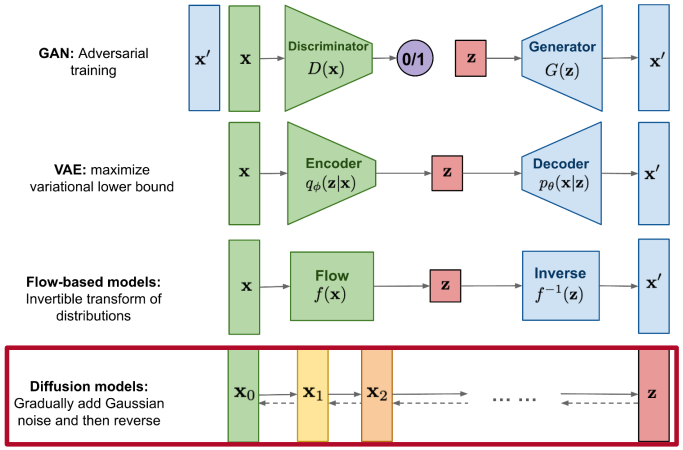
挑战2：现有架构、算法导致模型训练、推理低效，即：高成本、差体验



模型本身的大尺寸（更多的flops、内存占用）



每个新token都需要一次模型前传



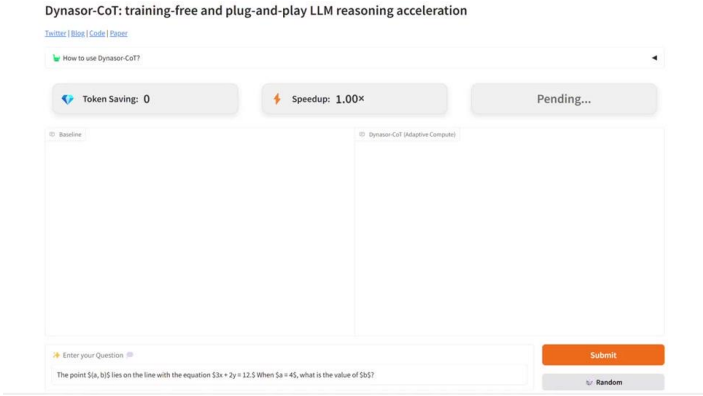
每个去噪步都需要一次模型前传

低效的顺序推理

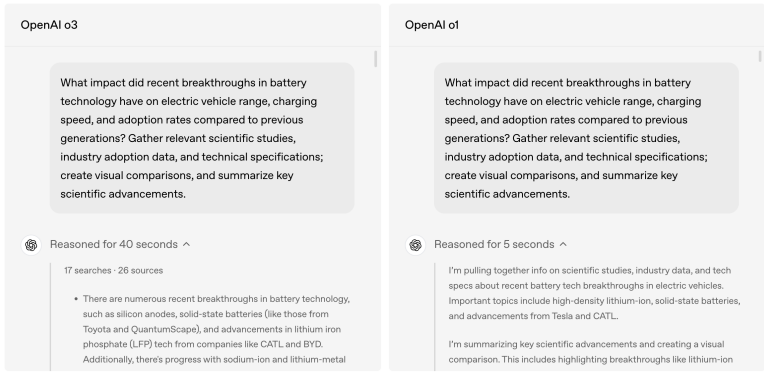
挑战3：对推理能力的兼顾为模型效率提出了新的要求

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	44.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

DeepSeek R1蒸馏的小模型推理能力弱，至少得上**7B**及以上的模型



生成的思维链中包含大量冗余“自我怀疑”



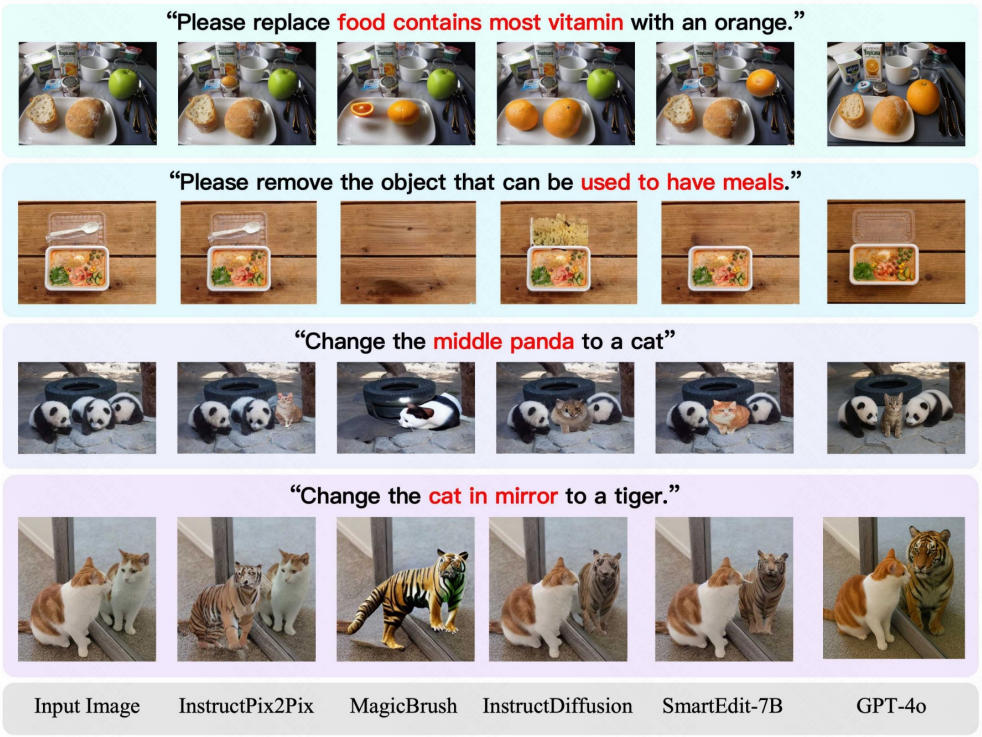
OpenAI o3 delivers a comprehensive, accurate, and insightful analysis of how recent battery technology breakthroughs are extending EV range, speeding up charging, and driving adoption, all backed by scientific studies and industry data. o1, while credible and on-topic, is less detailed and forward-looking, with minor inaccuracies or oversimplifications.

工具调用融合的推理有更高的latency

1、跨模态统一、兼具生成和理解能力的模型

GPT-4o: 跨模态生成理解统一是趋势

- 相较于专用的图像生成模型，统一语言和视觉建模有助于建立世界知识
  - 在传统编辑任务中的指令理解与跟随能力显著增强

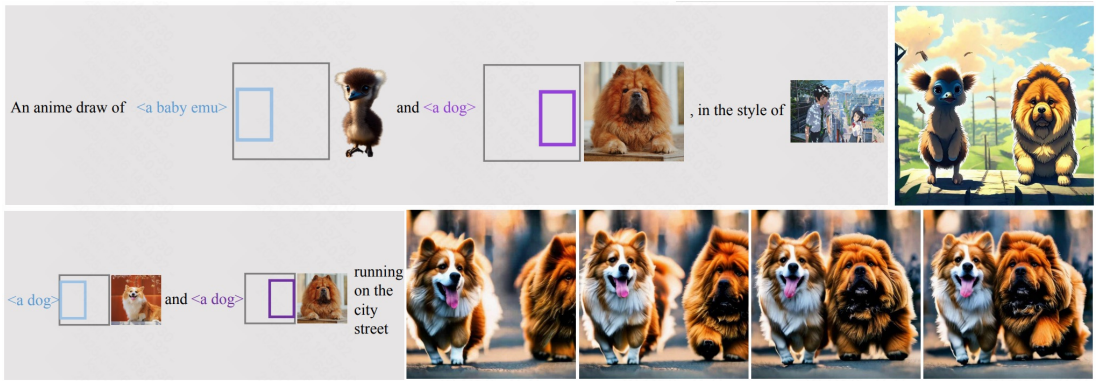


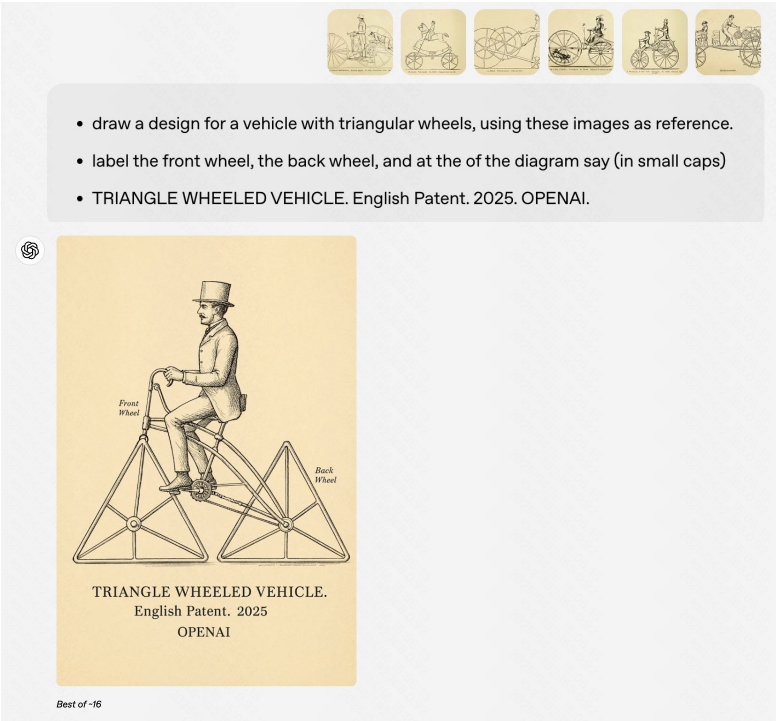
Yan, Zhiyuan, et al. "Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation." *arXiv:2504.02782* (2025).

GPT-4o: 跨模态生成理解统一是趋势

统一模型天然具备长上下文学习能力

处理多图与文本混合输入时，统一模型能够有效整合多模态信息，展现了控制精准、主体一致性强的生成效果





Sun, Quan, et al. "Generative multimodal models are in-context learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
https://openai.com/index/introducing-4o-image-generation/

如何将跨模态生成理解统一？模型桥接

- 模型桥接的方式能轻易实现跨模态生成

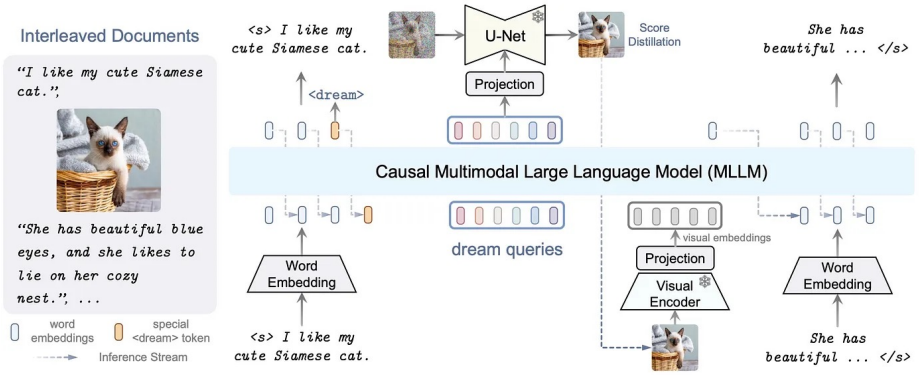
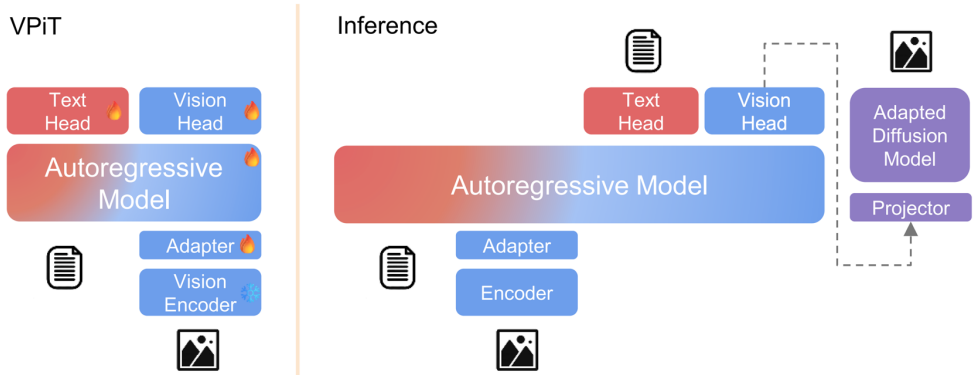


Figure 2: **Overview of of our DREAMLLM framework**. Interleaved documents serve as input, decoded to produce outputs. Both text and images are encoded into sequential, discrete token embeddings for the MLLM input. A special <dream> token predicts where to generate images. Subsequently, a series of *dream queries* are fed into the MLLM, capturing holistic historical semantics. The images are synthesized by the SD image decoder conditioned on queried semantics. The synthesized images are then fed back into the MLLM for subsequent comprehension.

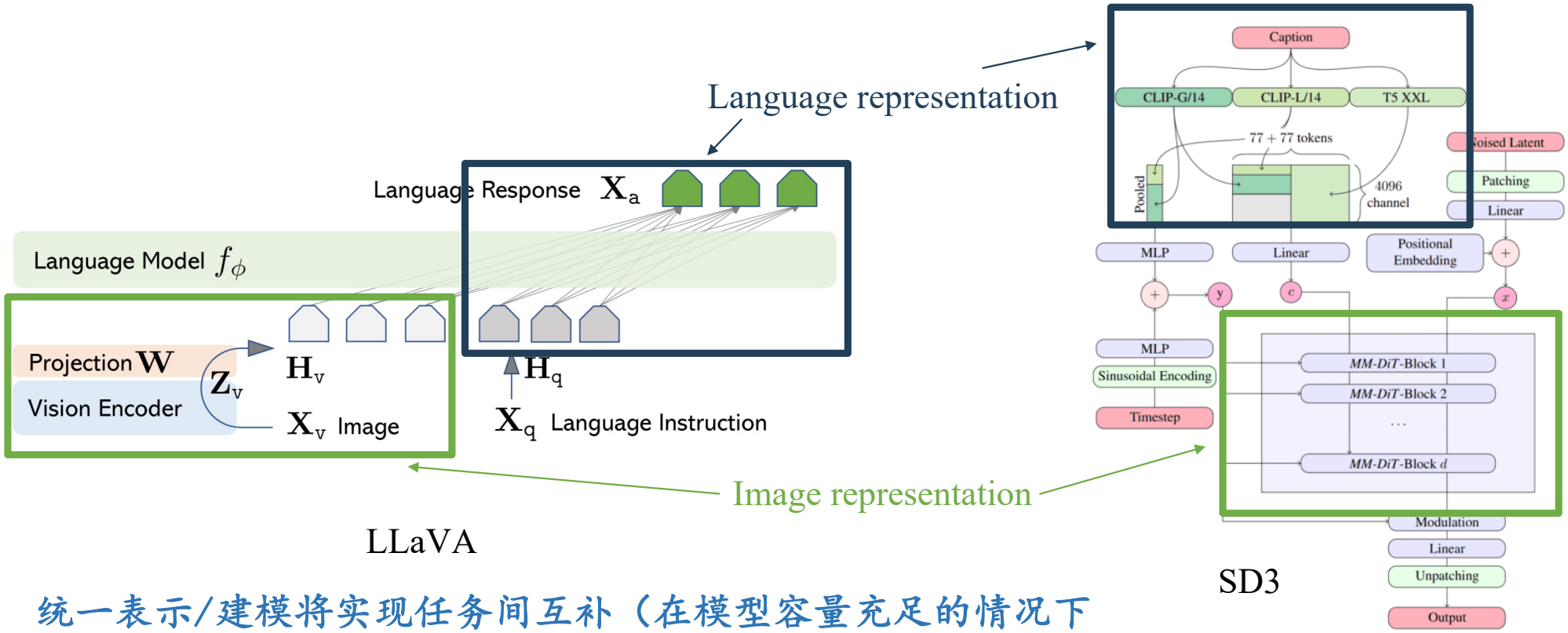
DreamLLM [https://arxiv.org/abs/2309.11499]



MetaMorph [https://arxiv.org/pdf/2412.14164]

如何将跨模态生成理解统一？模型桥接

- 扩散模型与语言模型中的图像/文本表示冗余



统一表示/建模将实现任务间互补（在模型容量充足的情况下）

如何将跨模态生成理解统一？自回归

- 图像离散化，统一自回归（但是离散化会丢信息）

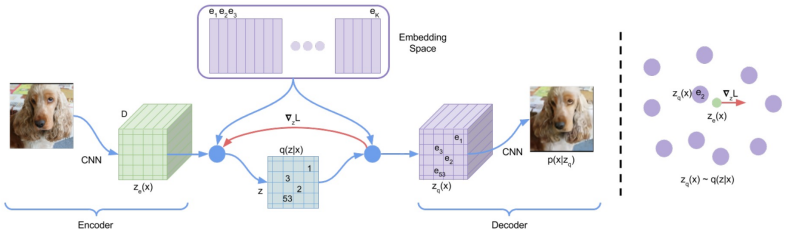
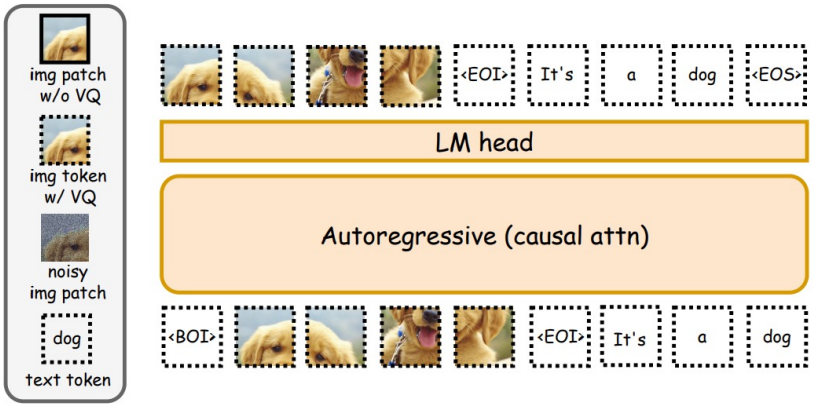


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

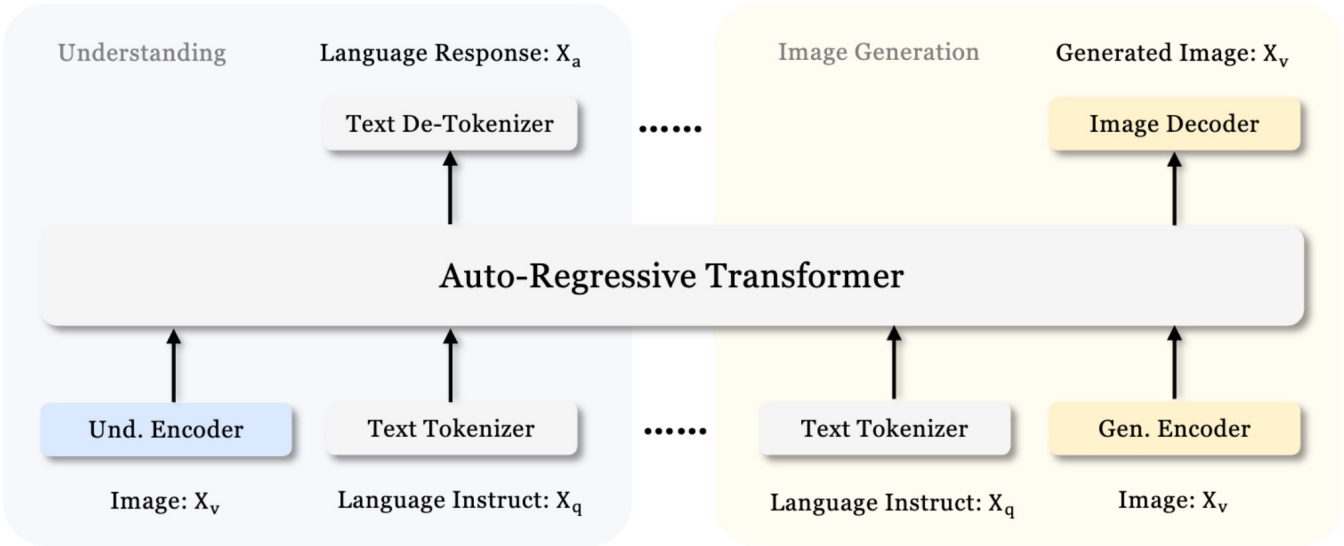
VQ-VAE编码丢失细节



Chameleon, EMU3, etc.

如何将跨模态生成理解统一？自回归

- 图像离散化，统一自回归（但是离散化会丢信息）
  - DeepSeek Janus-Pro：为图像理解和生成分别使用连续和离散编码器



如何将跨模态生成理解统一？参数共享的图像扩散+文本自回归

- 对于图像，扩散建模

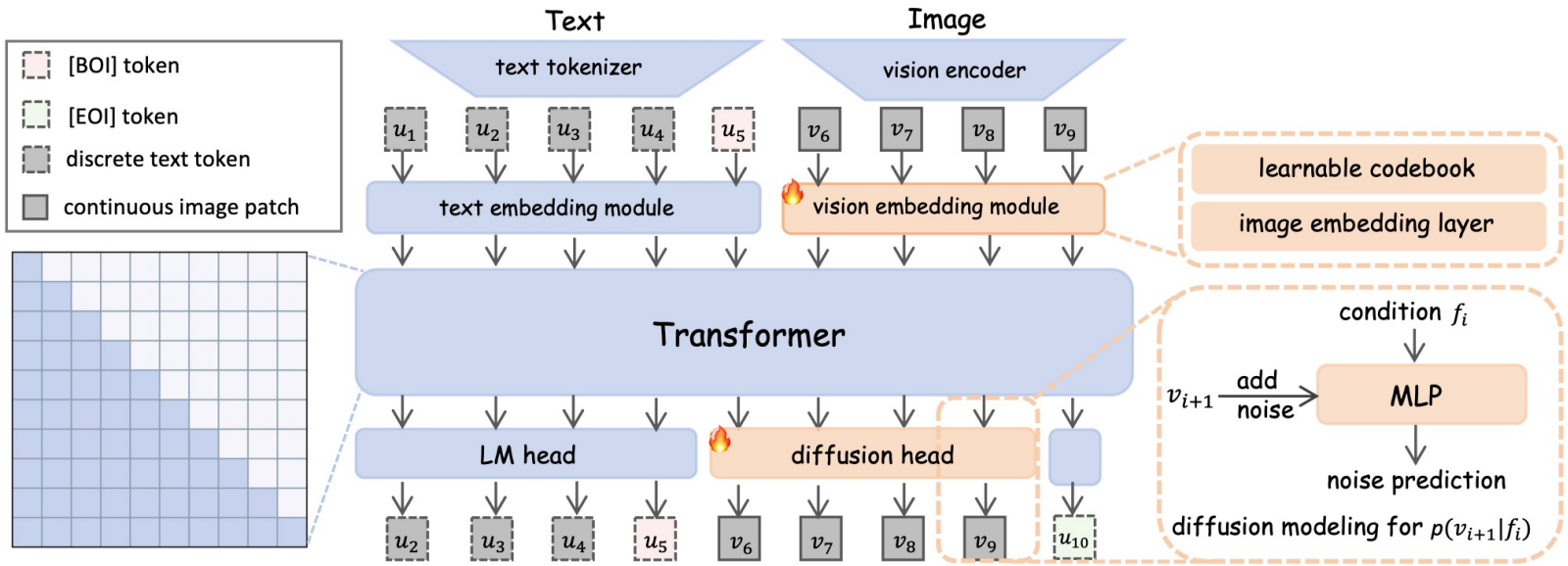
Transfusion

图文交错训练效率低：

- 图-文-图-文-图，只能在最后一张图算loss

图像生成不能使用**KV Cache**

如何将跨模态生成理解统一？Orthus!



- 自回归**Transformer**主干（拥抱**KV Cache**）
- 处理离散的文本**token**和连续的图像**feature**（基于连续VAE）
- 基于线性层定义的**language head**和**diffusion MLP**来分别生成文和图（逐**token/patch**）

Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

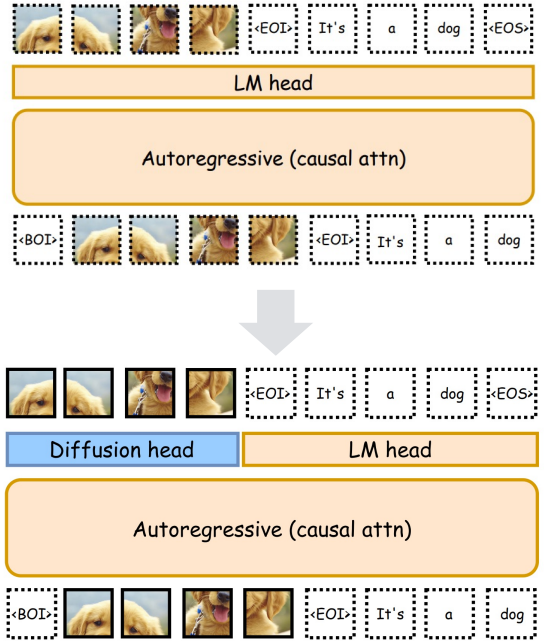
如何将跨模态生成理解统一？Orthus!

- 从离散图像特征到连续特征：

$$\boldsymbol{h}_i = \sum_j \boldsymbol{w}_j \mathbb{1}_{\tilde{v}_i=j},\ \tilde{v}_i = \underset{j\in\{1,\dots,K\}}{\arg\min}\ d(\boldsymbol{v}_i, \boldsymbol{c}_j)$$

$$\Rightarrow \quad \boldsymbol{h}_i = \sum_j \boldsymbol{w}_j \frac{e^{-d(\boldsymbol{v}_i,\boldsymbol{c}_j)/\tau}}{\sum_{k=1}^{K} e^{-d(\boldsymbol{v}_i,\boldsymbol{c}_k)/\tau}}$$

- 自回归统一模型（如：**Chameleon**）： $\tau = 0$
- **Orthus: $\tau = 1$**



- 从 $\tau = 0$的模型冷启动
  - 72个A100 GPU hours即可得到Orthus-7B-base
- 将涉及的**VQ-VAE**调成了**VAE**

Model	PSNR↑	SSIM [63]↑
VQ-VAE	23.7	0.80
Ours	26.1	0.84

Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

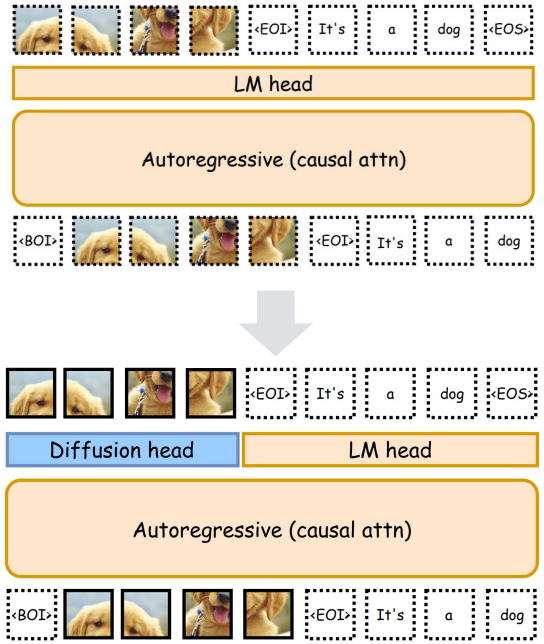
如何将跨模态生成理解统一？Orthus!

- **Diffusion head**训练：

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\epsilon,t}[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{v}_{i+1} + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \boldsymbol{f}_i)\|_2^2]$$

- 从文到图/图到文等不同任务学习有价值信号
  - 1:1混合LlaVA-v1.5-665K指令微调数据和高质量文生图数据JourneyDB、LAION-COCO-aesthetic (recaptioned from ShareGPT-4v)

$$\mathcal{L}_{Orthus} = \mathcal{L}_{\text{ar}} + \lambda\mathcal{L}_{\text{diff}}$$



Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

Orthus：文生图/图生文量化结果

- 在多个图像理解指标上超越了现有混合理解生成模型**Chameleon**和**Show-o**，并在文到图生成的**GenEval**指标上超过**SDXL**

Table 3. **Comparison with state-of-the-arts on visual generation benchmarks.** Model using external pre-trained diffusion model is marked with * and Chameleon† is post-trained with the same dataset as Orthus. The results in **bold** and underline are the best and second-best results, respectively.

Type	Model	Res.	GenEval	HPS
Gen. Only	SDv1.5 (Rombach et al., 2022)	512	0.43	27.0
	SDv2.1 (Rombach et al., 2022)	512	0.50	27.2
	DALL-E (Ramesh et al., 2022)	512	0.52	26.9
	Emu3-Gen (Wang et al., 2024)	512	0.54	-
	SDXL (Podell et al., 2023)	512	0.55	30.9
	SD3(d=30) (Esser et al., 2024)	512	0.64	-
Und. & Gen.	SEED-X* (Ge et al., 2024)	448	0.49	-
	LWM (Liu et al., 2024e)	256	0.47	26.1
	Show-o (Xie et al., 2024)	256	0.53	27.3
	Transfusion (Zhou et al., 2024)	256	**0.63**	-
	Chameleon†	512	0.43	26.9
	Orthus (Ours)	512	0.58	**28.2**

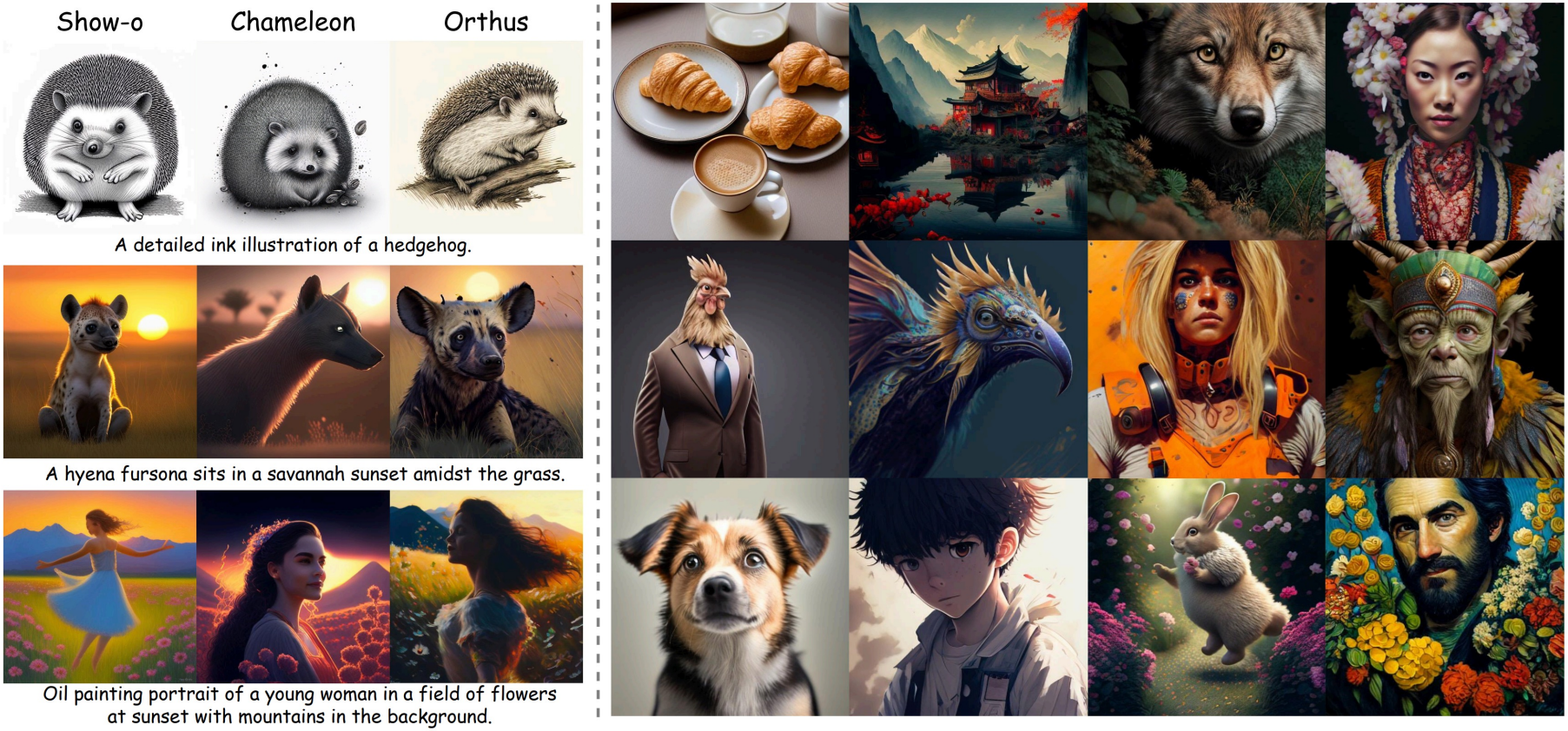
Table 2. **Evaluation on visual understanding benchmarks.** Und. and Gen. denote "understanding" and "generation", respectively. Models using external pre-trained diffusion models are marked with * and Chameleon† is post-trained with the same dataset as Orthus. The results in **bold** and underline are the best and second-best results, respectively. The results correspond to the exact match accuracy.

Type	Model	# Params	POPE↑	MME-P↑	VQAv2↑	GQA↑	MMMU↑
Und. Only	LlaVa (Liu et al., 2024d)	7B	76.3	809.6	-	-	-
	LlaVA-v1.5 (Liu et al., 2024b)	7B	85.9	1510.7	78.5	62.0	35.4
	InstructBLIP (Dai et al., 2023)	7B	-	-	-	49.2	-
	Qwen-VL-Chat (Bai et al., 2023)	7B	-	1487.5	78.2	57.5	-
	Emu3-Chat (Wang et al., 2024)	8B	85.2	1243.8	75.1	60.3	31.6
	InstructBLIP (Dai et al., 2023)	13B	78.9	1212.8	-	49.5	-
Und. and Gen.	Emu* (Sun et al., 2023)	13B	-	-	52.0	-	-
	NExT-GPT* (Wu et al., 2013)	13B	-	-	**66.7**	-	-
	Gemini-Nano-1 (Team et al., 2023)	1.8B	-	-	62.7	-	26.3
	Show-o (Xie et al., 2024)	1.3B	73.8	948.4	59.3	48.7	25.1
	LWM (Liu et al., 2024e)	7B	75.2	-	55.8	44.8	-
	Chameleon†	7B	77.8	1056.9	57.8	49.6	26.7
	Orthus (Ours)	7B	**79.6**	**1265.8**	63.2	**52.8**	**28.2**

Model	Res.	GenEval ↑	HPSv2↑	POPE↑	MME↑	GQA↑
Orthus	512	0.58	28.2	79.6	1265.8	52.8
VILA-U	256	0.40	25.3	83.9	1336.2	58.3
Janus	384	0.61	27.8	87.0	1338.0	59.1

Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

Orthus：文生图可视化结果



Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

Orthus：ablation结果

Table 4. Comparisons of the performance of Orthus via separate training and unified training across multimodal benchmarks.

Type	$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{ar}}$	POPE↑	MME-P↑	GQA↑	GenEval↑
Und. only	✗	✓	78.7	1244.2	51.9	-
Gen. only	✓	✗	-	-	-	0.56
Und. & Gen.	✓	✓	**79.6**	**1265.8**	**52.8**	**0.58**

同时从文到图和图到文数据学习可以实现1+1>2

Table 5. Ablation study on the choice of vision embedding modules on visual understanding tasks.

Type	POPE↑	MME-P↑	VQAv2↑	GQA↑	MMMU↑
softmax	**78.7**	**1244.2**	**60.8**	**51.9**	**28.0**
argmin	77.6	1064.8	57.9	50.1	26.7
linear	70.4	800.7	50.3	44.5	22.3

连续的图像特征对于视觉理解任务必要，但要避免冷启动

Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

Orthus：图文交错生成结果（图文->图、demo+图->图、图文->图文图文）

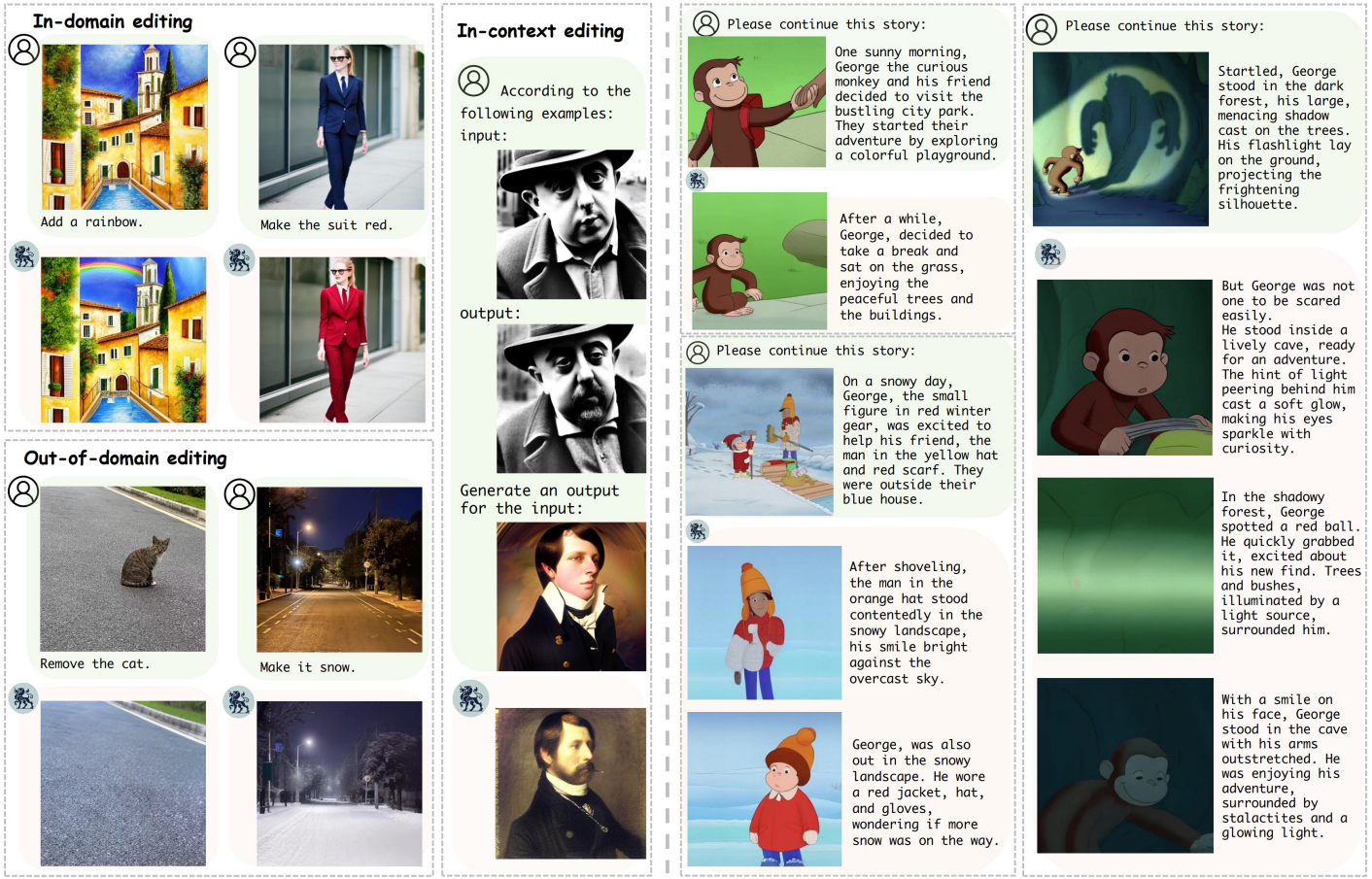


Table 1. Comparisons of CLIP similarities (Ruiz et al., 2023; Gal et al., 2022) between editing-specific diffusion models and Orthus on the test dataset of Instruct-Pix2Pix.

Model	-T↑	-I↑	-D↑
PnP (Tumanyan et al., 2023)	0.156	0.76	0.023
SDEdit (Meng et al.)	0.229	0.84	0.047
I-Pix2Pix (Brooks et al., 2023)	0.233	**0.88**	0.045
Orthus (Ours)	**0.238**	0.87	**0.049**

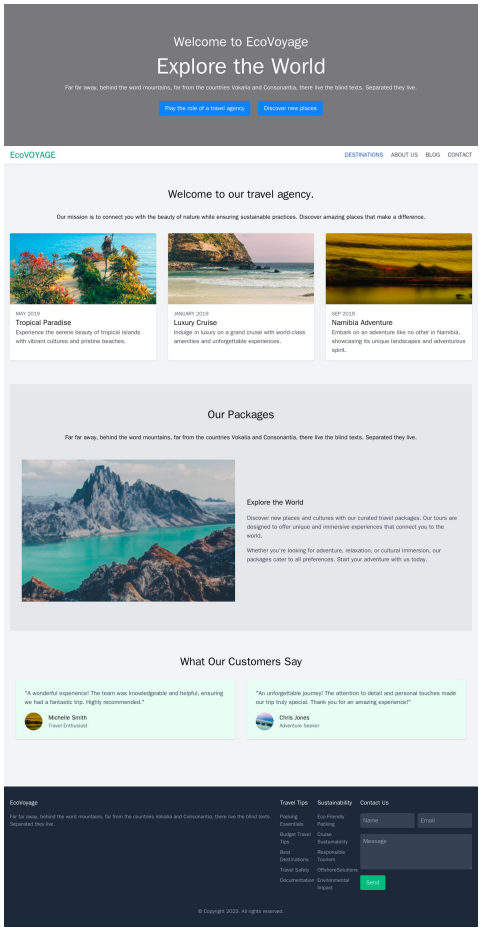
图像编辑能力指标

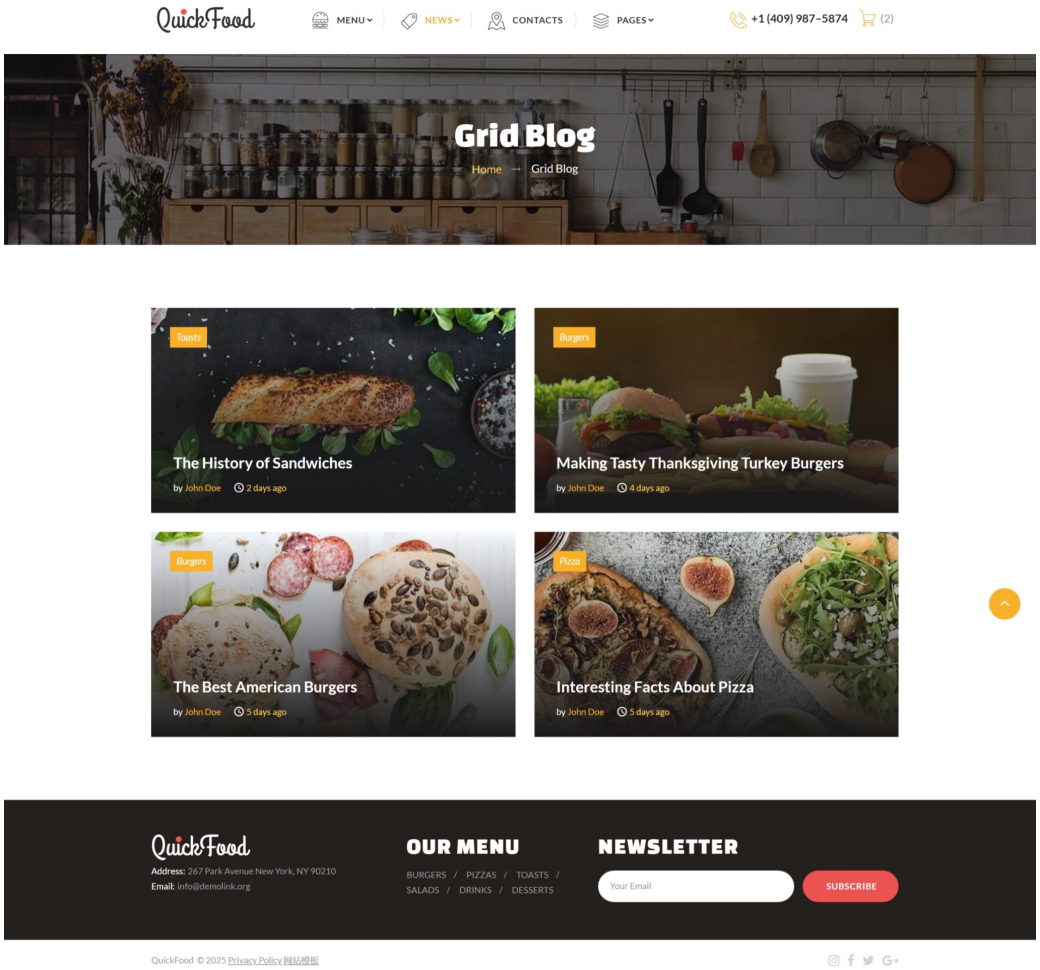
	Orthus	Show-o	NExT-GPT	MiniGPT-5	GILL	SEED-X
OpenING-IVD ↑	6.3	5.1	5.2	5.3	6.2	8.0

图文交错生成指标

Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

Orthus：图文交错的HTML网页生成





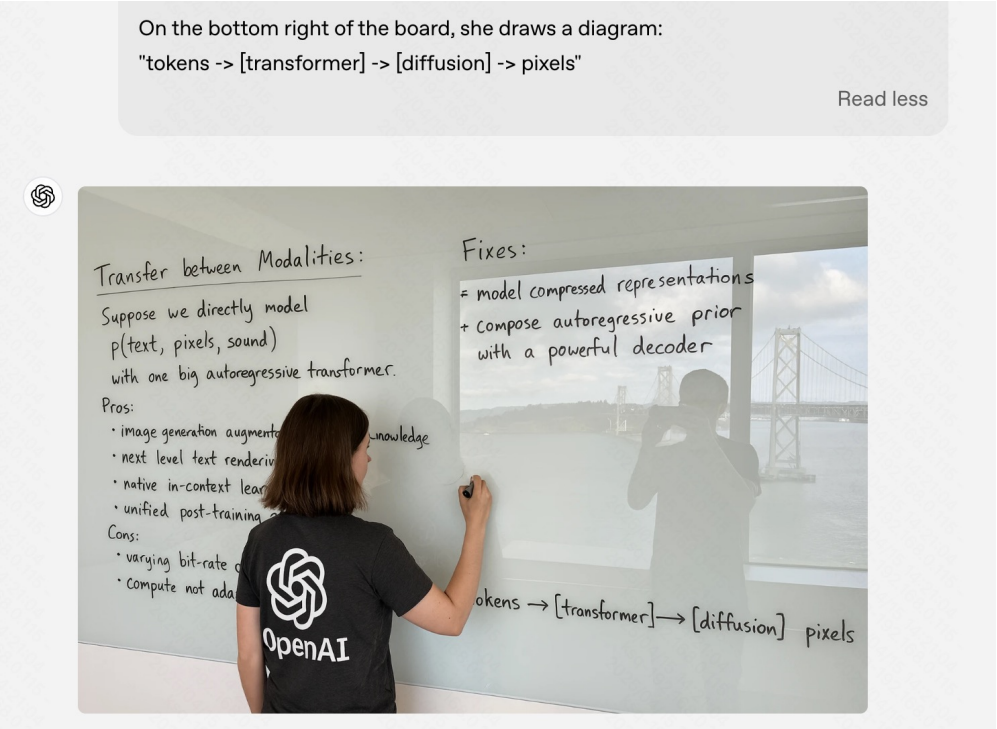
Kou, Jin, Liu, Ma, Jia, Chen, Jiang, Deng. Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads

如何将跨模态生成理解统一？

GPT-4o的彩蛋：tokens -> [transformer] -> [diffusion] -> pixels

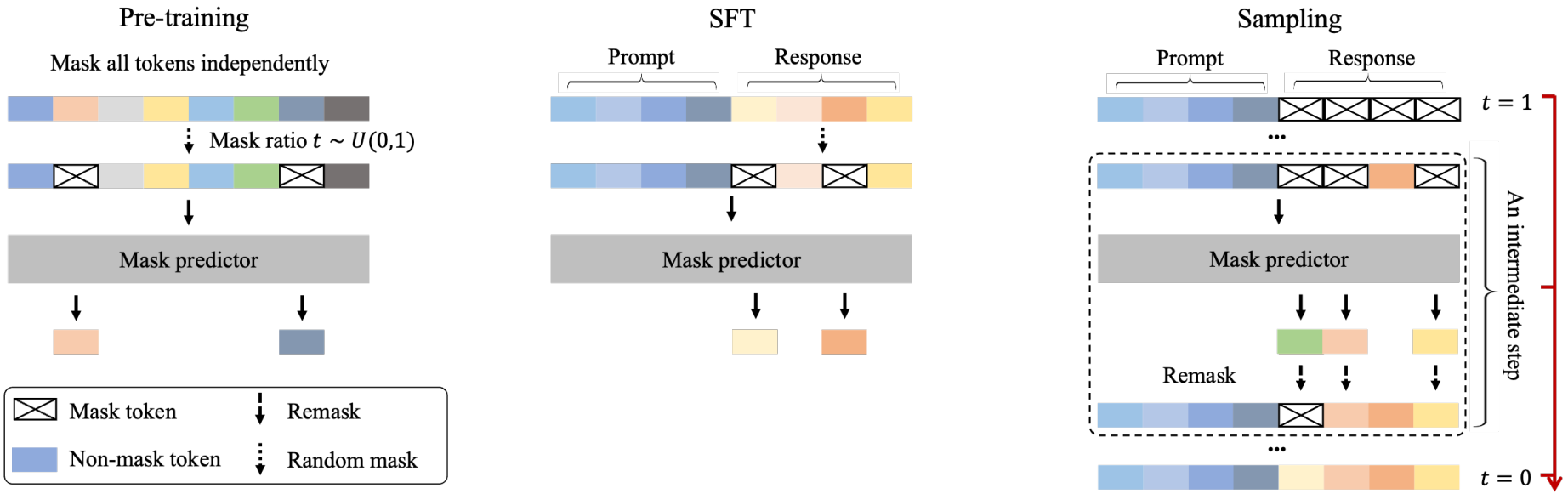
结合自回归模型的语义建模优势和扩散模型的细节建模优势

- 与Orthus大思路一致
  - 如何改进生成效率？Diffusion forcing？



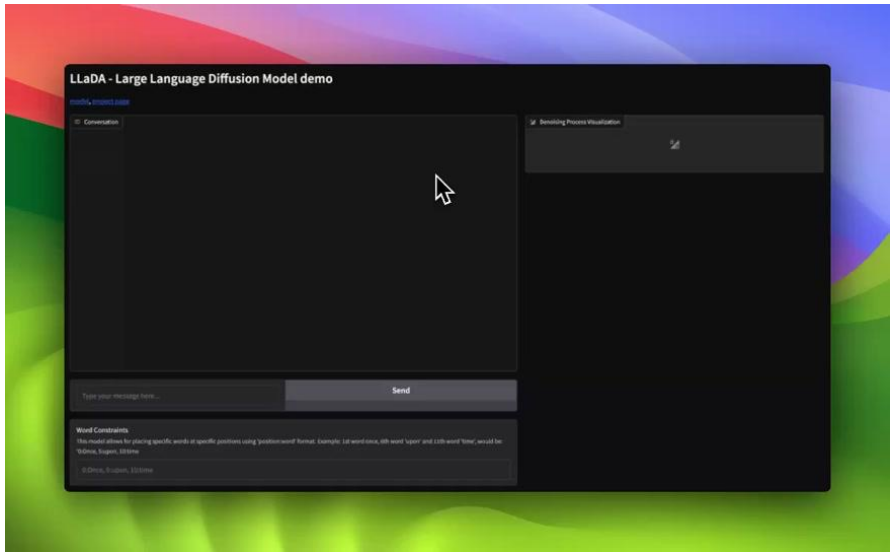
语言建模能否基于扩散模型？可以！

- 离散扩散模型（**discrete diffusion models**）

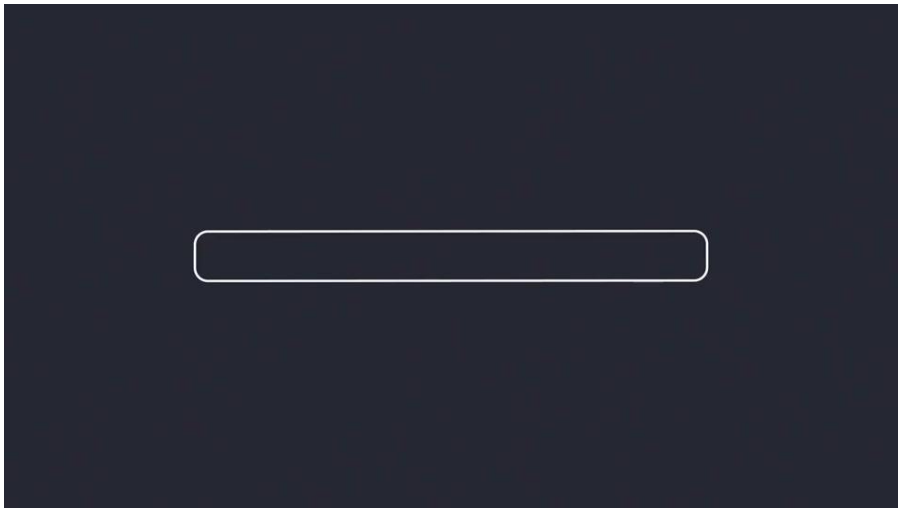


Nie et al., Large Language Diffusion Models

对离散扩散语言模型的期待：并行生成、推理加速



LLaDA (The first Large Language Diffusion Model, RUC)，可达LLaMA3级水平



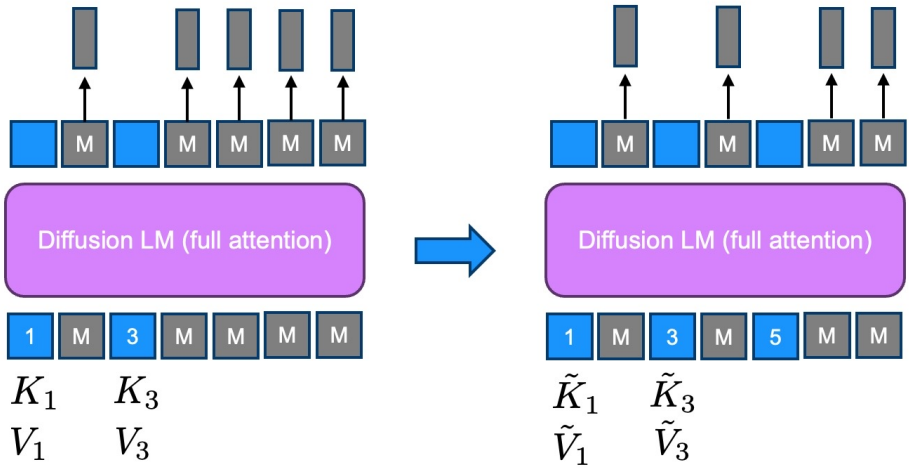
Mercury, 5-10x快的代码生成, InceptionAI Labs (co-founded by Stefano Ermon)

现有离散扩散语言模型的问题

- 需预先设定生成内容的总长度
- 无法应用KV cache：已经被解码的token的key和value始终会在去噪过程中发生变化
- 同时解码多个token时会产生潜在的训练-推理不一致的风险

现有离散扩散语言模型的问题

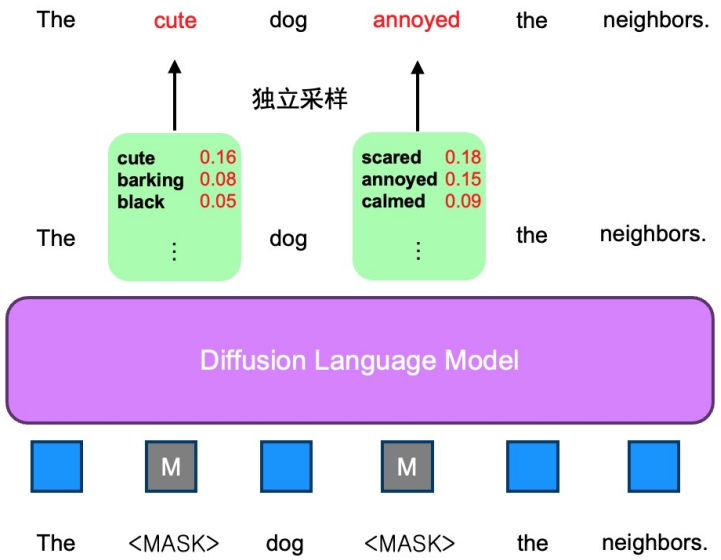
- 需预先设定生成内容的总长度
- 无法应用KV cache：已经被解码的token的key和value始终会在去噪过程中发生变化
- 同时解码多个token时会产生潜在的训练-推理不一致的风险



由于扩散模型的全局注意力机制，已经被解码的token的key和value依旧会在生成过程中产生变化，每一步迭代都需重新计算

现有离散扩散语言模型的问题

- 需预先设定生成内容的总长度
- 无法应用KV cache：已经被解码的token的key和value始终会在去噪过程中发生变化
- 同时解码多个token时会产生潜在的训练-推理不一致的风险

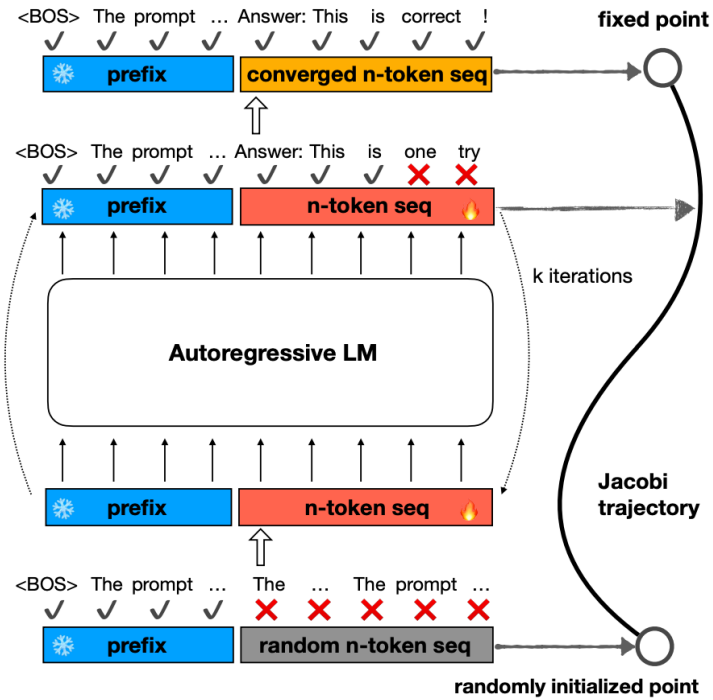


虽然diffusion LM可以一次预测多个token的logit，其预测之间相互独立，同时解码相关性强的多个token可能会产生冲突。

离散扩散：从自回归出发

- 自回归模型的不动点迭代解码（并行解码）

- 并行地对拟生成的**n**个**token**进行去噪
  - 每个迭代步，有至少一个正确token被生成
  - 自回归生成 => 离散去噪生成



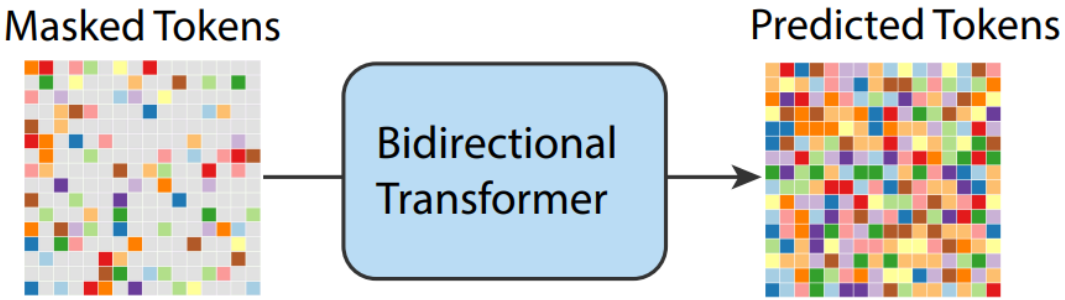
$$
\begin{aligned}
v_1^{k+1} &= \arg\max_v p_\theta(v|v_1^k, \mathbf{u}), \\
v_2^{k+1} &= \arg\max_v p_\theta(v|v_1^k, v_2^k, \mathbf{u}), \\
&\dots \\
v_n^{k+1} &= \arg\max_v p_\theta(v|v_1^k, \dots, v_{n-1}^k, \mathbf{u})
\end{aligned}
$$

Song, Y., Meng, C., Liao, R., and Ermon, S. Accelerating feedforward computation via parallel nonlinear equation solving. ICML 2021.
Santilli, A., Severino, S., Postolache, E., Maiorca, V., Mancusi, M., Marin, R., and Rodola, E. Accelerating transformer inference for translation via parallel decoding. ACL 2023.

离散扩散：图像生成的一大分支

- **MaskGIT** [Chang et al., 2022]

Masked Visual Token Modeling (MVTM)

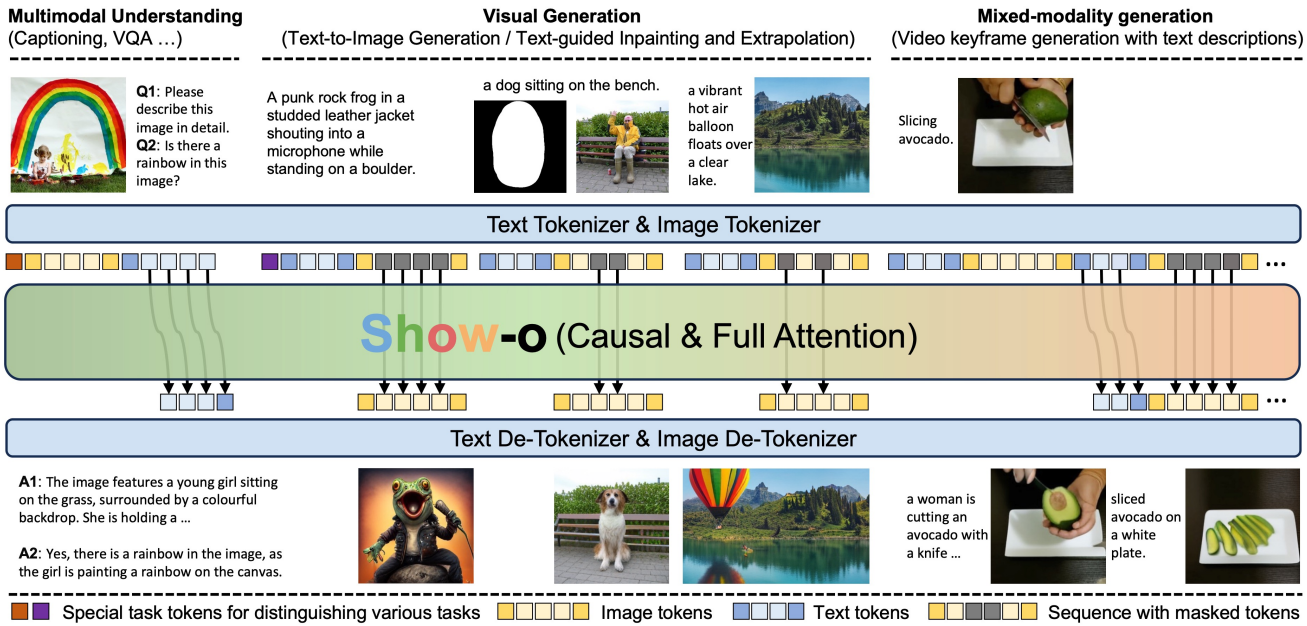


$$\mathcal{L}_{\text{mask}} = -\mathop{\mathbb{E}}_{\mathbf{Y}\in\mathcal{D}}\Big[\sum_{\forall i\in[1,N],m_i=1}\log p(y_i|Y_{\overline{\mathbf{M}}})\Big],$$

- 采样过程
  - 预测当前mask tokens的logits
  - 从logits进行采样替代mask tokens
  - 对新tokens进行remask

基于离散扩散的跨模态生成

- 将已有文本自回归+图像扩散的模型拓展为统一扩散模型？
  - Show-o的文本部分是自回归，可以被并行解码算法转化为离散扩散
  - Show-o的图像部分是类MaskGIT的离散扩散



Show-o architecture

基于离散扩散的跨模态生成

- 统一扩散的跨模态生成模型

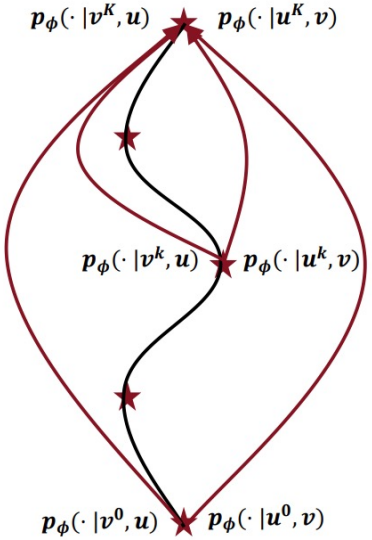
Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

基于离散扩散的跨模态生成及加速：Show-o Turbo

- 统一扩散的跨模态生成模型
  - 如何跨模态统一地减少“去噪”步？
    - 一致性蒸馏（consistency distillation）的离散扩散拓展

$$\mathcal{L}_c^u = \mathbb{E}_{k\sim\mathcal{U}(0,K)} d\left(p_{\phi^-}(\cdot|\mathbf{u}^K,\mathbf{v}), p_\phi(\cdot|\mathbf{u}^k,\mathbf{v})\right)$$

    - 必要的正则化



Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

Song, Dhariwal, Chen, Sutskever. Consistency Models.

Show-o Turbo：结果

User: Describe in one sentence what the person in the picture is doing.

Show-o Turbo:

0:	in	' '	sentence	IST	.	<|eoi|>	one	'\n'	<|mmu|>	doing	in	' '	is	the	one	<|mmu|>
1:	The	person	in	the	person	picture	a	a	a	a	a	a	a	a	a	a
2:	The	person	in	the	picture	is	is	a	a	a	a	a	a	a	a	a
3:	The	person	in	the	picture	is	riding	a	dirt	dirt	,	a	a	a	a	a
4:	The	person	in	the	picture	is	riding	a	dirt	bike	on	a	dirt	dirt	,	a
5:	The	person	in	the	picture	is	riding	a	dirt	bike	on	a	dirt	track	,	covered
6:	The	person	in	the	picture	is	riding	a	dirt	bike	on	a	dirt	track	,	kicking

User: Please explain what sport the person in the picture is doing?

Show-o Turbo:

0:	is	USER	the	' '	sport	IST	ASS	<|eoi|>	what	' '	<|mmu|>	doing	the	' '	is	person
1:	The	person	in	the	the	is	is	is	a	a	a	a	a	a	a	a
2:	The	person	in	the	picture	is	doing	surfing	surfing	surfing	surfing	.	.	.	.	a
3:	The	person	in	the	picture	is	surfing	surfing	,	riding	a	a	riding	a	a	a
4:	The	person	in	the	picture	is	surfing	,	riding	a	wave	wave	<|end|>	.	.	.
5:	The	person	in	the	picture	is	surfing	,	riding	a	wave	on	a	surf	.	.
6:	The	person	in	the	picture	is	surfing	,	riding	a	wave	on	a	surf	board	in

User: What scene does this image depict?

Show-o Turbo:

0:	<|eoi|>	<|eoi|>	does	:	this	depict	'\s'	image	:	scene	ANT	IST	<|soi|>	What	:	<|mmu|>
1:	The	scene	scene	the	a	a	a	a	large	large	large	large	large	large	large	large
2:	The	image	depicts	a	image	of	a	large	large	large	large	large	large	large	large	large
3:	The	image	depicts	a	beautiful	scene	with	with	with	with	with	,	,	large	large	large
4:	The	image	depicts	a	beautiful	scene	of	a	large	,	a	large	large	large	large	large
5:	The	image	depicts	a	beautiful	scene	of	a	large	city	modern	,	,	with	a	a
6:	The	image	depicts	a	beautiful	scene	of	a	large	city	skyline	a	a	,	,	a
7:	The	image	depicts	a	beautiful	scene	of	a	large	city	skyline	with	a	bridge	,	with
8:	The	image	depicts	a	beautiful	scene	of	a	large	city	skyline	with	a	prominent	in	a
9:	The	image	depicts	a	beautiful	scene	of	a	large	city	skyline	with	a	prominent	building	,
10:	The	image	depicts	a	beautiful	scene	of	a	large	city	skyline	with	a	prominent	building	,

兼顾文生图和图生文的能力，只需要4步去噪即可达到**0.625**的**GenEval**分数，超越**SDXL**；在图像理解任务加速**1.5**倍以上

Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

Show-o Turbo：文生图结果

Steps	Model	CFG	GenEval ↑							HPS ↑	IR ↑	CS ↑	Time (sec) ↓
			AVG	TO	CT	P	CL	SO	CA				
16	Show-o	10	**0.674**	0.823	0.647	0.288	0.838	0.984	0.463	**0.277**	**0.992**	**0.318**	1.39
	Show-o	5	0.672	0.778	0.666	0.293	0.835	0.991	0.468	0.270	0.885	**0.318**	1.39
	Show-o Turbo*	0	0.649	0.793	0.644	0.253	0.809	0.956	0.440	0.266	0.768	0.315	0.77
	Show-o Turbo	0	0.646	0.818	0.597	0.218	0.827	0.984	0.430	0.273	0.925	**0.318**	0.77
8	Show-o	10	0.578	0.631	0.519	0.235	0.811	0.991	0.280	0.257	0.672	0.313	0.76
	Show-o	5	0.580	0.647	0.584	0.225	0.766	0.984	0.275	0.255	0.632	0.313	0.76
	Show-o Turbo*	0	**0.642**	0.788	0.631	0.253	0.787	0.981	0.413	0.264	0.800	0.315	0.46
	Show-o Turbo	0	0.638	0.813	0.541	0.250	0.814	0.991	0.420	**0.273**	**0.963**	**0.318**	0.46
4	Show-o	10	0.353	0.237	0.325	0.095	0.540	0.863	0.060	0.197	-0.560	0.283	0.44
	Show-o	5	0.396	0.298	0.334	0.158	0.572	0.925	0.088	0.207	-0.300	0.294	0.44
	Show-o Turbo*	0	0.596	0.692	0.553	0.218	0.758	0.978	0.375	0.249	0.633	0.312	0.30
	Show-o Turbo	0	**0.625**	0.770	0.553	0.245	0.806	0.978	0.398	**0.269**	**0.934**	**0.318**	0.30
2	Show-o	10	0.181	0.025	0.131	0.008	0.327	0.588	0.008	0.140	-1.756	0.246	0.29
	Show-o	5	0.251	0.051	0.188	0.038	0.442	0.778	0.010	0.152	-1.456	0.260	0.29
	Show-o Turbo*	0	0.459	0.407	0.422	0.148	0.668	0.925	0.185	0.201	-0.259	0.295	0.22
	Show-o Turbo	0	**0.557**	0.614	0.478	0.180	0.793	0.972	0.305	**0.247**	**0.680**	**0.312**	0.22

Table 1. **Comparison of T2I performance at the resolution of 512 × 512 based on GenEval, HPS, IR, and CS.** AVG: average, TO: Two Object, CT: Counting, P: Position, CL: colors, SO: Single Object, CLA: Color Attr.

不带CFG的4步生成超过原始模型带CFG的8步生成，

实际加速接近3倍

Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

Show-o Turbo： inpainting结果

User : In the distance, a small white sailboat was parked between the mountains and the water.




Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

Show-o Turbo： extrapolation结果

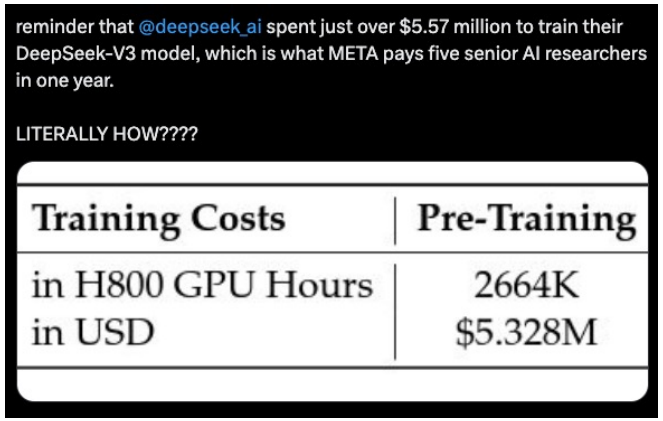
User : The mountains and jungles are covered with thin mist.

User : A serene natural land-scape featuring a clear lake surrounded by lush trees.

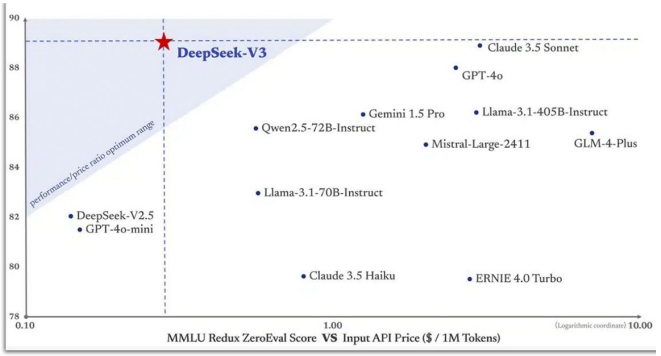
Xu, Wang, Liao, Li, Hou, Deng. Show-o Turbo: Towards Accelerated Unified Multimodal Understanding and Generation

2、大模型推理加速

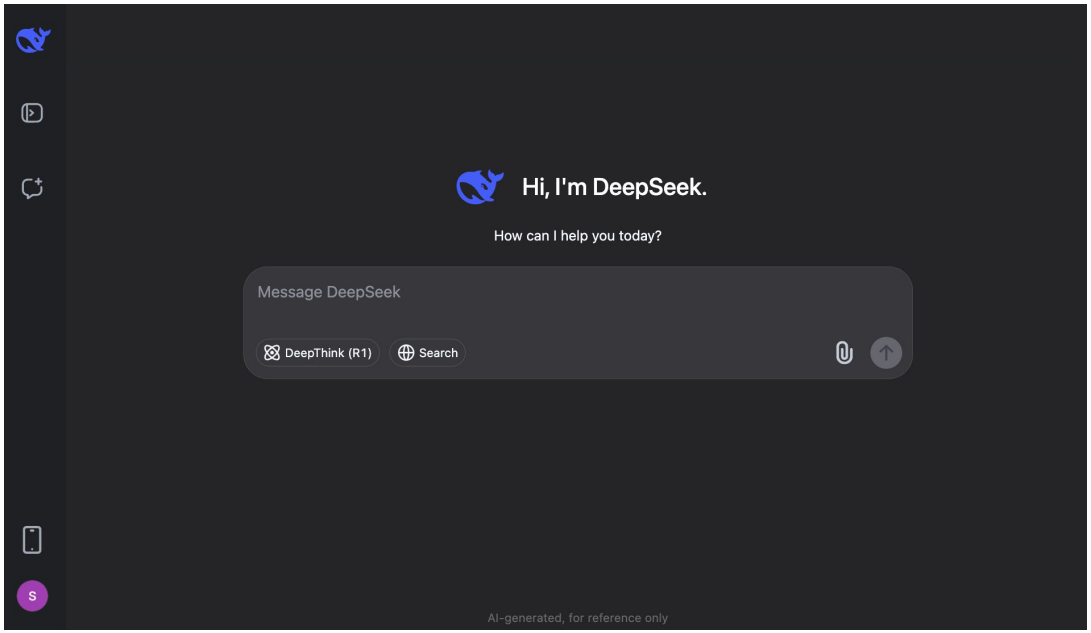
DeepSeek给我们上的一课：只有模型好不够，降低成本才是应用的关键



DeepSeek v3的训练成本仅等于META五位研究员的年薪



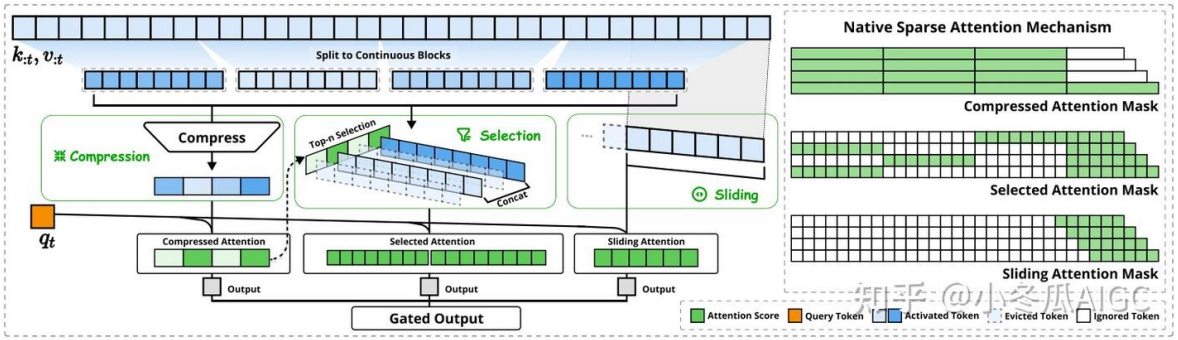
DeepSeek v3的API调用成本也大大低于几大公司竞品



OpenAI o1: $60.00 per 1M output tokens
DeepSeek R1: $2.19 per 1M output tokens

DeepSeek如何降低成本?

- 模型&算法侧: **DeepSeekMoE & MLA & NSA**



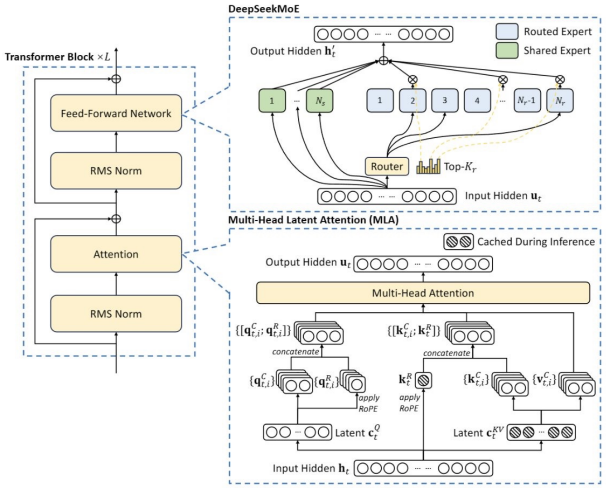
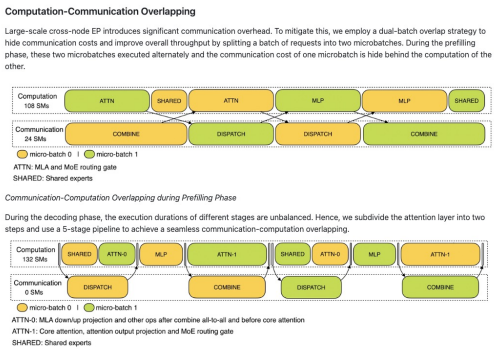
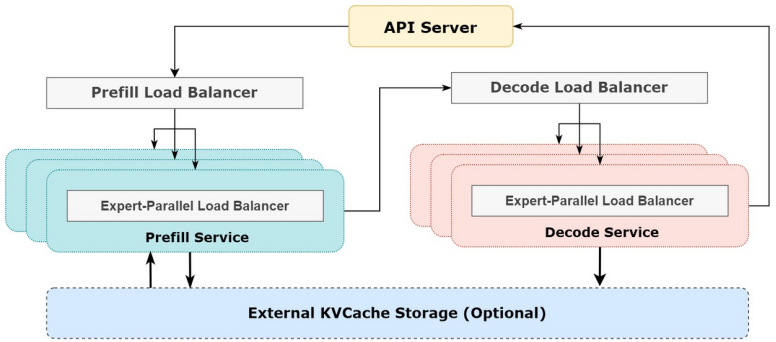
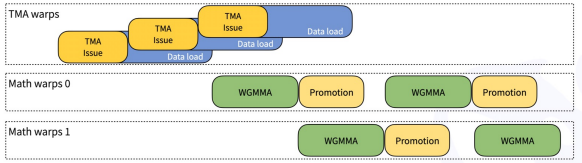


Figure 1 | Illustration of the architecture of DeepSeek-V2. MLA ensures efficient inference by significantly reducing the KV cache for generation, and DeepSeekMoE enables training strong models at an economical cost through the sparse architecture.

- 底层实现侧: 专家并行、计算/通信重叠、负载均衡、极致代码优化、etc.

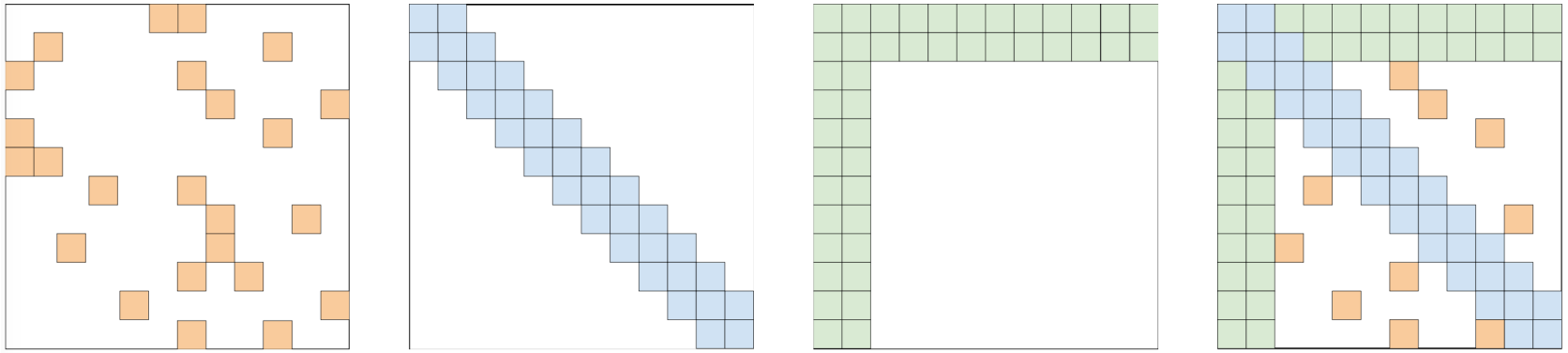






注意力稀疏化

- 标准注意力机制的瓶颈主要在平方的计算复杂度 & 线性的**KV-Cache**存储
- 注意力机制的稀疏化通过引入结构化的遮掩范式，从而降低计算复杂度/显存需求



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

一些经典策略：随机采样、局部窗口、全局 token，以及结合它们

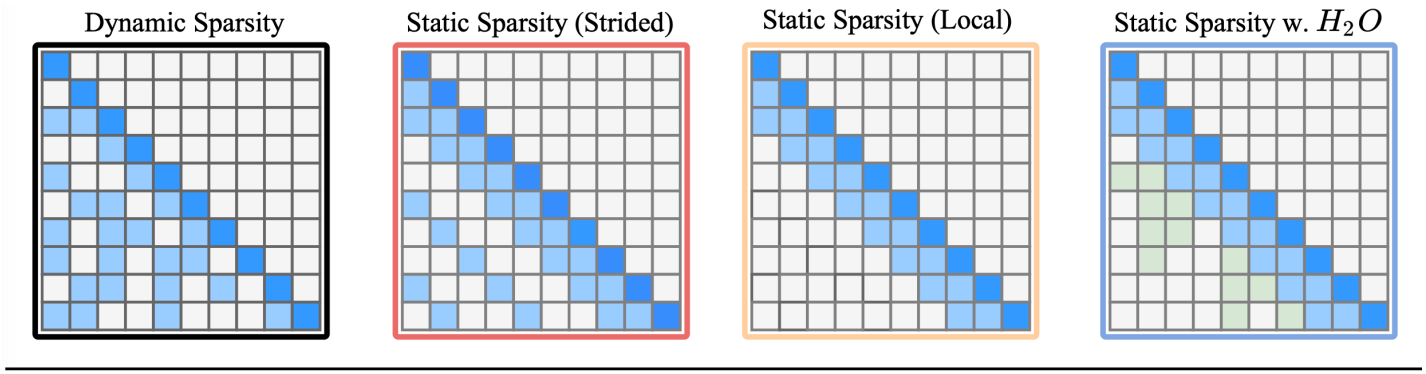
Child, Rewon, et al. "Generating long sequences with sparse transformers." arXiv:1904.10509 (2019).

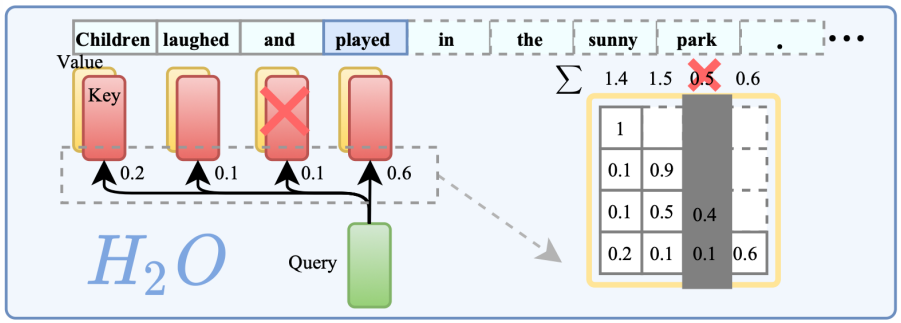
Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." Neurips 2020.

Xiao, Guangxuan, et al. "Efficient streaming language models with attention sinks." arXiv:2309.17453 (2023).

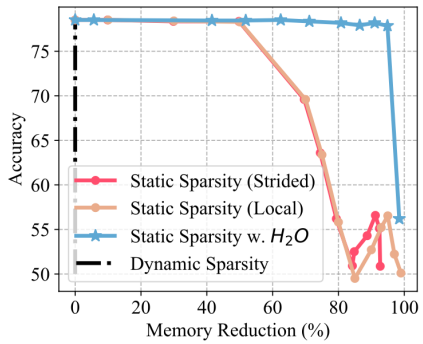
注意力稀疏化：从静态向动态转化

- **H2O**：在**decoding**阶段动态识别关键**token**（**Heavy Hitters**）





基于累计注意力分数



至多90%的**memory**减少

Zhang, Zhenyu, et al. "H2o: Heavy-hitter oracle for efficient generative inference of large language models." Neurips 2023.

注意力稀疏化：从局部向全局转化

H2O：基于局部累积注意力得分

↓

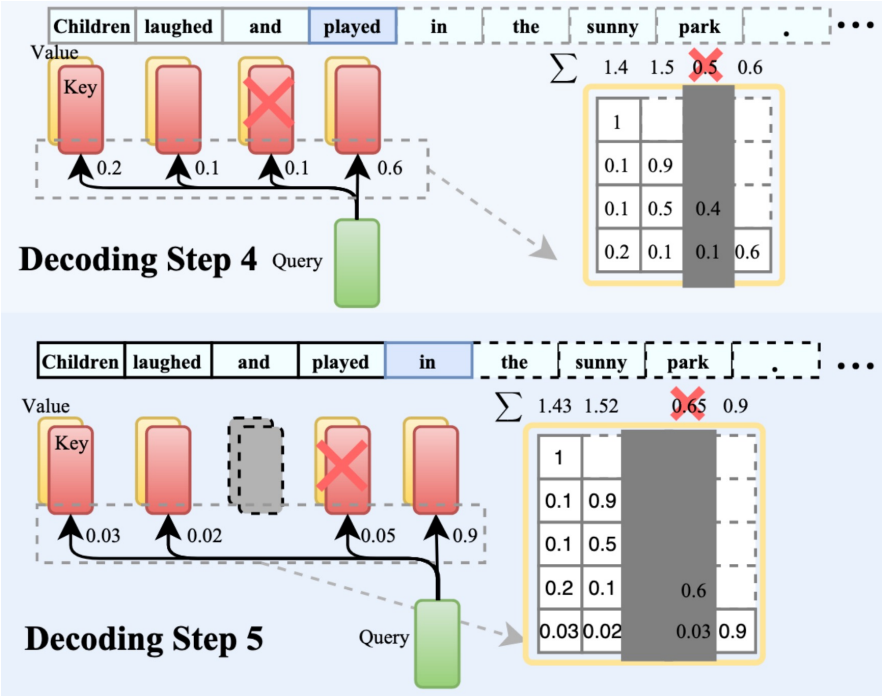
采用局部信息不能最优判断丢弃对象，
导致 **attention bias** 问题

↓

如何才能采用全局信息来判断丢弃对象

↓

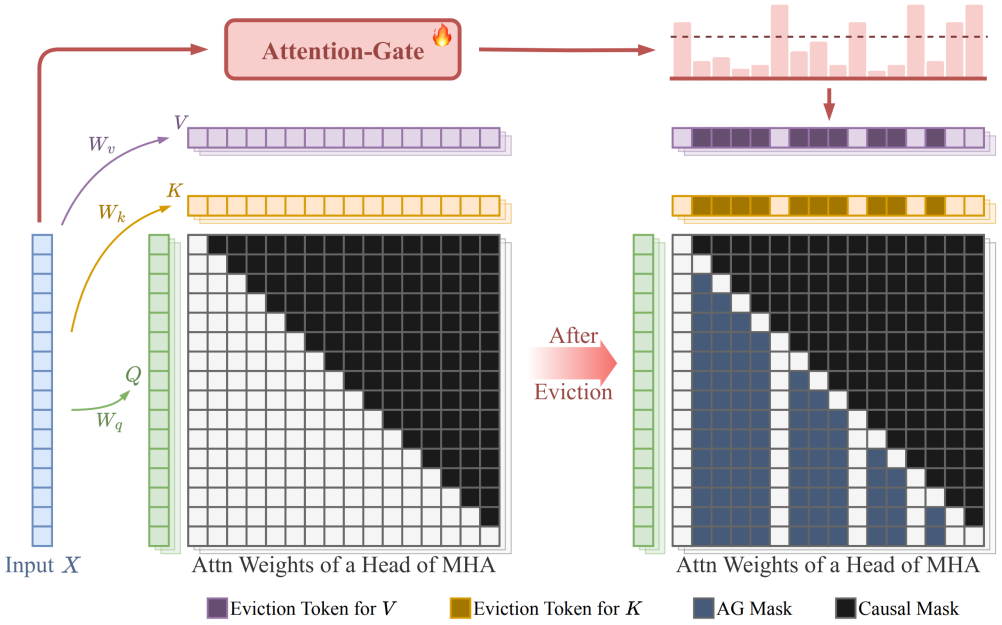
Attention-Gate !



H2O 采用局部的累积注意力得分来判断丢弃对象
即：之前丢弃的 **token**，后续评估将不再考虑

Zhang, Zhenyu, et al. "H2o: Heavy-hitter oracle for efficient generative inference of large language models." Neurips 2023.

Attention-Gate：可后训练的稀疏自注意力

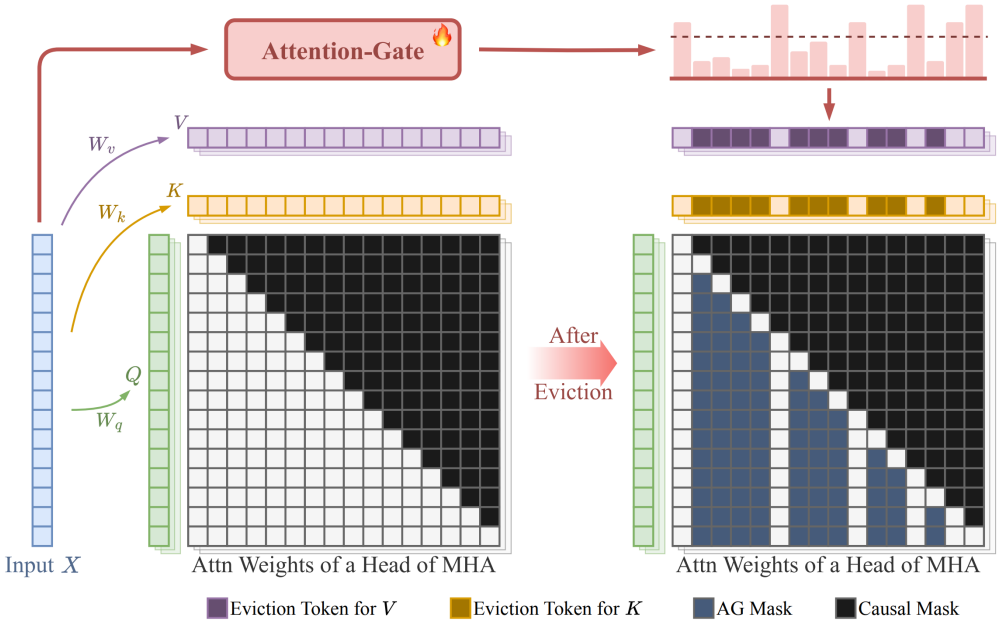


- 稀疏自注意力需要是**in-context**学习的！
  - 基于一个light attention生成主attention的掩码
  - 使用Straight-Through Estimator实现二值化决策的梯度回传，支持端到端微调
  - 每个head、layer都生成单独的掩码

- **AG**有4个**head** vs. 原**LLaMA-2-7B**有32个**head**

Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

Attention-Gate：可后训练的稀疏自注意力



- 控制稀疏率

$$\ell_{\text{evict}} = \alpha \cdot \left| \overline{\text{AG}} - \beta \right|$$

  - 让所有AG模块的平均预测接近预设阈值 β

Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

Attention-Gate：结果

Table 1: Performance comparison of Llama2-7B and various KV-Cache eviction strategies *after continual pre-training*. For baselines, (W_q, W_k, W_v, W_o) are made trainable, while in our method, the AG module is also trainable. Higher values indicate better performance for all metrics. Acc. refers to accuracy. %Evict. refers to the mean KV-Cache eviction ratio, representing the percentage of tokens evicted from KV-Cache. The eviction ratio is fixed at 50% for the baseline methods. In contrast, our method achieves better performance (average accuracy and score) while maintaining a higher average %Evict..

	Metric	PIQA	ARC-C	ARC-E	RTE	COPA	BoolQ	HellaSwag	MMLU	Avg.	Metric	LongBench
Llama2-7B-cpt	Acc.	72.69	32.88	50.62	50.54	57.00	64.77	42.19	26.64	49.67	Score	23.42
StreamingLLM	Acc.	72.42	31.53	**49.74**	50.90	54.00	61.31	37.75	26.66	48.04	Score	4.61
H2O	Acc.	72.20	30.85	49.38	**51.99**	55.00	**62.42**	41.45	26.45	48.72	Score	4.85
Ours	Acc.	**76.33**	**32.20**	48.32	50.18	**59.00**	60.46	**64.23**	**28.54**	**52.41**	Score	**13.71**
	%Evict.	43.12	46.54	45.15	48.60	**55.37**	**50.16**	**61.10**	**70.36**	**52.55**	%Evict.	**68.55**

基于4张4090进行持续预训练，实现52.55% **attention**稀疏率，和~3%的平均性能提升

Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

Attention-Gate：结果

- 显著减少**prefilling**的**peak memory**和时间开销

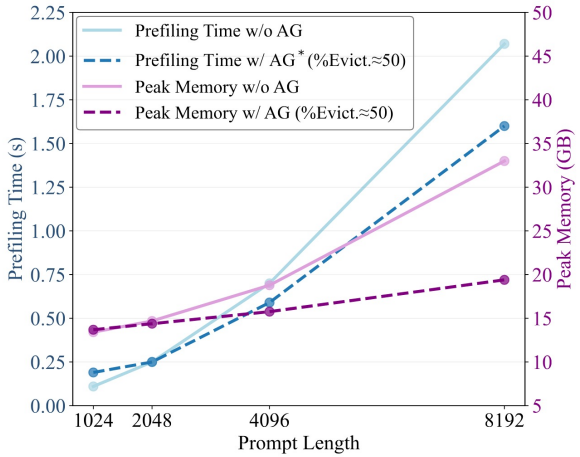


Figure 3: Comparison of peak memory usage and prefilling time between the LLaMA2-7B model (without AG) and the proposed implementation (with AG and ∼50% eviction) across varying prompt lengths. The results show significant improvements in memory efficiency with AG, especially as prompt length increases. Prefilling time is not the primary focus, and the current implementation (marked with * in the legend) relies on a suboptimal for-loop over attention heads. Even so, the method maintains stable prefilling time and shows a clear reduction trend with longer prompts.

Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

Attention-Gate：结果

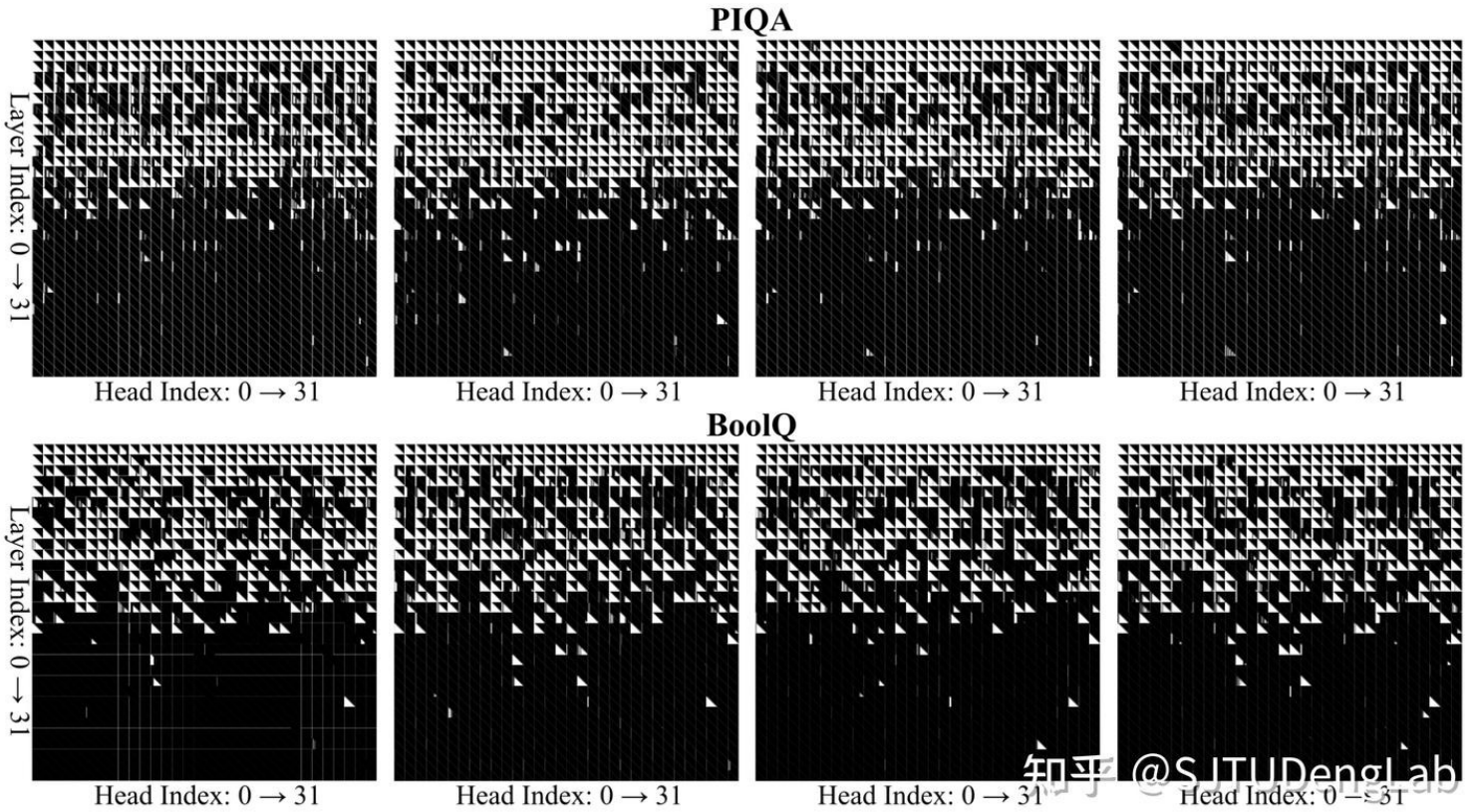
- AG可以非常轻量化，甚至可以pre-compute（跟前一层的自注意力计算并行）

		Metric	PIQA	ARC-C	RTE	COPA	BoolQ	OBQA	Avg.
Vanilla AG	(1)	Acc.	82.15	59.66	64.26	93.00	86.82	78.80	77.45
		%Evict.	66.16	48.31	65.47	45.40	67.46	67.17	60.00
减少head数目	(2-1)	Acc.	81.88	57.63	65.70	91.00	87.52	77.40	76.86
		%Evict.	63.92	36.38	62.73	24.38	65.22	63.57	52.70
	(2-2)	Acc.	82.15	53.90	62.45	89.00	87.31	77.40	75.37
		%Evict.	58.97	31.47	59.77	20.32	63.02	59.17	48.79
降低dim	(3-1)	Acc.	81.45	53.36	58.84	88.00	86.73	78.40	74.46
		%Evict.	61.75	33.55	61.34	19.24	64.59	59.59	50.01
	(3-2)	Acc.	83.03	53.90	59.93	89.00	87.16	76.40	74.90
		%Evict.	58.68	24.23	32.23	12.28	59.40	55.54	40.39
使用前1层预测	(4-1)	Acc.	81.66	55.25	66.06	88.00	86.85	78.00	75.97
		%Evict.	49.52	36.92	46.85	28.74	56.02	60.32	46.40
	(4-2)	Acc.	82.75	55.93	79.06	82.00	86.33	78.40	77.41
		%Evict.	53.31	44.38	51.20	47.95	61.98	61.73	53.43
AG换成线性gate	(5)	Acc.	82.54	54.58	57.40	81.00	87.71	74.80	73.01
		%Evict.	1.06	0.46	0.81	0.26	1.38	1.16	0.86

Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

Attention-Gate：结果

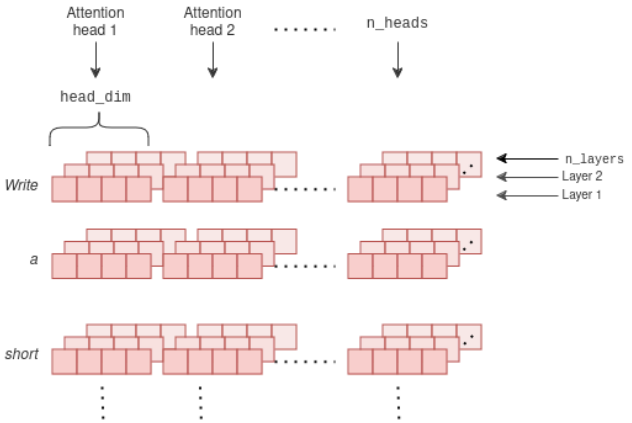
- 模型更深的层可以被更显著稀疏化、不同任务、层、头的稀疏模式不一

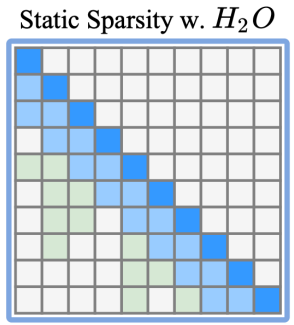


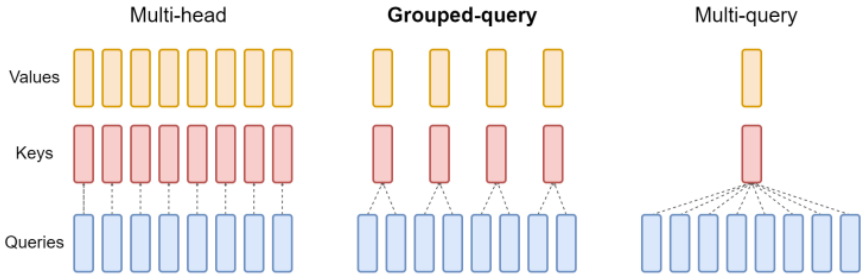
Zeng et al. IN-CONTEXT KV-CACHE EVICTION FOR LLMS VIA ATTENTION-GATE. https://arxiv.org/pdf/2410.12876

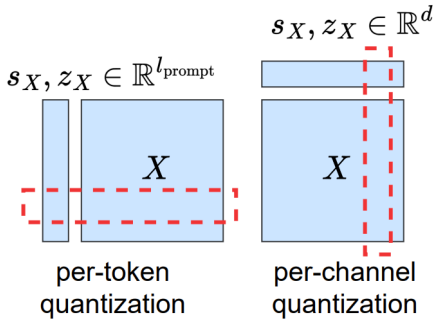
KV Cache压缩的多个维度

- **KV Cache: num_layers * num_heads * seq_len * feature_dim**
  - 针对序列长度：H2O、FastGen、etc.
  - 针对注意力头数量：如MQA和GQA
  - 针对存储精度：量化，如KIVI实现2-bit量化









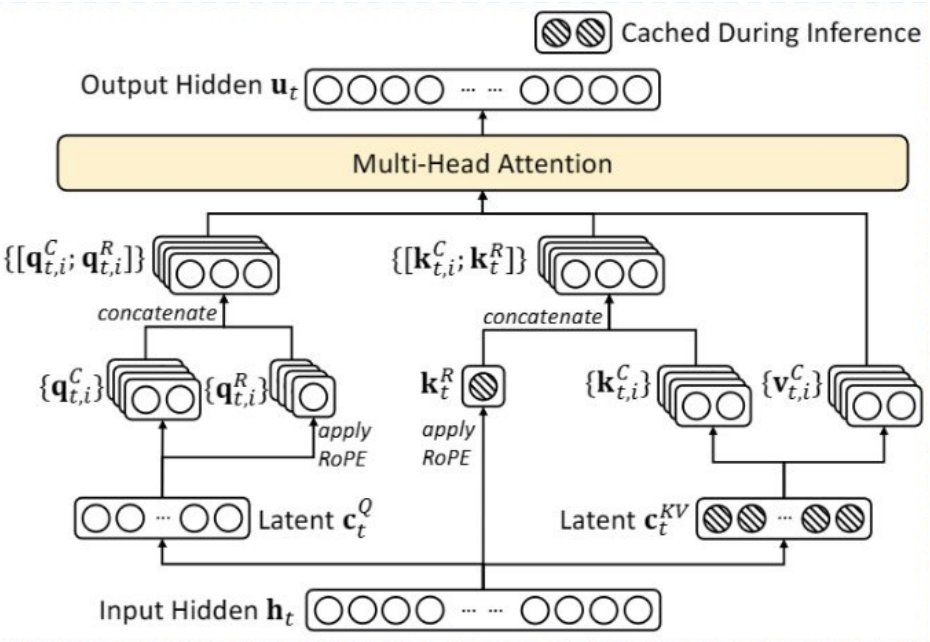
能否针对特征维度？

KV Cache特征维度的压缩

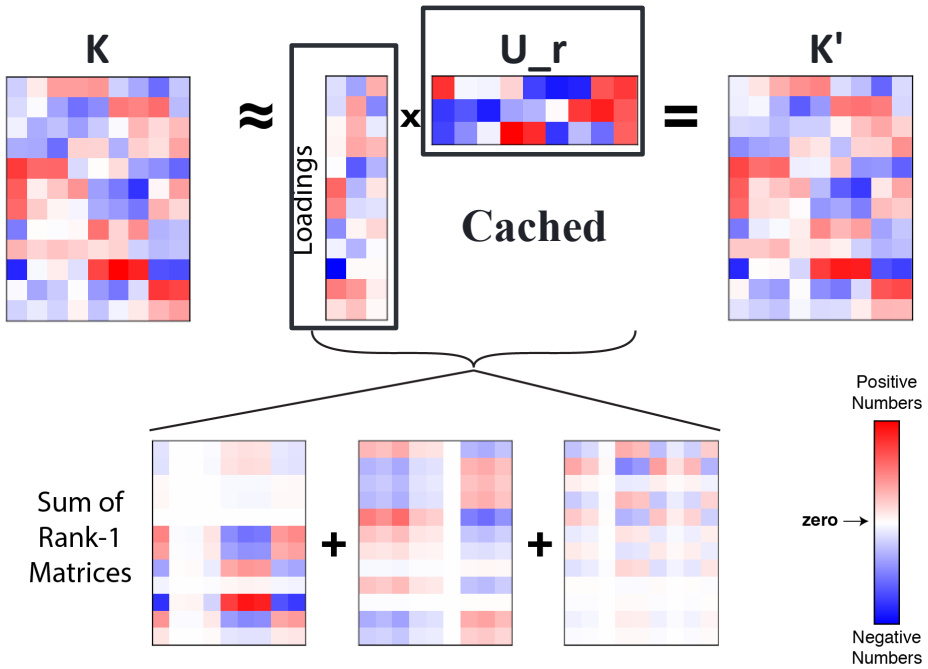
😃 **Multi-head Latent Attention (MLA)**

- 提升 DeepSeek 模型 KV Cache 的效率，降低通信开销，提高推理速度。
- 😢 需要昂贵的预训练才能实现。

🤔 如何以低成本为其他开源模型（比如**LLaMA, Mistral**）引入类 **MLA** 机制？



KV Cache特征维度的压缩：对K/V做PCA？

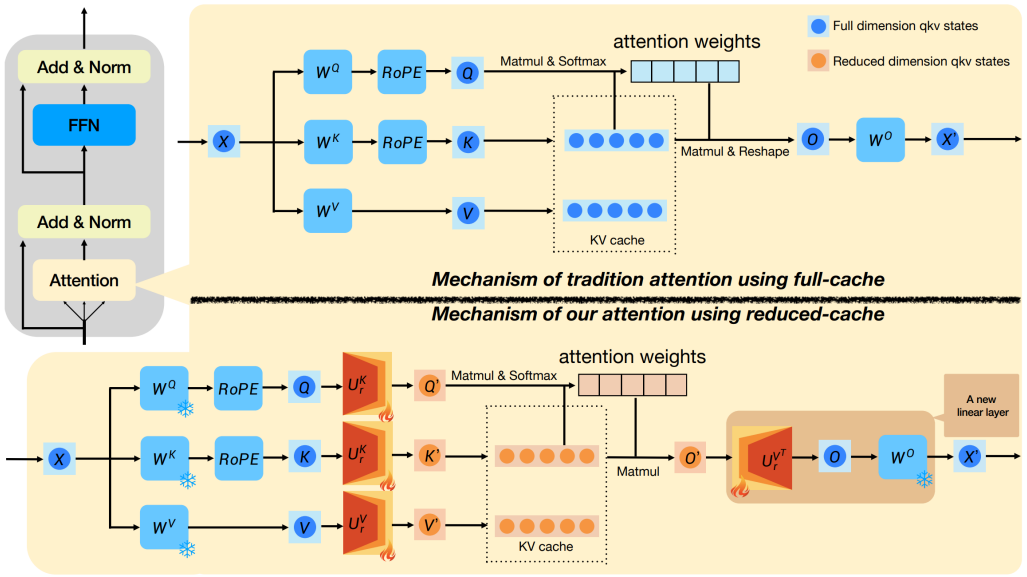


- 然而，**PCA**只在局部、线性的情况下最优
  - Transoformer中，多层自注意力非线性耦合，PCA的近似误差被显著放大

Lin et al. MATRYOSHKAKV: ADAPTIVE KV COMPRESSION VIA TRAINABLE ORTHOGONAL PROJECTION. ICLR 2025.

MatryoshkaKV: 面向预训练模型的MLA

- 正交特征投影矩阵
  - 参考PCA，最大化特征利用效率
- 数据驱动的持续预训练
  - 套娃学习，保证投影矩阵列间的主次关系
- 推理时压缩率搜索
  - 基于calibration dataset，搜索不同层、不同head的最大投影矩阵“秩”，从而更好地适应不同层和头的需求。



Lin et al. MATRYOSHKAKV: ADAPTIVE KV COMPRESSION VIA TRAINABLE ORTHOGONAL PROJECTION. ICLR 2025.

MatryoshkaKV: 结果

- 一个训练后的模型，可以面向不同的budget，利用不同的压缩率
- 不压缩时保持与原模型相同的性能
- 压缩50%维度，性能下降3%以内

Model	Budget	Method	HLSG	ARCC	ARCE	PIQA	WG	CSQA	Avg.
LLaMA2 7B-base	100.0%	Baseline	74.00	35.93	50.97	78.50	61.64	65.93	61.16
		PCA	72.04	36.95	52.38	76.66	61.72	67.24	61.17
		MKV	72.05	37.29	52.38	76.66	61.72	67.32	61.24
	87.5%	PCA	71.91	35.93	53.97	76.66	61.40	67.65	61.25
		MKV	71.58	37.97	53.26	75.95	62.12	69.57	61.74
	75.0%	PCA	70.99	35.59	54.14	76.22	60.06	66.99	60.67
		MKV	71.58	38.31	55.56	76.01	61.09	66.75	61.55
	62.5%	PCA	67.16	34.24	54.85	74.76	57.77	61.10	58.31
		MKV	68.03	37.97	56.08	75.12	60.30	65.44	60.49
	50.0%	PCA	42.11	29.83	35.10	58.16	52.57	40.62	43.07
		MKV	66.78	36.61	55.91	74.32	59.12	61.92	59.11
	37.5%	PCA	24.24	26.44	26.63	51.25	50.36	19.90	33.14
		MKV	63.97	33.90	51.68	74.97	57.92	59.21	56.94
	25.0%	PCA	23.98	29.49	26.28	51.20	50.36	16.22	32.92
		MKV	51.91	27.46	44.44	69.64	54.54	44.39	48.73
Mistral-v0.3 7B-base	100.0%	Baseline	75.50	42.03	63.14	80.25	65.43	70.68	66.17
		PCA	75.46	42.03	62.96	80.25	65.35	70.27	66.05
		MKV	75.44	42.03	62.96	80.25	65.51	70.27	66.08
	87.5%	PCA	73.46	42.71	63.32	79.54	63.93	70.76	65.92
		MKV	75.63	42.03	64.37	79.71	65.51	70.35	66.27
	75.0%	PCA	70.75	37.63	61.73	78.18	62.59	68.47	63.23
		MKV	75.29	43.39	63.14	79.54	64.96	69.12	65.90
	62.5%	PCA	63.48	34.24	55.73	75.90	60.77	62.24	58.73
		MKV	74.23	40.34	62.96	79.33	64.25	68.63	64.96
	50.0%	PCA	28.12	22.71	28.40	58.16	49.64	22.85	34.98
		MKV	73.32	38.98	62.08	79.16	61.88	67.08	63.75
	37.5%	PCA	25.04	22.03	28.04	53.86	49.25	21.21	33.24
		MKV	70.40	35.93	58.91	77.91	60.30	64.29	61.29
	25.0%	PCA	24.91	26.10	25.40	52.67	48.30	19.74	32.85
		MKV	59.21	25.42	48.68	73.83	54.30	45.13	51.10

Lin et al. MATRYOSHKAKV: ADAPTIVE KV COMPRESSION VIA TRAINABLE ORTHOGONAL PROJECTION. ICLR 2025.

MatryoshkaKV: 结果

- 训练完的模型，在不同数据集上搜索出来的不同层、不同**head**上的最大压缩率具有显著区别
  - 深层压缩率高

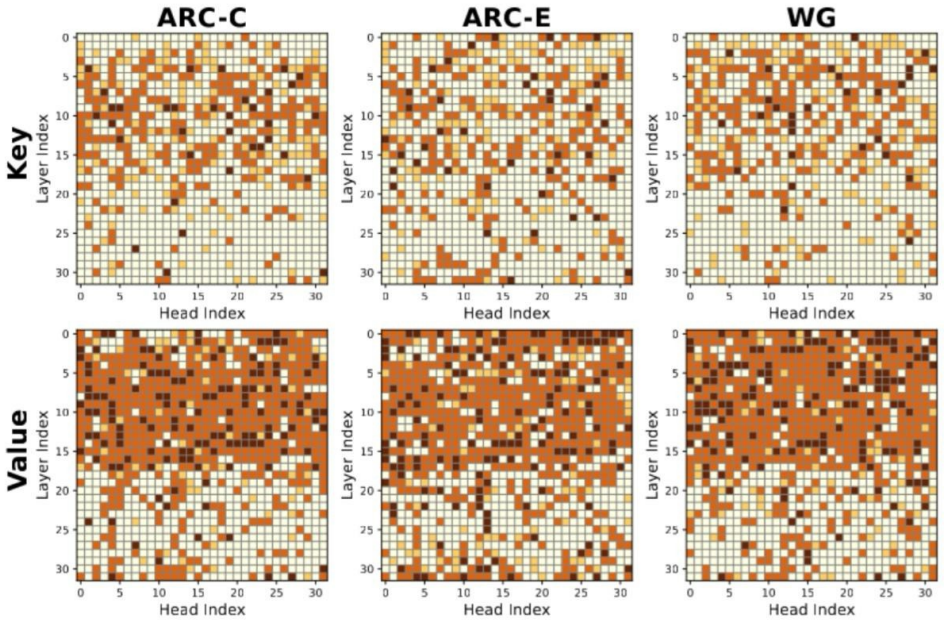


Figure 1: Visualization of the feasible compression level for the keys and values in our model distilled from the LLaMA2-7B-base model. We individually leverage samples in ARC-challenge (ARC-C), ARC-easy (ARC-E) (Clark et al., 2018), and Winogrande (WG) (Sakaguchi et al., 2019) to determine the compression level. Lighter colors indicate higher compression levels. As shown, our approach enables the use of various compression strategies for various tasks.

Lin et al. MATRYOSHKAKV: ADAPTIVE KV COMPRESSION VIA TRAINABLE ORTHOGONAL PROJECTION. ICLR 2025.

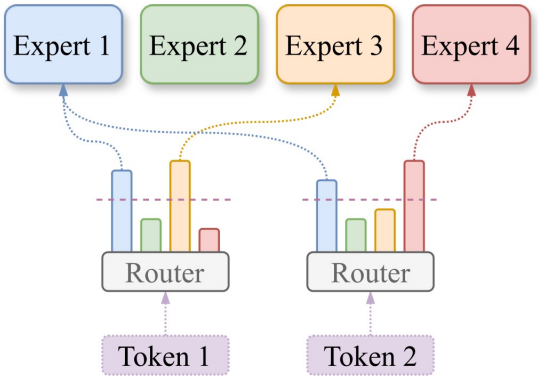
MoE架构将是大模型继续scaling的方向？

- **MoE** 架构 $y = \sum_{i=1}^{n} G(x)_i \, E_i(x)$

  一个路由 + 若干专家

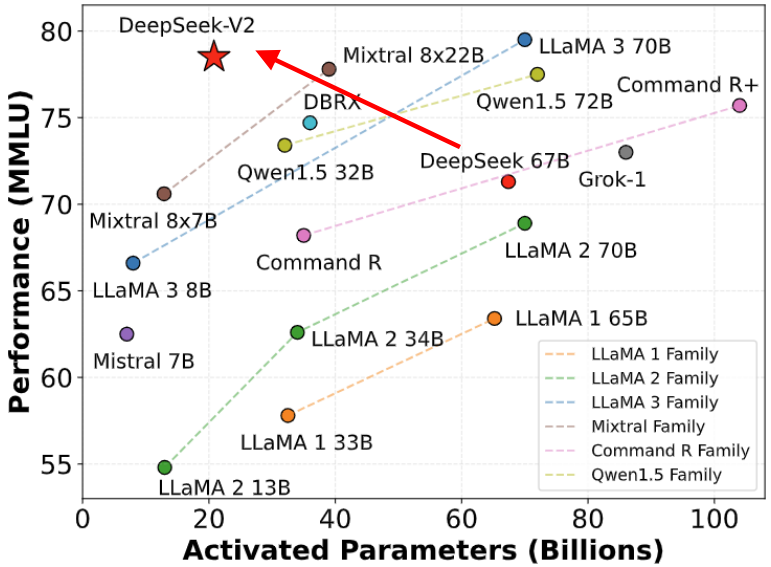
  每次选择路由打分最高的 k 个专家

$$G(x) = \text{Softmax}\left(\text{TopK}\left(x \cdot W_g, k\right)\right)$$



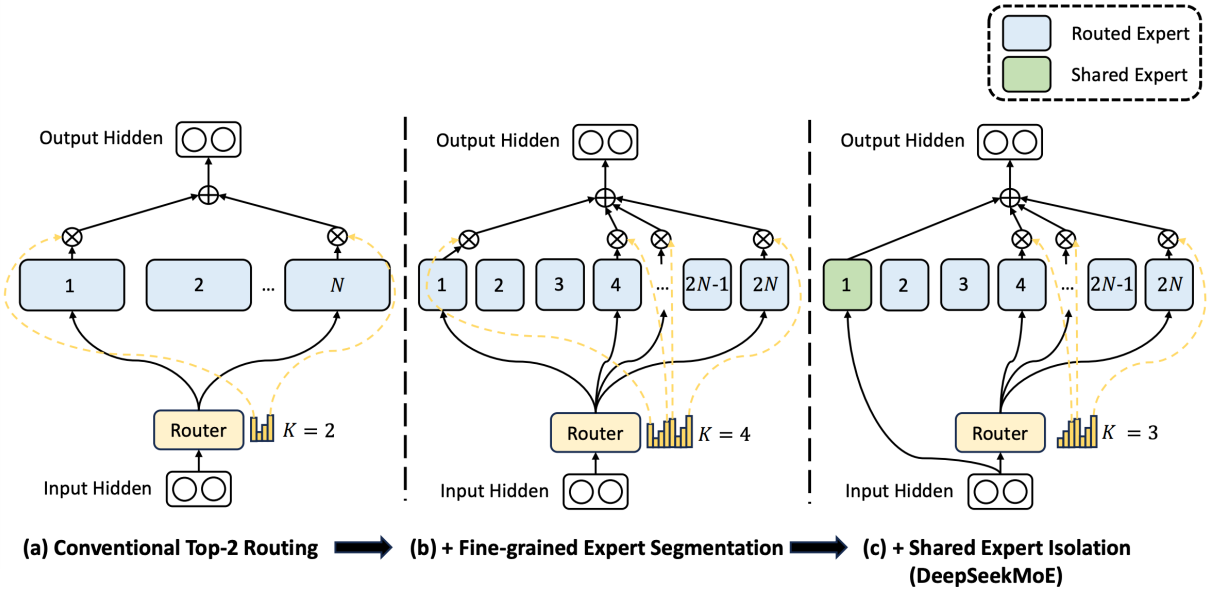
- **MoE** 模型相比 **dense** 模型的优势
  - 只激活小部分参数，推理计算量更低，速度更快
  - 在相同的计算预算下，MoE 性能表现更优

- 最新进展：
  - Llama 4、DeepSeek-V3/R1
  - 商汤 SenseNova V6...



优化 MoE 架构的效率

- **DeepSeekMoE**的精髓在于
  共享专家+领域专家的组合



但是，为什么坚持使用top-k路由机制，将不同token都交给固定数目的专家来处理？

- 直觉上，easy token可以简单地被少量专家处理，hard token才需要更多专家

优化 MoE 架构的效率

- 但是，为什么坚持使用top-k路由机制，将不同token都交给固定数目的专家来处理？
  - 在SocialIQA数据集上，Mixtral-8x7B不同层里 不同tokens的路由概率差距明显

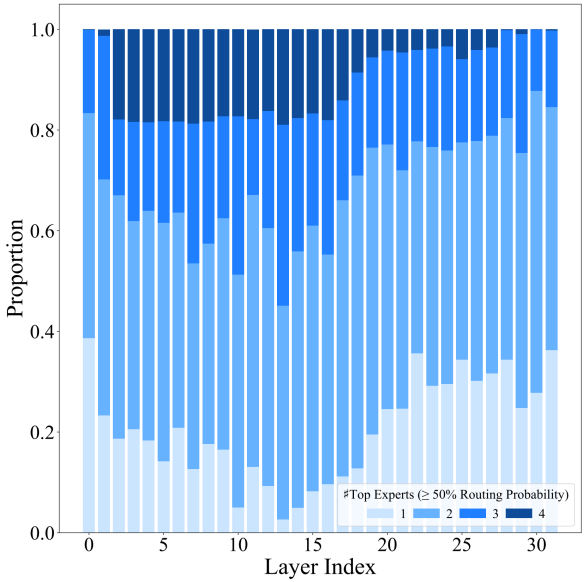
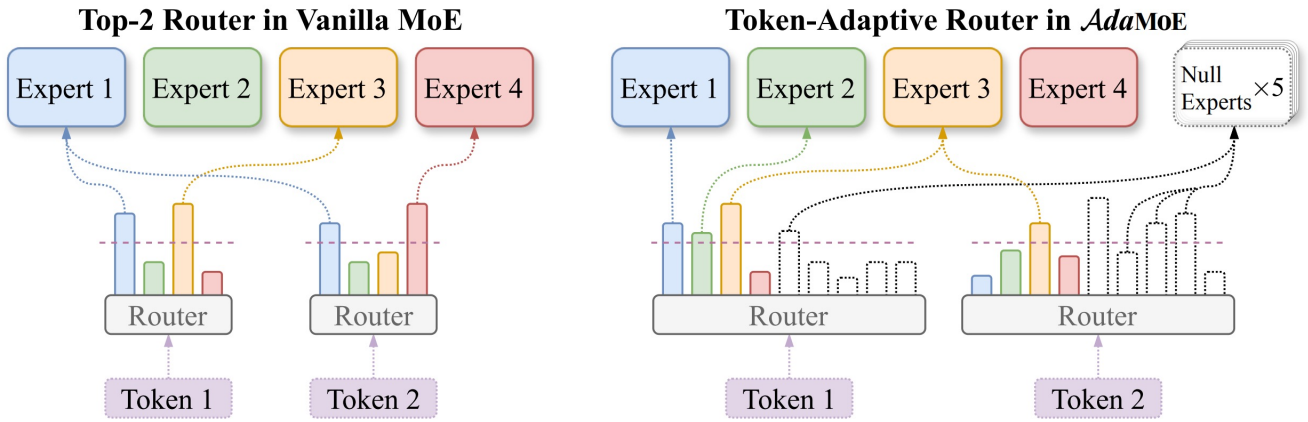


Figure 3: Proportions of the number of top experts with cumulative routing probabilities exceeding 50% for tokens in the SocialIQA dataset. Each bar represents the proportion of different counts of tokens at the corresponding MoE layer in Mixtral-8x7B.

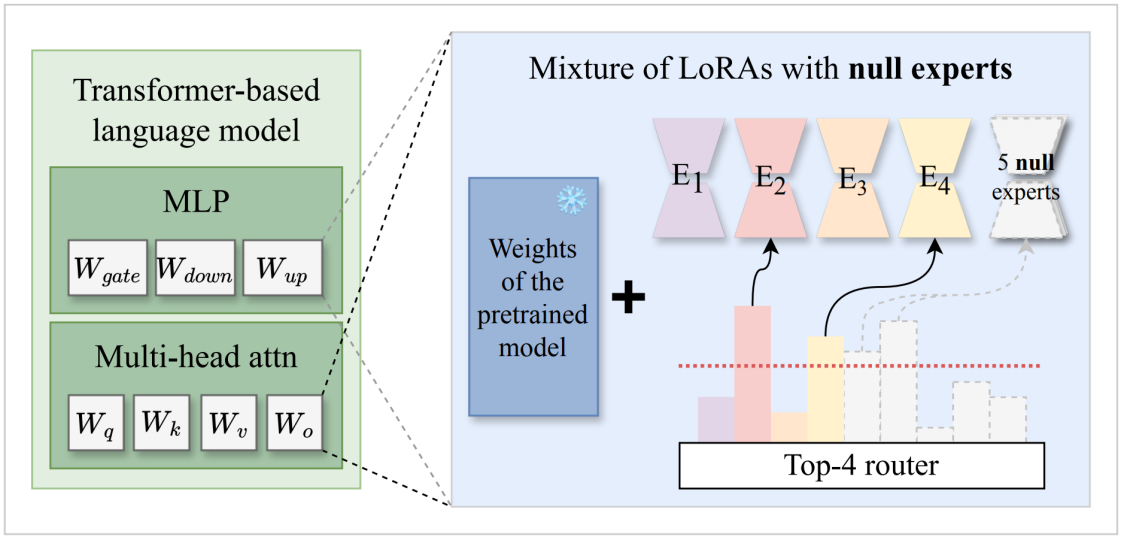
AdaMoE：动态MoE，引入空专家，不同tokens可选不同数目的专家



- 引入空专家，沿用**top-k**选择，可实现不同**tokens**选不同数目的真专家
- 通过后训练的方式即可提高预训练**MoE**模型（例如：**Mixtral-8x7B**）的计算效率

AdaMOE: Token-Adaptive Routing with Null Experts for Mixture-of-Experts Language Models. Zeng, Miao, Gao, Zhang, Deng. EMNLP 2024 Findings.

AdaMoE：动态MoE，拓展到Mixture of LoRA



- 负载均衡损失：对所有空专家一视同仁

$$\ell_{null} = \alpha \cdot (n+m) \cdot \sum_{i=1}^{n+m} \tilde{f}_i \cdot P_i \qquad \tilde{f}_i = \begin{cases} f_i & \text{if } i \leq n \\ \frac{1}{m} \sum_{j=n+1}^{n+m} f_j & \text{if } i > n \end{cases}$$

AdaMOE: Token-Adaptive Routing with Null Experts for Mixture-of-Experts Language Models. Zeng, Miao, Gao, Zhang, Deng. EMNLP 2024 Findings.

AdaMoE: 结果

	Metric	WINO	HELLA	PIQA	SIQA	OQA	ARC-C	Avg.
Original Mixtral-8x7B	Acc.	55.96	53.62	68.06	64.59	65.40	83.73	65.23
Fine-tuned Mixtral-8x7B	Acc.	80.43	84.10	**90.48**	76.36	**89.00**	87.46	84.64
*Ada*MoE	Acc.	**81.93**	**85.50**	90.32	**76.97**	88.20	**89.15**	**85.35**
	%FLOPs↓	14.99	14.10	18.07	16.31	13.22	14.55	15.21
	Load	1.66	1.68	1.59	1.63	1.70	1.67	1.66

Table 1: Comparison of performance and computational efficiency across six datasets: WINO, HELLA, PIQA, SIQA, OQA and ARC-C. Metrics include Acc. (accuracy), %FLOPs↓ (percentage of FLOPs reduction by *Ada*MoE compared to the baselines), and Load (the average number of experts used per MoE/*Ada*MoE layer). The baselines are original/fine-tuned Mixtral-8x7B, both using the top-2 routing strategy (Load = 2.00). *Ada*MoE not only reduces FLOPs but also achieves better accuracy across most datasets compared to the fine-tuned Mixtral-8x7B with LoRA.

在**ARC-C**上，将**Mixtral-8x7B**的**FLOPs**降低**14.5%,** 准确率提升**1.69%**

AdaMOE: Token-Adaptive Routing with Null Experts for Mixture-of-Experts Language Models. Zeng, Miao, Gao, Zhang, Deng. EMNLP 2024 Findings.

AdaMoE: 结果

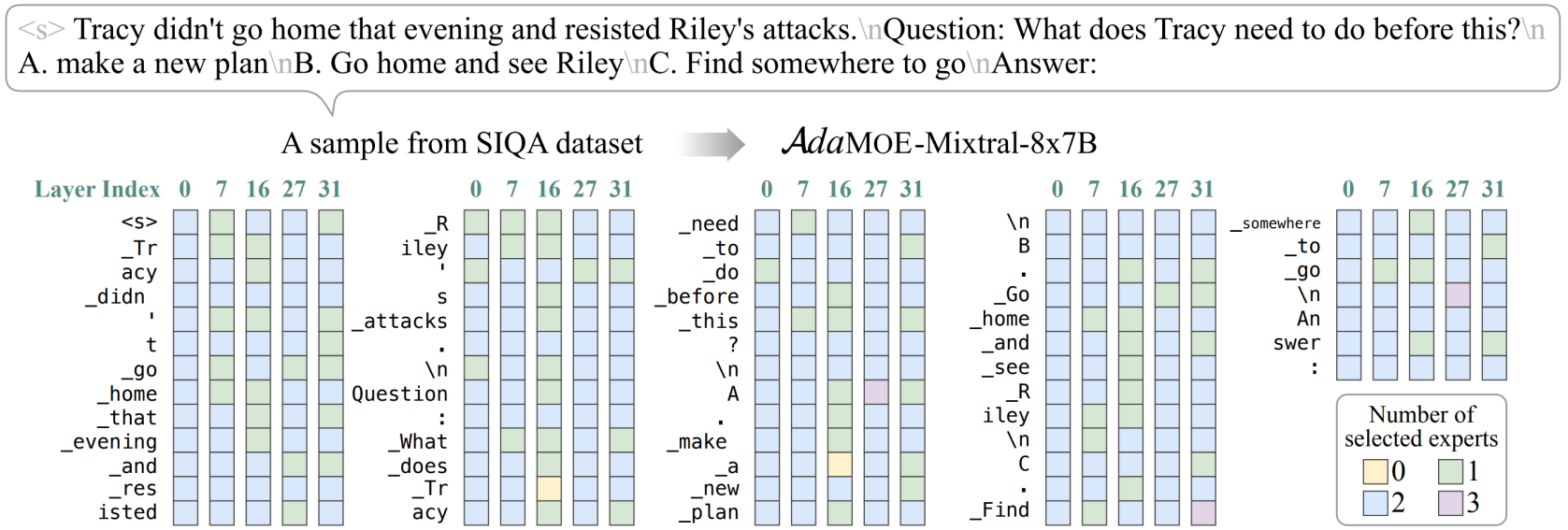


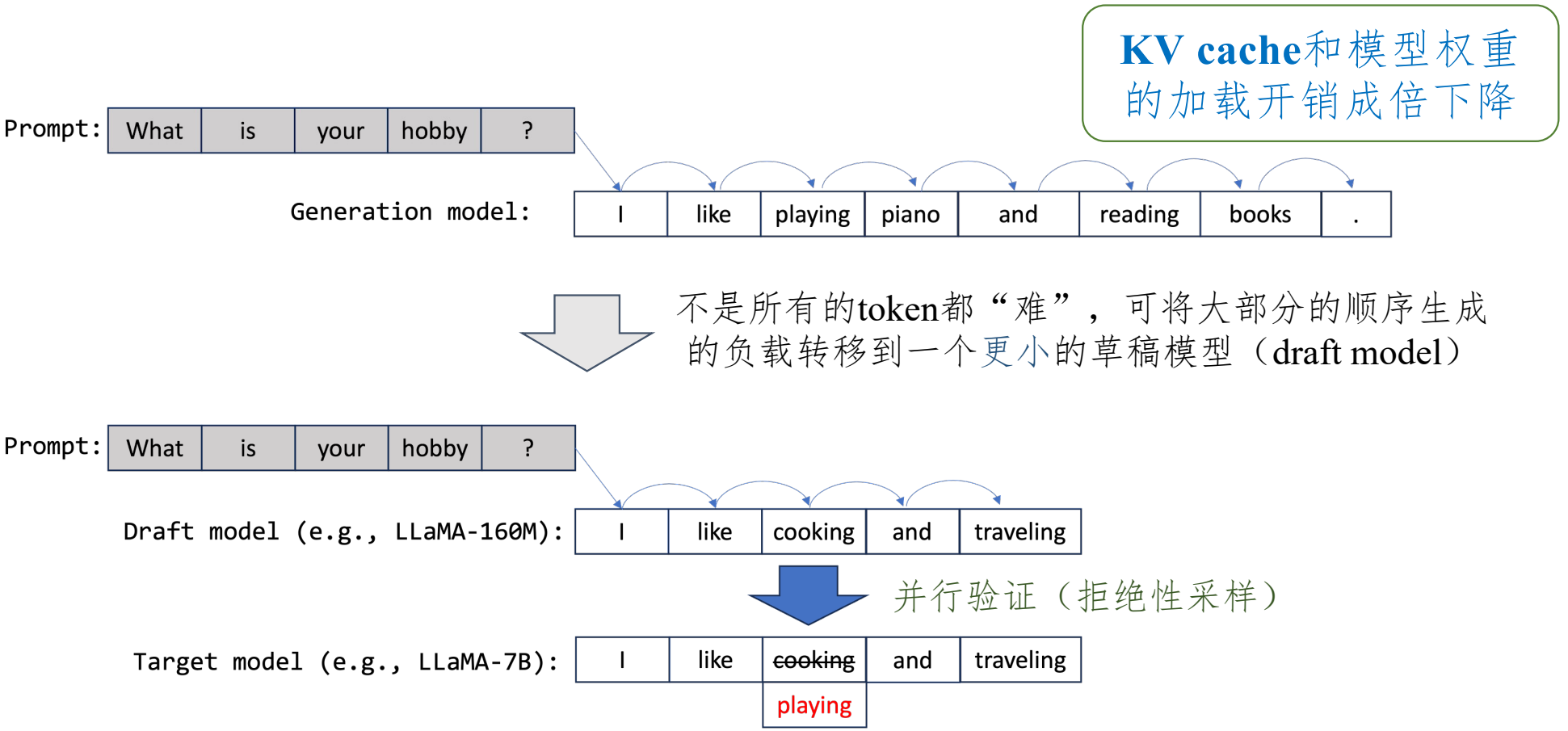
Figure 1: The number of selected experts for various tokens in an *Ada*MoE variant of Mixtral-8x7b. As shown, after applying *Ada*MoE, the model possesses the ability to perform token-adaptive routing. Also note that some tokens only require 1 expert for feature abstraction, which offers the opportunity for inference acceleration.

「关键**token**重计算，平凡**token**轻处理」

AdaMOE: Token-Adaptive Routing with Null Experts for Mixture-of-Experts Language Models. Zeng, Miao, Gao, Zhang, Deng. EMNLP 2024 Findings.

算法侧推理加速

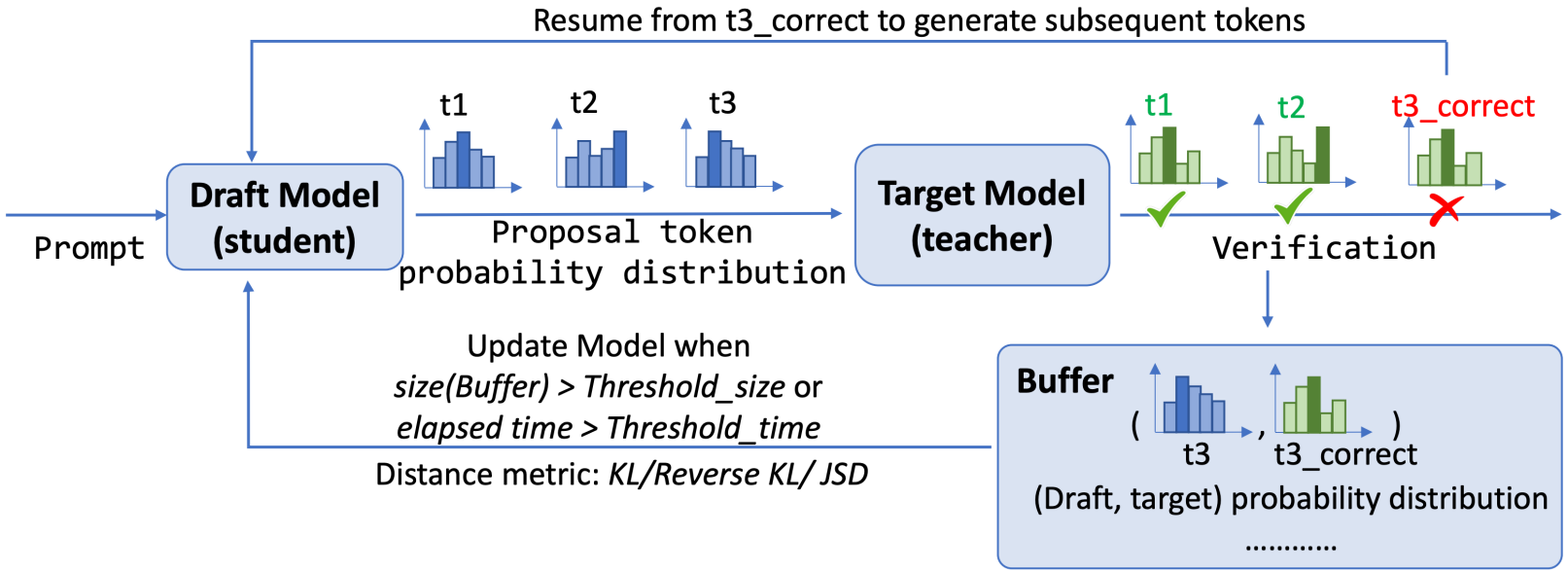
LLMs推理加速：投机解码将大模型的计算开销卸载到小模型



Fast Inference from Transformers via Speculative Decoding. Yaniv Leviathan, Matan Kalman, Yossi Matias

在线投机解码：在线蒸馏+投机解码

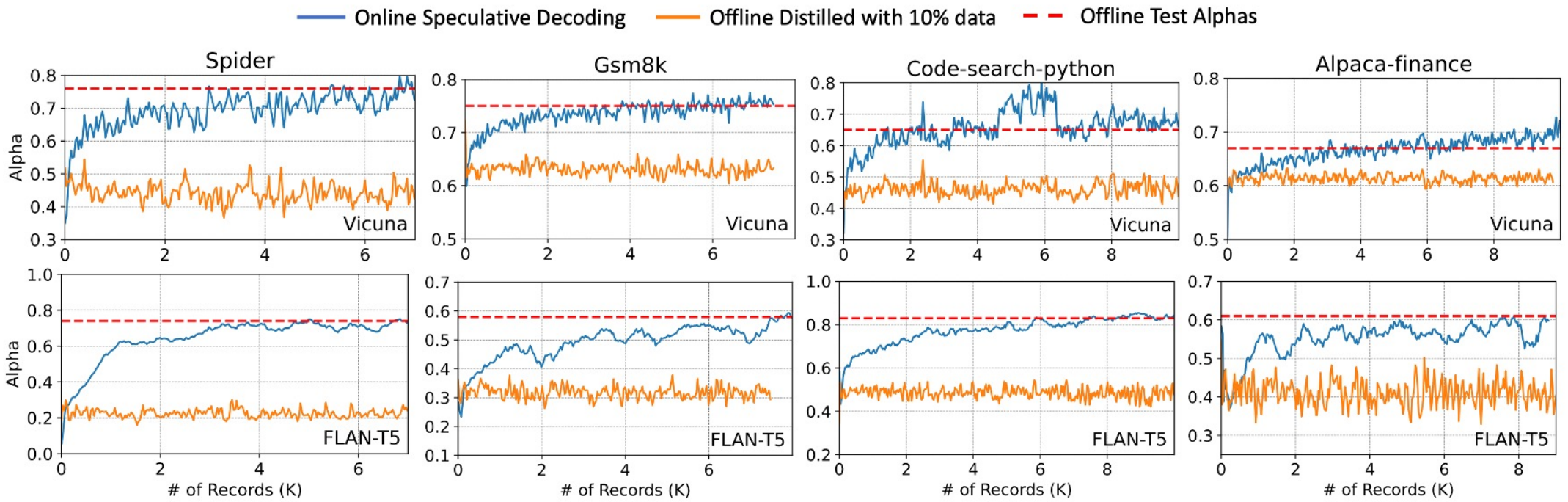
Open domain情况下，draft model快速适配query dist.



- 将草稿模型的错误预测和目标模型的校正结果存储在**buffer**中
- 当**buffer**打满，基于在线蒸馏损失函数更新草稿模型

Liu, Hu, Bailis, Cheung, Deng, Stoica & Zhang. Online Speculative Decoding. ICML 2024.

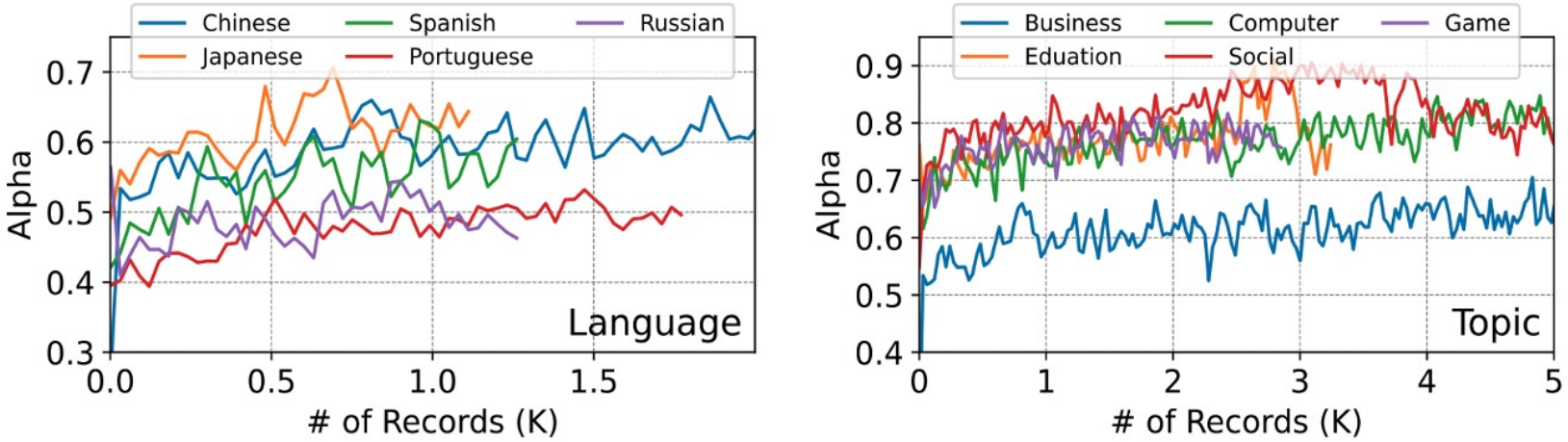
在线投机解码：结果（快速提高草稿模型的token acceptance rate）



- 在线部署场景中，草稿模型渐渐适配数据分布，不断变准，因此加速效率不断提升

Liu, Hu, Bailis, Cheung, Deng, Stoica & Zhang. Online Speculative Decoding. ICML 2024.

在线投机解码：结果（结合基于语言/主题的路由）



- 使用多个草稿模型独立处理不同语言/主题，相对于单个草稿模型，进一步提高tokens acceptance rate
- 可拓展为基于用户的路由，为每一个用户部署一个草稿模型
-

Liu, Hu, Bailis, Cheung, Deng, Stoica & Zhang. Online Speculative Decoding. ICML 2024.

在线投机解码：结果

Dataset	Spider	Chatbot Arena	Extra Parameters (B)
Medusa-7B	1.34×	2.03×	0.44
Medusa-7B + OSD	2.01×	2.38×	0.44
Draft model + OSD	2.17×	1.51×	0.16

超越/结合**Medusa**（一个代表性的多头LLM）

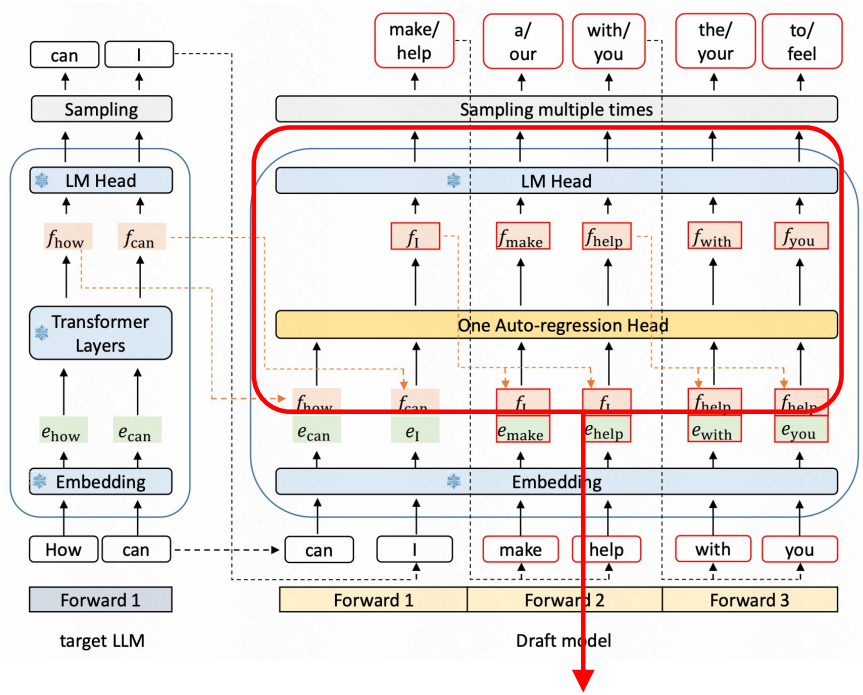
Dataset	**Spider**	**Gsm8k**	**Alpaca-Finance**	**Code-Python**
Tokens with the greatest precision increase	AV, SELECT, first, ⟨EOS⟩, template, SUM, G, COUNT, \n, city, WHERE, ';, (, IST, id	⟨EOS⟩, >>, +, To, <<, this, =, %, know, are, We, calculate, be, The, have	1, Here, (, :, provide, depends, However, goals, amount, 3, there, The, \n, personal, will	"', (, Here, python, ', how, doc, snippet, import, based, {, Python, This, :, you
Tokens with the greatest recall increase	SELECT, *, FROM, (, IST, *), \n, COUNT, G, first, WHERE, ⟨EOS⟩, IN, ;, MAX, ';	start, >>, <<, +, find, how, we, =, fore, To, so, \, ⟨EOS⟩, then, let	general, 1, several, This, depends, Here, provide, However, goals, over, (, If, amount, it, can	Here, This, snippet, "', ', how, python, (, takes, Python, you, doc, an, import, def

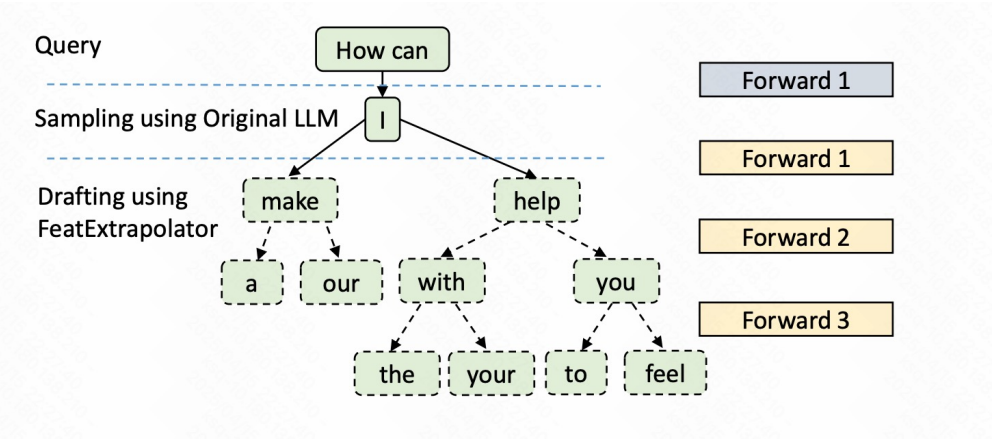
Token acceptance rate 提升最多的**tokens**

Liu, Hu, Bailis, Cheung, Deng, Stoica & Zhang. Online Speculative Decoding. ICML 2024.

轻量化draft model的投机解码：Eagle家族

- 核心思想：引入一个新的轻量化自回归头生成草稿序列





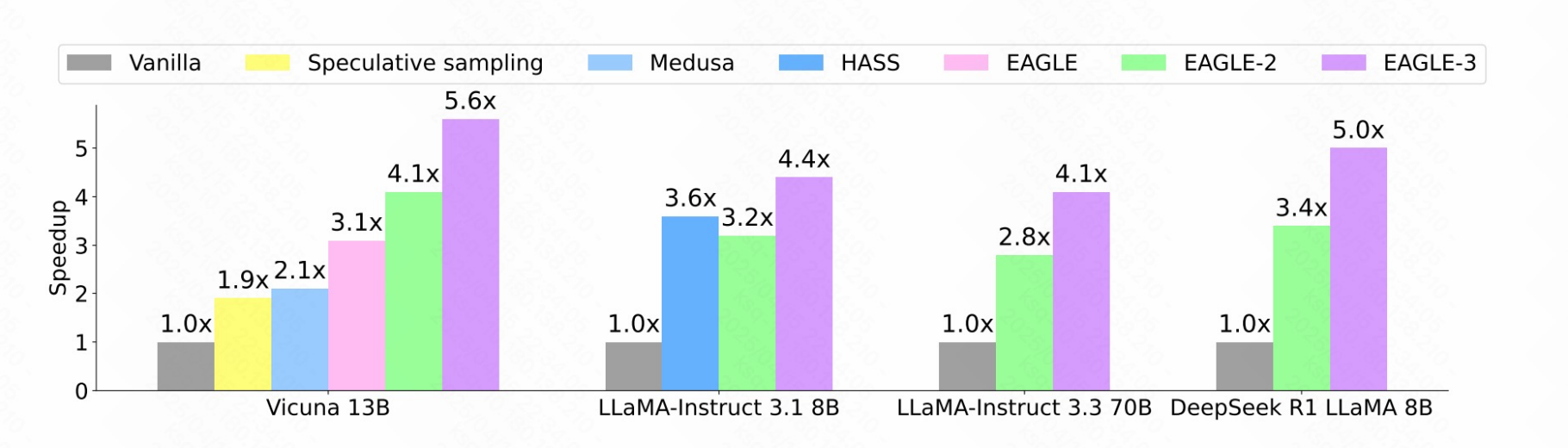
利用树形注意力并行验证草稿序列

与Medusa的主要区别：自回归地预测隐藏特征而非直接预测Token

Li, Yuhui, et al. "EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty." Forty-first International Conference on Machine Learning.

轻量化draft model的投机解码：Eagle家族

- 25年3月发布的Eagle3模型实现4-5倍加速



- 主要改进：
  - 除最后一层特征外，concat中间层特征作为自回归头预测特征的输入
  - Scale up草稿模型的训练数据，实验验证接收率与训练数据间的scaling law

Li, Yuhui, et al. "EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test." *arXiv:2503.01840* (2025).

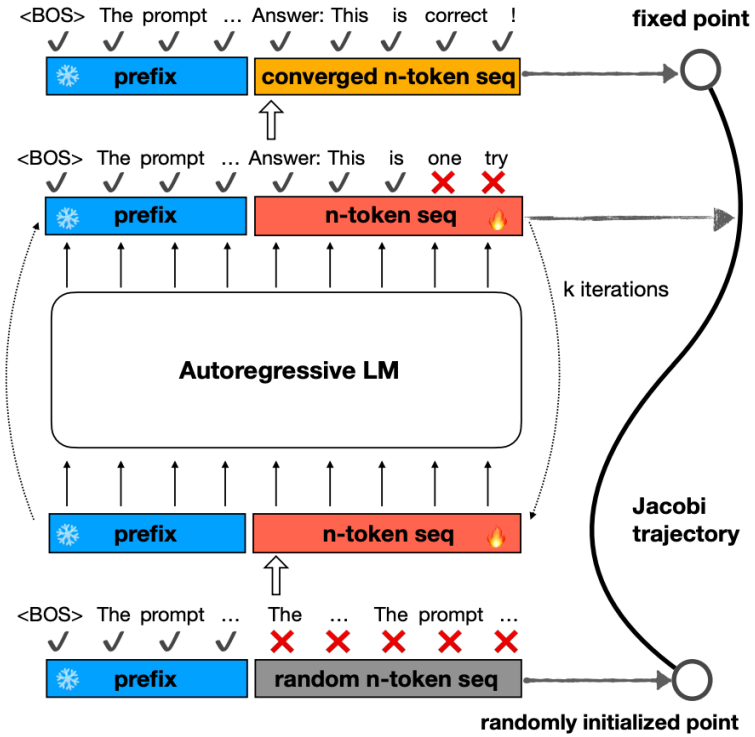
大语言模型并行解码的初试: Jacobi decoding

- 给定大语言模型，同时预测 **n**个**token**等价于求解：

  其中 $f(y_i, \boldsymbol{y}_{<i}, \boldsymbol{x}) = 0 \text{ for } i = 1, \ldots, n.$

- 这$f(y_i, \boldsymbol{y}_{<i}, \boldsymbol{x}) := y_i - \arg\max_y p(y|\boldsymbol{y}_{<i}, \boldsymbol{x})$并行求解，步数不超过**n**，生成质量可保证

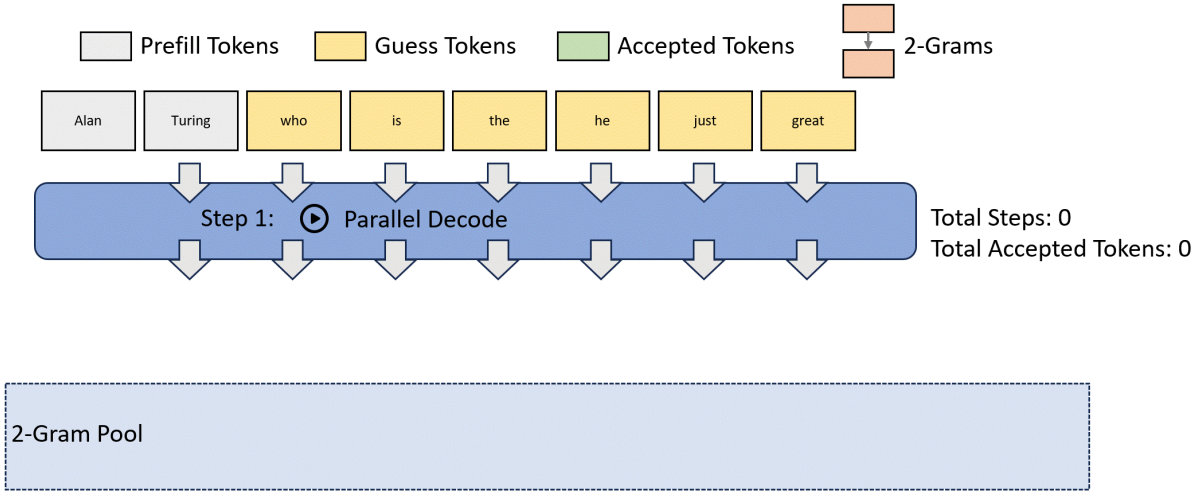
- 但实际效果差（如：仅**1.05**倍提升）
  - 原因：模型训练时未学过如何预测多个**tokens**



Song, Y., Meng, C., Liao, R., and Ermon, S. Accelerating feedforward computation via parallel nonlinear equation solving. ICML 2021.
Santilli, A., Severino, S., Postolache, E., Maiorca, V., Mancusi, M., Marin, R., and Rodola, E. Accelerating transformer inference for translation via parallel decoding. ACL 2023.

无需draft model的投机解码：Lookahead Decoding

- 通过收集和缓存**Jacobi**迭代过程中的**n-gram**进行**drafting**
  - 并行验证n-gram pool中的多组候选词元，若匹配则直接填充至生成序列中



Pros

- 简化部署、即插即用

- Cons
  - 只适用于贪婪采样策略

Fu, Yichao, et al. "Break the Sequential Dependency of LLM Inference Using Lookahead Decoding." *Forty-first International Conference on Machine Learning*.

Lookahead Decoding：结果

- 实现至多2.25倍的thoughput加速

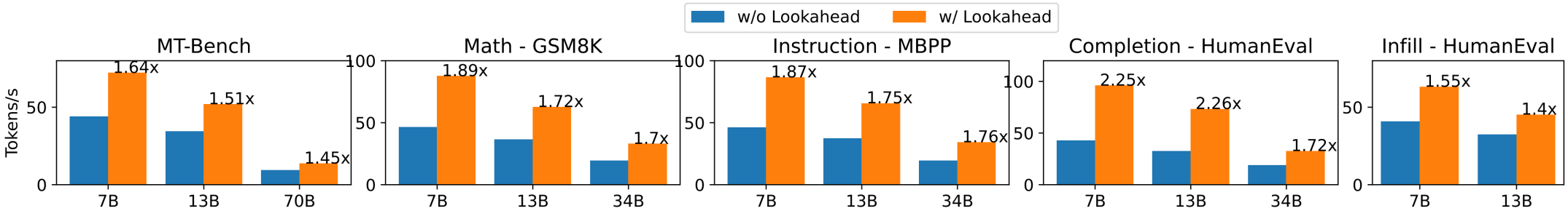
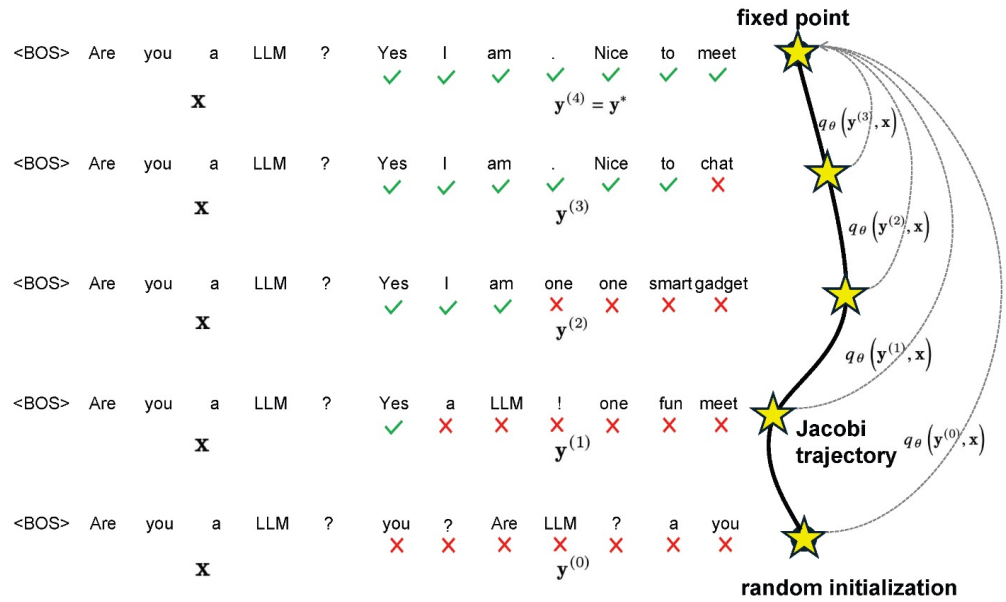


Figure 5: Throughput of LOOKAHEAD DECODING on various dataset without FlashAttention and distributed serving

Fu, Yichao, et al. "Break the Sequential Dependency of LLM Inference Using Lookahead Decoding." *Forty-first International Conference on Machine Learning*.

Jacobi decoding能否加速？一致性大语言模型（Consistency LLMs, CLLMs）

- 通过训练习得预测**n**个**tokens**的能力
  - 从随机初始化的起点预测fixed point？
    - 不行，问题太难，训练难收敛
  - 从Jacobi解码轨迹上的任意点预测fixed point？
    - 可以，形成一系列从简单到困难的学习问题，有助于模型收敛



Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性大语言模型

fixed point

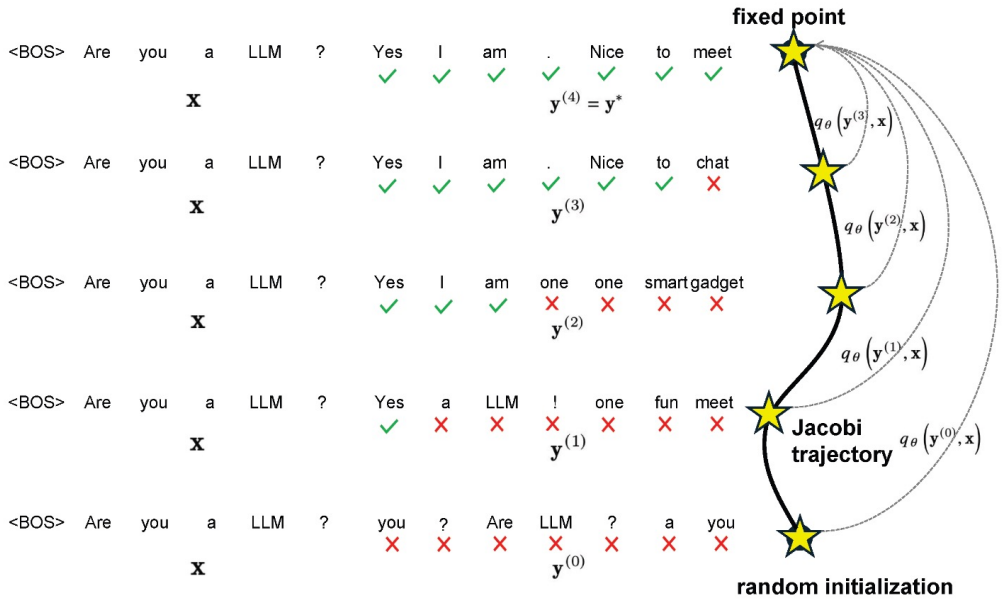
$$\mathcal{L}_{\mathrm{GC}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathcal{J})\sim\mathcal{D},\mathbf{y}\sim\mathcal{J}}\Big[\sum_{i=1}^{n} D(q_{\theta^-}(\cdot|\mathbf{y}^*_{:i},\mathbf{x}))||q_\theta(\cdot|\mathbf{y}_{:i},\mathbf{x})\Big]$$

$$\mathcal{L}_{\mathrm{LC}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathcal{J})\sim\mathcal{D},(\mathbf{y}^{(j)},\mathbf{y}^{(j+1)})\sim\mathcal{J}}\Big[\sum_{i=1}^{n} D(q_{\theta^-}(\cdot|\mathbf{y}^{(j+1)}_{:i},\mathbf{x}))||q_\theta(\cdot|\mathbf{y}^{(j)}_{:i},\mathbf{x})\Big]$$

$D(\cdot||\cdot)$ 是分布间距离度量

$$\mathcal{L}_{\mathrm{AR}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{l})\sim\mathcal{D}}\Big[-\sum_{i=1}^{N}\log q_\theta(l_i|\mathbf{l}_{:i},\mathbf{x})\Big]$$

自回归损失防止模型退化



Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性大语言模型：结果

聊天，**Vicuna-7B**，**2.4**倍加速

数学，**Abel-7B-001**，**3**倍加速

代码，**Deepseek-coder-7B**，**3.4**倍加速

Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性大语言模型：结果

- 至多3.6倍加速

相比于Medusa2、Eagle3，不需要模型架构上的改变

生成质量极少损失

Methods	Speed (tokens/s)	Speedup	Metric	Size
GSM8K				
Fine-tuned LLaMA2-7B (Chern et al.)				
+ AR	43.5	1.0×	59.1	
+ Jacobi	45.7	1.1×	59.1	6.7B
+ lookahead	74.8	1.7×	59.1	
CLLM-LLaMA2-7B				
+ AR	43.5	1.0×	56.4	
+ Jacobi	132.4	**3.0**×	56.4	6.7B
+ lookahead	125.2	2.9×	56.4	
Medusa-2 + LLaMA2-7B				
+ typical	70.2	1.6×	51.3	8.3B
Fine-tuned LLaMA2-7B + distilled LLaMA-160m				
+ speculative	73.8	1.7×	59.1	6.8B
ShareGPT (MT-Bench)				
Fine-tuned LLaMA2-7B				
+ AR	37.6	1.0×	6.5	
+ Jacobi	39.9	1.1×	6.5	6.7B
+ lookahead	60.8	1.6×	6.5	
CLLM-LLaMA2-7B				
+ AR	36.7	1.0×	6.4	
+ Jacobi	88.4	2.4×	6.4	6.7B
+ lookahead	95.0	**2.5**×	6.4	
Medusa-2 + LLaMA2-7B				
+ typical	102.5	2.7×	6.4	8.3B
Fine-tuned LLaMA2-7B + distilled LLaMA-160m				
+ speculative	51.3	1.4×	6.5	6.8B

Methods	Speed (tokens/s)	Speedup	Metric	Size
Spider				
Fine-tuned Deepseek-7B				
+ AR	38.0	1.0×	70.0	
+ Jacobi	39.5	1.0×	70.0	6.7B
+ lookahead	55.3	1.5×	70.0	
CLLM-Deepseek-7B				
+ AR	38.0	1.0×	69.3	
+ Jacobi	127.4	3.4×	69.3	6.7B
+ lookahead	135.2	**3.6**×	69.3	
Medusa-2 + Deepseek-7B				
+ typical	104.2	2.7×	66.4	8.3B
Fine-tuned Deepseek-7B + distilled LLaMA-160m				
+ speculative	66.8	1.8×	70.0	6.8B
Code-Search-Net Python				
Fine-tuned Deepseek-7B				
+ AR	40.1	1.0×	60.4	
+ Jacobi	43.2	1.1×	60.4	6.7B
+ lookahead	68.0	1.7×	60.0	
CLLM-Deepseek-7B				
+ AR	38.5	1.0×	59.2	
+ Jacobi	102.1	2.5×	59.2	6.7B
+ lookahead	115.7	**2.9**×	59.2	
Medusa-2 + Deepseek-7B				
+ typical	128.0	3.2×	48.3	8.3B
Fine-tuned Deepseek-7B + distilled LLaMA-160m				
+ speculative	59.3	1.5×	60.4	6.8B

Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性大语言模型：结果

- 训练的开销很低

Table 9. Computation required for consistency training.

Dataset	Training time	% of pre-training cost	Training resources
Spider	2 hours	$< 0.01\%$	8 A100 40GB GPUs
GSM8K	12 hours	$\sim 0.01\%$	8 A100 40GB GPUs
CodeSearchNet-Python	22 hours	$\sim 0.1\%$	8 A100 40GB GPUs
ShareGPT	30 hours	$\sim 0.2\ \%$	8 A100 40GB GPUs

Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性大语言模型：为什么加速？

- **Fast forwarding**：多个连续的tokens在单次网络传播中被正确预测

Stationary tokens：提前被正确预测，并在后续迭代中保持不变，即使之前的tokens存在错误

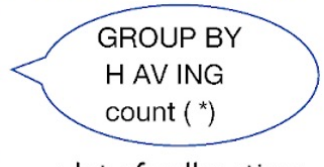
Target LLM

```
0: with ## __ ' , Manager table have number ID ' Could Player or ruction ['
1: SELECT T as ## as '' , as columns _ _ C I you _ ''
2: SELECT country 2 T FROM ## Country COUNT ## FROM name name ub _ t FROM
3: SELECT country FROM FROM 1 player T AS ( WHERE player AS GROUP WHERE id WHERE
4: SELECT country FROM player player WHERE GROUP 1 T SELECT T . T BY player AS
5: SELECT country FROM player GROUP _ country _ . 1 country GROUP country GROUP COUNT GROUP
6: SELECT country FROM player GROUP BY CON H count H H H BY H BY (
7: SELECT country FROM player GROUP BY country TRY I L H AV AV country H H
8: SELECT country FROM player GROUP BY country H A I A V AV ING G AV
9: SELECT country FROM player GROUP BY country H AV ING V ING COUNT COUNT COUNT >
10: SELECT country FROM player GROUP BY country H AV ING count E _ ( ( (
11: SELECT country FROM player GROUP BY country H AV ING count ( ( number *) *)
12: SELECT country FROM player GROUP BY country H AV ING count ( *) *) *) >
13: SELECT country FROM player GROUP BY country H AV ING count ( *) > *) >
14: SELECT country FROM player GROUP BY country H AV ING count ( *) > 1 '\n'
```

Consistency LLM

```
0: with ## __ ' , Manager table have number ID ' Could Player or ruction ['
1: SELECT country country FROM player GROUP BY country H H AV AV AV ING *) *)
2: SELECT country FROM player BY BY H AV AV AV ING count ( *) > ' '
3: SELECT country FROM player GROUP BY H AV ING count ( *) > 1 ' ' ' '
4: SELECT country FROM player GROUP BY country AV AV count ( *) > 1 ' ' '\n'
5: SELECT country FROM player GROUP BY country H AV ING count *) > 1 ' ' '\n'
6: SELECT country FROM player GROUP BY country H AV ING count ( *) > 1 '\n'
```

Target LLM



a lot of collocations

CLLM

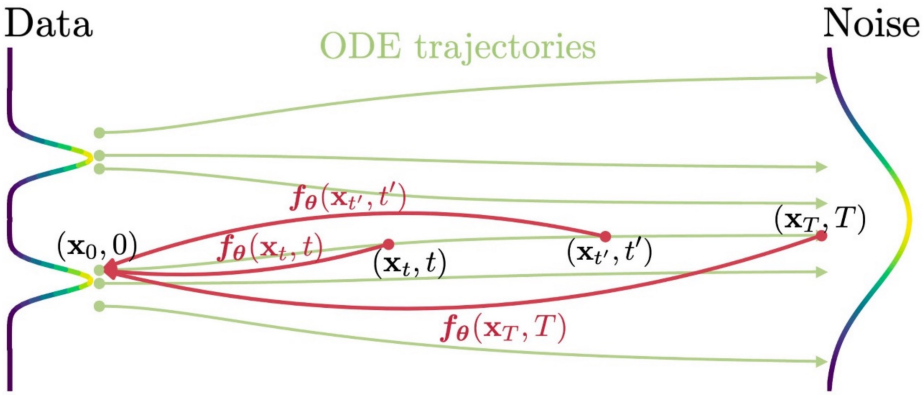
Kou, Hu, He, Deng & Zhang. CLLMs: Consistency Large Language Models. ICML 2024.

一致性蒸馏加速算法

一致性蒸馏（Consistency Distillation）

- 概率流ODE定义了从噪声到数据的一一映射

  =>直接建模此映射 $\forall t \in [0,T] : \boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \mathbf{x}_0$

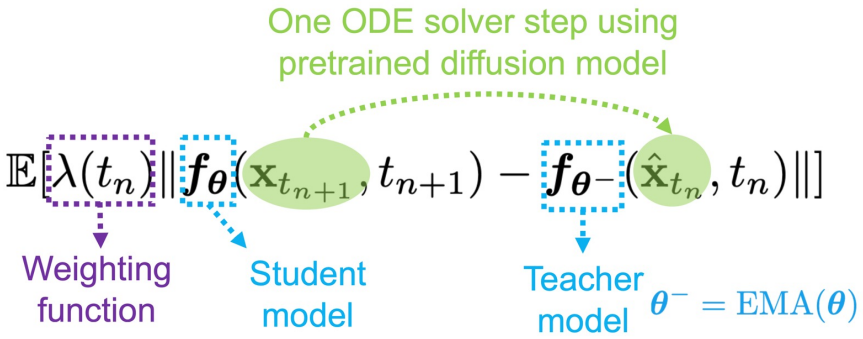


- 模型参数化：需保证边界条件（t=0）

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)F_{\boldsymbol{\theta}}(\mathbf{x}, t)$$

$$c_{\text{skip}}(0) = 1 \qquad c_{\text{out}}(0) = 0$$

- 训练：



Algorithm 1 Multistep Consistency Sampling

Input: Consistency model $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot,\cdot)$, sequence of time points $\tau_1 > \tau_2 > \cdots > \tau_{N-1}$, initial noise $\hat{\mathbf{x}}_T$

$\mathbf{x} \leftarrow \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_T, T)$

for $n = 1$ **to** $N-1$ **do**

  Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

  $\hat{\mathbf{x}}_{\tau_n} \leftarrow \mathbf{x} + \sqrt{\tau_n^2 - \epsilon^2}\mathbf{z}$

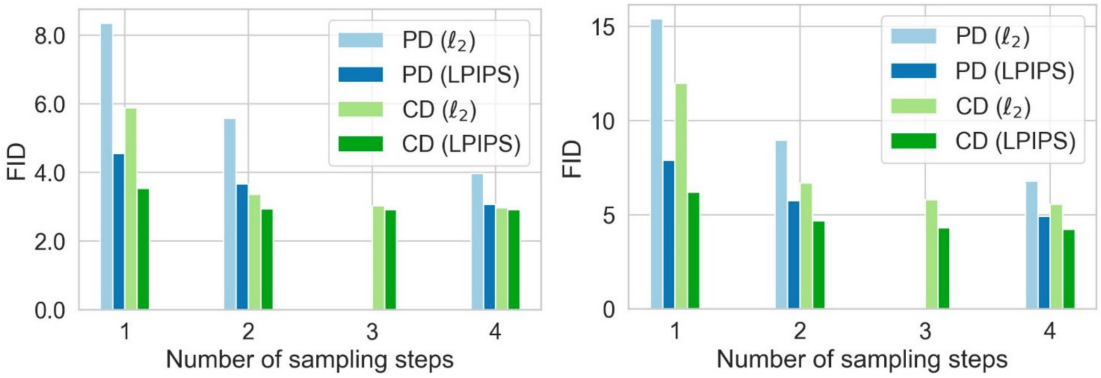
  $\mathbf{x} \leftarrow \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{\tau_n}, \tau_n)$

end for

Output: $\mathbf{x}$

Song et al. Consistency Models. ICML 2023

一致性蒸馏：结果



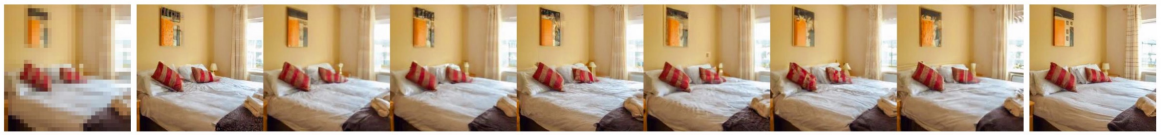
(a) CIFAR-10　　(b) ImageNet 64 × 64

一步生成的**SOTA FID**:

- 3.55 on CIFAR-10
- 6.20 on ImageNet 64

可用于零样本图像编辑等应用

(a) *Left*: The gray-scale image. *Middle*: Colorized images. *Right*: The ground-truth image.



(b) *Left*: The downsampled image (32 × 32). *Middle*: Full resolution images (256 × 256). *Right*: The ground-truth image (256 × 256).

(c) *Left*: A stroke input provided by users. *Right*: Stroke-guided image generation.

Song et al. Consistency Models. ICML 2023

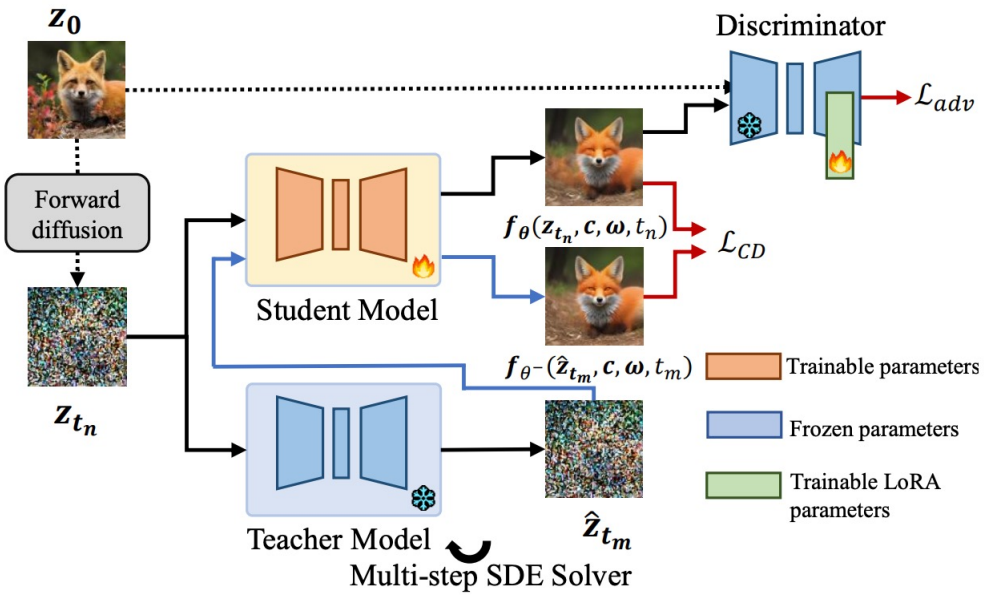
一致性蒸馏：用于DALL-E 3

In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.

A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. the grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone.

https://cdn.openai.com/papers/dall-e-3.pdf

随机一致性蒸馏



$$\min_{\theta} \mathcal{L}_{CD}(\theta) = \mathbb{E}_{n, z_{t_n}} \left[\lambda(t_n) \left\| f_\theta(z_{t_n}, t_n) - f_{\theta^-}(\hat{z}_{t_m}, t_m) \right\|_2^2 \right]$$

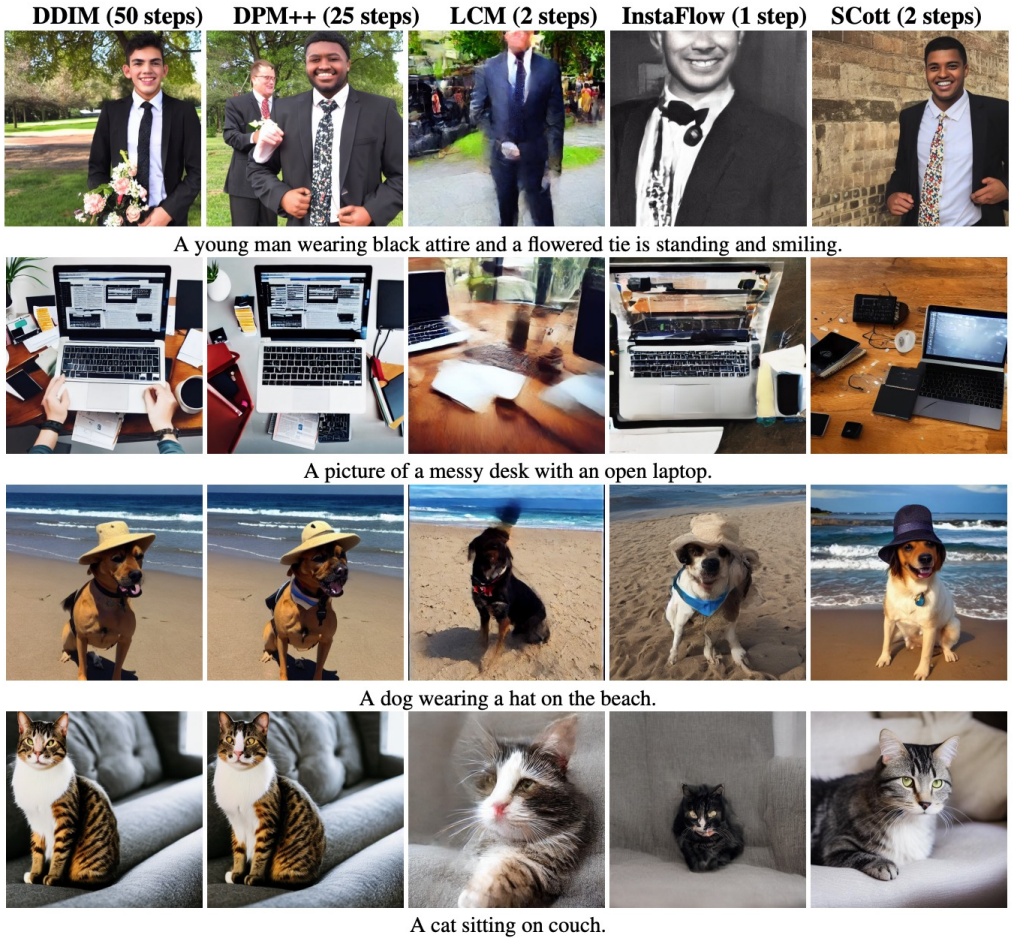
- 基于随机微分方程（SDE）采样器构建蒸馏教师模型
- 多步SDE采样矫正累计误差
- 引入的噪声可以视作数据增广

Liu, Xie, Ye, Deng, etc. SCott: Accelerating Diffusion Models with Stochastic Consistency Distillation. AAAI

随机一致性蒸馏：结果

- 两步采样达到21.9的FID
  - 显著超越InstaFlow 和UFOGen

Figure 1: 512 × 512 resolution images generated by SCott using 2 sampling steps. SCott is trained based on Realistic-Vision-v51.



Liu, Xie, Ye, Deng, etc. SCott: Accelerating Diffusion Models with Stochastic Consistency Distillation. AAAI

分段一致性蒸馏

- 直接学习从噪声到图像的一致性映射太难，可将**ODE**轨迹切段分别处理
  - 课程学习思想：渐进减少切段数目
  - 基于教师模型的采样进行学习，避免对高质量训练集的依赖
  - 引入偏好学习损失，兼顾加速和对齐

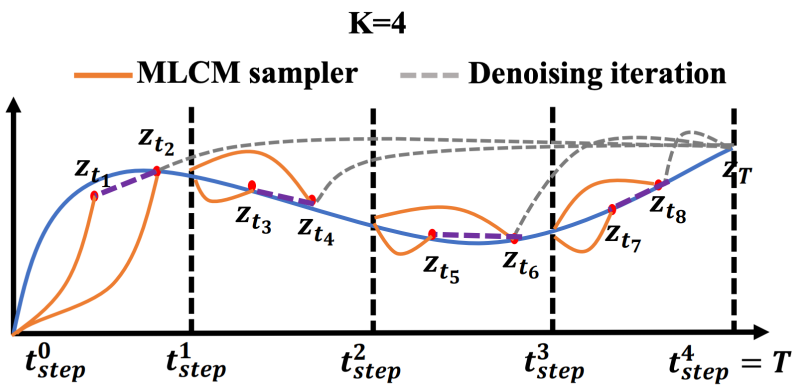


Image-free MLCD in the first stage

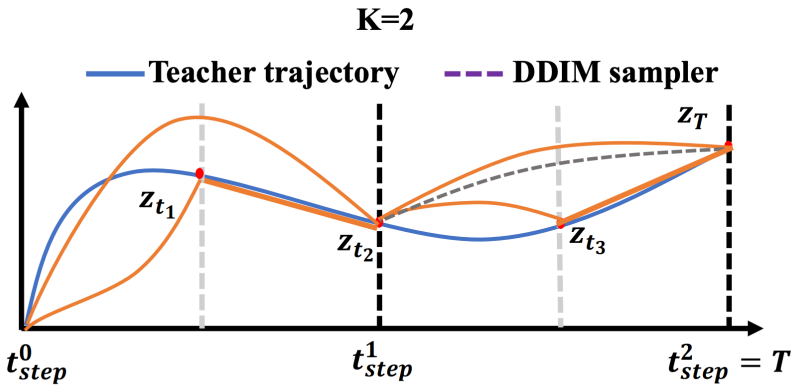


Image-free MLCD in the progressive training stage

Xie, Liao, Deng, Chen, Tang & Lu. MLCM: Multistep Consistency Distillation of Latent Diffusion Model. Arxiv 2406.05768.

分段一致性蒸馏：结果

Figure 1: 1024 × 1024 image samples from MLCM, distilled from SDXL-base-1.0 [32] based on LoRA [8]. From top to bottom, **2**, **3**, and **4** sampling steps are adopted, respectively. Apart from such good visual quality, MLCM can also yield improved metrics compared to strong baselines.

Table 1: Quantitative comparisons on MSCOCO-2017 5K validation datasets. All models adopts SDXL architecture.

Method	Step	CS	AS	IR	PS
DDIM [42]	25	33.36	5.54	0.87	0.229
LCM [23]	4	32.53	5.42	0.48	0.224
SDXL-Turbo [38]	4	33.30	5.64	0.83	0.226
SDXL-Lightning [17]	4	32.40	5.63	0.72	0.229
SDXL-Lightning [17]	8	32.73	5.95	0.71	0.227
HyperSD [33]	4	32.64	5.52	1.15	**0.234**
HyperSD [33]	8	32.41	5.83	1.14	0.233
MLCM	2	33.02	6.10	1.10	0.227
MLCM	3	33.24	6.17	1.18	0.232
MLCM	4	33.30	6.19	1.20	0.233
MLCM	8	**33.48**	**6.20**	**1.22**	0.233

- 无需训练图片，统一模型，兼容不同采样步数
- **2**步采样即可实现高质量的**1024²**的图像生成
  - 远超**SDXL, HyperSD**等开源模型

Xie, Liao, Deng, Chen, Tang & Lu. MLCM: Multistep Consistency Distillation of Latent Diffusion Model. Arxiv 2406.05768.

分段一致性蒸馏：结果

Figure 4: MLCM with ControlNet. Our MLCM can be incorporated into ControlNet pipeline and produce satisfactory results with 2 steps sampling.

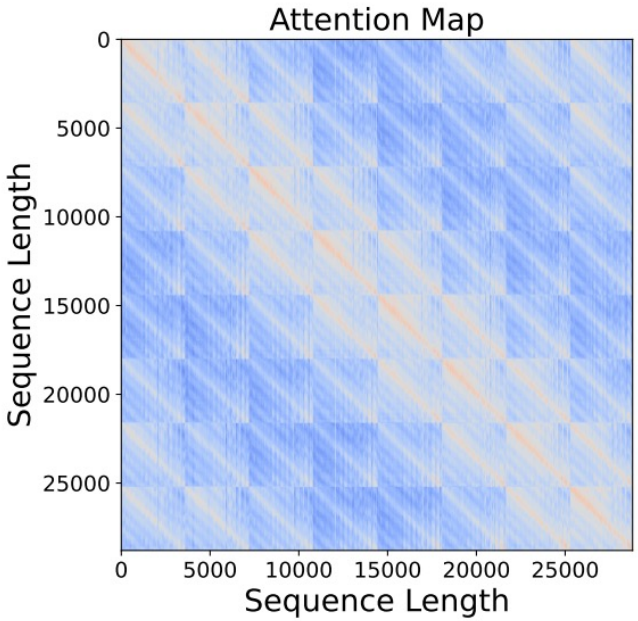
与ControlNet无缝结合：2步采样结果

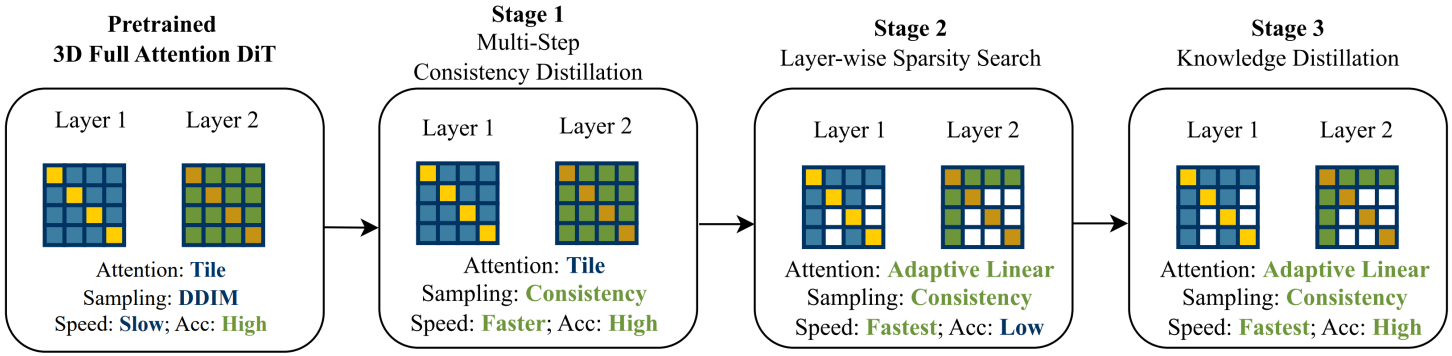
Figure 6: MLCM for Chinese-to-image generation. With 3 steps sampling, our MLCM model can produce images that align with Chinese semantic meaning. The first line presents images in general Chinese contexts, while the second line showcases images in specific Chinese cultural settings.

赋能中文-图像生成：3步采样结果

Xie, Liao, Deng, Chen, Tang & Lu. MLCM: Multistep Consistency Distillation of Latent Diffusion Model. Arxiv 2406.05768.

分段一致性蒸馏+注意力稀疏：5-10倍视频生成加速





Ding et al. EFFICIENT-VDIT: EFFICIENT VIDEO DIFFUSION TRANSFORMERS WITH Attention Tile.

3、如何 高效地构建/构建高效的 深度推理模型？

深度推理: the new frontier of AIGC?

deepseek-r1开源复现方法整理

yeyan，中国地质大学 工程硕士

deepseek-r1持续火热，估计会掀起一波复现其训练过程的热潮，先简单整理下目前看到的。

目录 **方法**

*** 1.1 open-r1**

由huggingface组建，目前刚上线2周，发布了最新进展open-r1/update-1，在MATH-500任务上接近deepseek的指标，可以在open-r1/open-r1-eval-leaderboard查看指标的排行榜。

Model	MATH-500 (HF lighteval)	MATH-500 (DeepSeek Reported)
DeepSeek-R1-Distill-Qwen-1.5B	81.6	83.9
DeepSeek-R1-Distill-Qwen-7B	91.8	92.8
DeepSeek-R1-Distill-Qwen-14B	94.2	93.9
DeepSeek-R1-Distill-Qwen-32B	95.0	94.3
DeepSeek-R1-Distill-Llama-8B	85.8	89.1
DeepSeek-R1-Distill-Llama-70B	93.4	94.5

知乎 @yeyan

1.2 mini-deepseek-r1

用 GRPO 和倒计时游戏复制出一个简单版本的 R1。

在大约 50 步时，模型学会了正确的格式，即<think>...</think>\n<answer>...</answer>;在 100 步时，解方程的成功率约为 25%，并且模型开始用文字进行 “推理”;在 200 步时，收敛变慢，成功率约为 40%。模型开始学习一种新的“格式”，它通过尝试不同的组合并检查结果来解方程，这种方式类似于编程解决问题的方式：在 450 步时，解方程的成功率为 50%，性能仍然在缓慢提升，并且模型保持了从 200 步开始的新格式。

*** 1.3 open-thoughts**

目标是整理一个推理数据集，用于训练最先进的小型推理模型，使其在数学和代码推理基准测试中超越 DeepSeek-R1-Distill-Qwen-32B 和 DeepSeek-R1-Distill-Qwen-7B。

目前已在以下领域生成数据：代码（Code）领域；数学（Math）领域；科学（Science）领域；谜题（Puzzle）领域。

目前已经发布了openthinker开源模型。

*** 1.4 TinyZero**

目录 ，3B 基础语言模型（LM）自行发展出自我验证和搜索能力。可以低于 30 美元的价格体验到"Ahah moment"。

User: Using the numbers [19, 36, 55, 7], create an equation that equals 65.
Assistant: Let me solve this step by step.
<think> We need to find an equation using the numbers 19, 36, 55, and 7 exactly once, with basic arithmetic operations, that equals 65. One possible combination is 55 + 36 - 19 + 7. Let's check: 55 + 36 = 91, 91 - 19 = 72, and 72 + 7 = 79. However, this doesn't equal 65. Another combination is 55 + 36 + 7 - 19. Let's check: 55 + 36 = 91, 91 + 7 = 98, and 98 - 19 = 79. This also doesn't equal 65. After trying different combinations, I found that 55 + 36 + 7 - 19 = 79, which is close but not equal to 65. Finally, I found that 55 + 36 - 19 + 7 = 79, which is still not equal to 65. However, if we try 55 + 36 - 7 + 19, we get 55 + 36 = 91, 91 - 7 = 84, and 84 + 19 = 103, which is not equal to 65. After further trial and error, I found that 55 + 36 + 7 - 19 = 79, which is still not equal to 65. However, if we try 55 + 36 - 7 - 19, we get 55 + 36 = 91, 91 - 7 = 84, and 84 - 19 = 65. This equals 65. </think>
<answer> 55 + 36 - 7 - 19 </answer>

知乎 @yeyan

*** 1.5 simpleRL-reason**

DeepSeek-R1 和 Kimi-k1.5 使用简单的强化学习算法来学习新兴的长思维链（CoT）和自我反思模式，并取得了良好的结果，其中没有使用 MCTS 和奖励模型。然而，他们的实验是基于大规模强化学习环境中的大型模型。目前尚不清楚小型模型是否能表现出类似的行为，需要多少数据，以及定量结果与其他方法相比如何。simpleRL-reason重现了 DeepSeek-R1-Zero* 和 DeepSeek-R1 用于复杂数学推理的训练，从 Qwen-2.5-Math-7B*（基础模型）开始，并且仅使用来自原始数学数据集的 8K（查询、最终答案）示例，平均获得了近 20 个绝对百分点的提升。

All results are in pass@1 accuracy

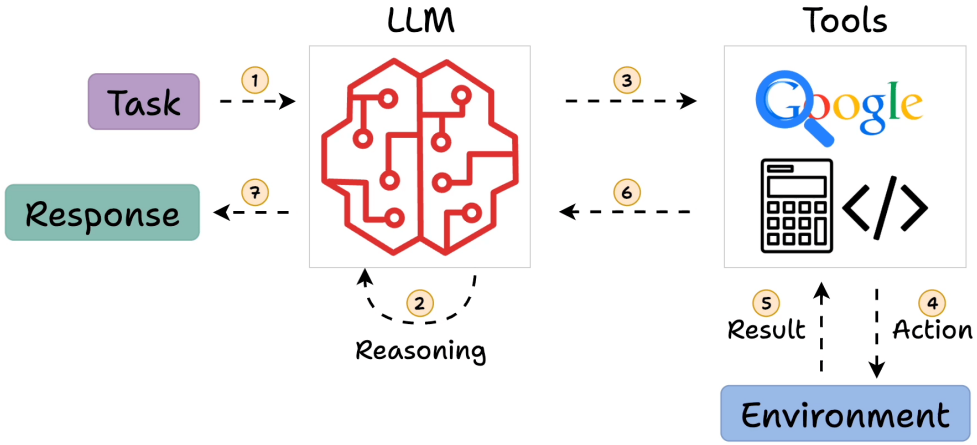
	AIME 2024	MATH 500	AMC	Minerva Math	OlympiadBench	Avg.
Qwen2.5-Math-7B-Base	16.7	52.4	52.5	12.9	16.4	30.2
Qwen2.5-Math-7B-Base + 8K MATH SFT	3.3	54.6	22.5	32.7	19.6	26.5
Qwen-2.5-Math-7B-Instruct	13.3	79.8	50.6	34.6	40.7	43.8
Llama-3.1-70B-Instruct	16.73	64.6	30.1	35.3	31.9	35.7
rStar-Math-7B	26.7	78.4	47.5	-	47.1	-
Eurus-2-7B-PRIME	26.7	79.2	57.8	38.6	42.1	48.9
Qwen2.5-7B-SimpleRL-Zero	33.3	77.2	62.5	33.5	37.6	48.8
Qwen2.5-7B-SimpleRL	26.7	82.4	62.5	39.7	43.3	50.9

知乎 @yeyan

1.6 RAGEN

RAGEN 是用于训练智能体模型的 DeepSeek-R1 (-Zero) 方法的首次复现，主要在gym-sokoban

开源社区对**DeepSeek R1**的复现如火如荼



推理能力是**Agent**够不够“聪明”的关键

提高推理模型的token利用率：Dr. GRPO

- **GRPO**为响应长度和题目难度引入了双重偏差：
  - 响应长度偏差：对正确回答偏向短句，对错误回答偏向长句
  - 难度归一化偏差：标准差小的问题影响力放大，标准差大的问题被压缩权重

GRPO

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right]\right\},$$

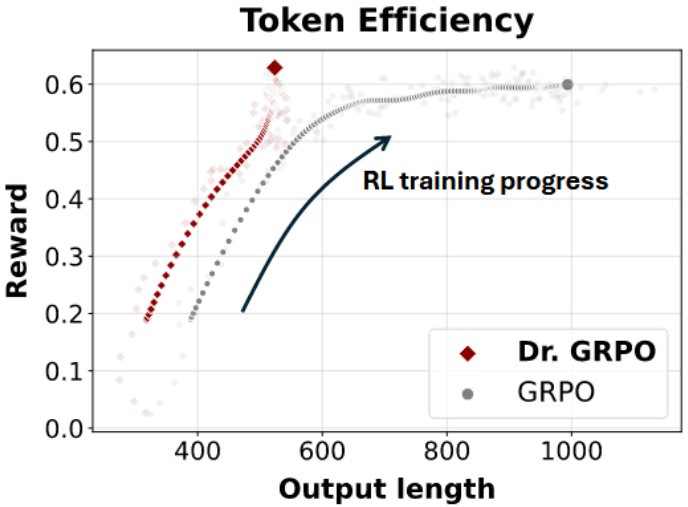
$$\text{where } \hat{A}_{i,t}=\frac{R(\mathbf{q},\mathbf{o}_i)-\text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\})}.$$

Dr. GRPO

GRPO Done Right (without bias)

$$\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t},\text{clip}\left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right]\right\},$$

$$\text{where } \hat{A}_{i,t}=R(\mathbf{q},\mathbf{o}_i)-\text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\}).$$



- 删除长度和奖励归一化项，消除**GRPO**中的偏差，减少错误输出的长度，提升**token**使用效率

Liu et al. Understanding R1-Zero-Like Training: A Critical Perspective. https://arxiv.org/pdf/2503.20783

提高推理模型的训练效率：DAPO

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(o|q)} \left\{ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{ \min\left[\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}A_i, \text{clip}\left(\frac{\pi_\theta\left(o_{i,t}\mid q,o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t}\mid q,o_{i,<t}\right)}, 1-\epsilon, 1+\epsilon \right) A_i \right] - \beta\mathbb{D}_{KL}^{(i,t)}\left[\pi_\theta\|\pi_{\text{ref}}\right] \right\} \right\}$$

1. 去掉 KL 散度
2. 分开设定 clip 上下阈值（ε_{low}, $\varepsilon_{\text{high}}$）
3. 去掉全对 or 全错的 sample 参与梯度更新
4. 优化 response-level 的长度偏差
5. 软的长度惩罚

GRPO 是先每条样例求 token-level 的平均，
再对所有样例求平均
DAPO 是对所有样例的所有 token 求平均
好处是优化长度偏差

clip-higher 就是分开设置上下阈值，阈值设宽点
让模型更能探索

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(o|q)} \left\{ \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left\{ \min\left[\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}A_i, \text{clip}\left(\frac{\pi_\theta\left(o_{i,t}\mid q,o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t}\mid q,o_{i,<t}\right)}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}} \right) A_i \right] \right\} \right\}$$

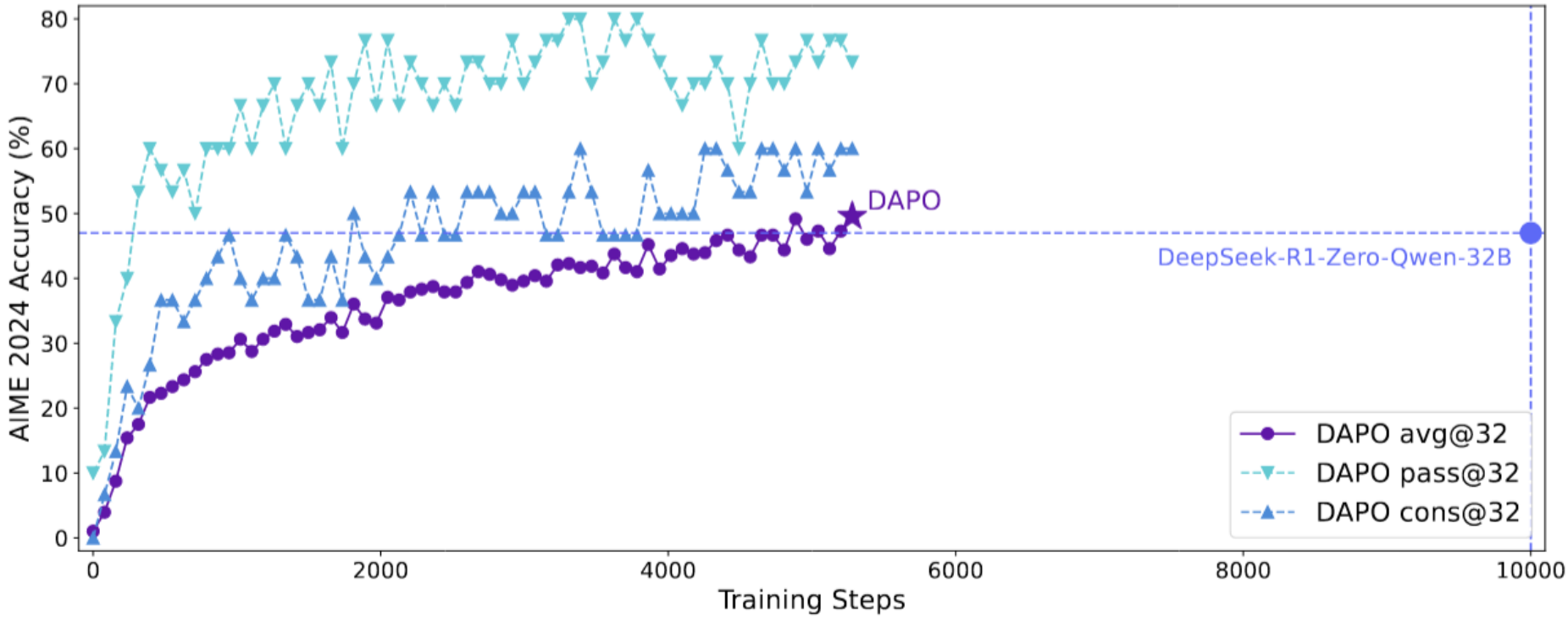
s.t. $0 < \left|\{o_i \mid o_i \text{ is correct}\}\right| < G$

去掉全对 or 全错的 sample，提高训练效率

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \le L_{\max} - L_{\text{cache}} \\ \frac{(L_{\max}-L_{\text{cache}})-|y|}{L_{\text{cache}}}, & L_{\max}-L_{\text{cache}} < |y| \le L_{\max} \\ -1, & L_{\max} < |y| \end{cases}$$

ByteDance Seed et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. https://arxiv.org/pdf/2503.14476

DAPO：结果



ByteDance Seed et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. https://arxiv.org/pdf/2503.14476

高效激发视觉空间推理能力：vsGRPO

- 视觉空间推理（**visual-spatial reasoning**）是最重要的多模态能力之一，对**VLA**等至关重要
- 初步发现：
  - 常用CoT策略对开源多模态模型在空间智能上性能无益

- Think-mode: `Let's think step by step and then answer the question using a single word or phrase.`
- Observe-mode: `Please observe the video first and then answer the question using a single word or phrase.`
- Vanilla-mode: `Please answer the question using a single word or phrase.`

Backbone	Methods	Avg	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Qwen2-VL-2B	Think-mode	22.9	18.4	4.3	31.5	17.3	28.3	22.9	26.2	16.8
	Observe-mode	21.8	16.8	1.7	32.7	22.7	28.8	27.6	26.2	18.1
	Vanilla-mode	23.3	21.4	3.4	32.3	31.1	26.7	27.7	24.7	18.9
Qwen2-VL-7B	Think-mode	31.3	44.8	26.1	25.3	23.4	34.7	30.9	32.9	31.5
	Observe-mode	32.0	29.9	19.0	39.6	32.0	34.6	40.0	36.0	24.4
	Vanilla-mode	32.2	39.4	25.0	25.8	43.2	32.6	30.9	27.8	32.6

Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., & Deng, Z. (2025). Improved Visual-Spatial Reasoning via R1-Zero-Like Training

高效激发视觉空间推理能力：vsGRPO

- 视觉空间推理（**visual-spatial reasoning**）是最重要的多模态能力之一，对**VLA**等至关重要
- 挑战：
  - 空间智能相关的视频-问题对缺失
  - 解决：基于**ScanNet**[Dai et al., 2017]构造**VSI-100k**，包括100k视频-问答对

Object Count
Question: How many trash can(s)in the room?
Answer: 3

Relative Direction
Question: If I am standing by the shelf and facing the shower is the bicycle to the left or the right of the shower?
Answer: left

Object Size
Question: What is the length of the longest dimension (length,width, or height) of the coffee table, measured in centimeters?
Answer: 113

Absolute Distance
Question: What is the distance between the shower and the kitchen counter (in meters)?
Answer: 6.1

Room Size
Question: What is the size of this room (in square meters)?
Answer: 47.9

Relative Distance
Question: Which of these objects (sink, pillow, bed, guitar)is the closest to the bicycle?
Answer: sink

- 不包括route planning和appearance order
  - 二者的构造需超越静态3D信息
  - 可测试任务泛化能力

Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., & Deng, Z. (2025). Improved Visual-Spatial Reasoning via R1-Zero-Like Training

vsGRPO：结果

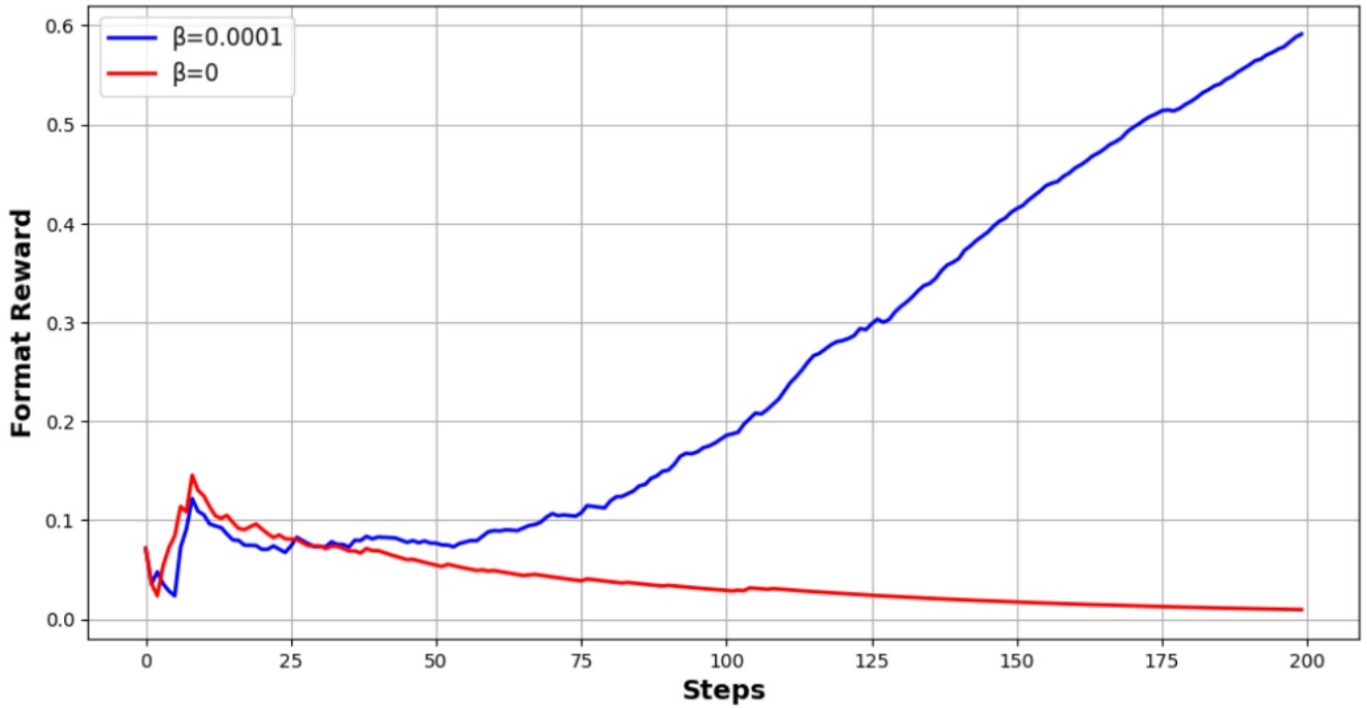
- 相比传统**SFT**和**DPO**有明显提升
- 微调2B模型表现超过了闭源的 **GPT-4o**
- 微调7B模型表现接近开源最佳的72B模型

Methods	Eval. Mode	Avg	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Open-source										
Qwen2-VL-2B	V	23.3	21.4	3.4	32.3	31.1	26.7	27.7	24.7	18.9
+ SFT	V	29.6	29.6	23.5	47.4	33.5	26.9	28.3	28.8	18.6
+ DPO	V	23.9	21.7	3.7	34.8	32.4	27.1	28.5	24.2	18.6
+ vsGRPO-T	V	26.1	24.7	10.7	37.4	36.2	27.3	29.5	25.7	17.9
+ vsGRPO-O	V	28.0	26.2	16.4	44.8	38.2	27.0	29.3	24.2	18.2
+ vsGRPO-T	T	29.6	35.0	28.2	34.7	25.2	28.0	38.5	28.5	18.7
+ vsGRPO-O	O	31.2	34.6	22.5	44.8	33.7	29.4	41.8	26.8	15.8
+ vsGRPO-V	V	35.4	53.6	29.0	52.7	43.4	28.1	30.9	26.8	18.9
Qwen2-VL-7B	V	32.2	39.4	25.0	25.8	43.2	32.6	30.9	27.8	32.6
+ SFT	V	38.1	44.7	27.6	46.1	50.4	34.0	35.7	33.0	33.4
+ DPO	V	32.6	39.1	25.2	26.5	44.2	32.6	30.9	29.3	33.3
+ vsGRPO-V	V	40.7	**59.9**	**29.6**	50.8	48.3	35.4	35.6	34.0	31.5
IVL2-2B	V	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
LNV-7B	V	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
IVL2-40B	V	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LNV-72B	V	**40.9**	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Close-source										
GPT-4o	V	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Pro	V	**48.8**	49.6	28.8	**58.6**	**49.4**	**46.0**	**48.1**	**42.0**	**68.0**

Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., & Deng, Z. (2025). Improved Visual-Spatial Reasoning via R1-Zero-Like Training

vsGRPO：结果

- **KL penalty**很重要



Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., & Deng, Z. (2025). Improved Visual-Spatial Reasoning via R1-Zero-Like Training

推理模型加速?

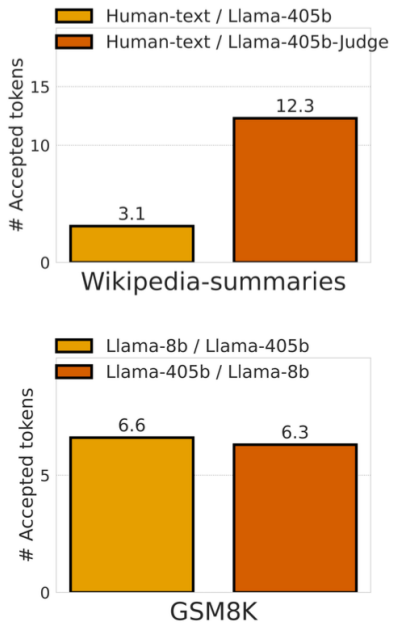
- 深度推理模型（LRM）：长链思维链（CoT）
  - **Pros**: 显著提升复杂任务性能
  - **Cons**: 导致推理延迟大幅增加，Reasoning阶段通常占用60-80%的总时间，相当于传统任务3至5倍的token量

可行方向：

1. 新型投机解码机制
2. CoT压缩：减少**Reasoning token**数量

投机解码：问题

- 标准投机解码要求输出分布对齐才能接受**proposal**
  - 导致低的token接受率、平均单次接受token数



将Llama-8b/Llama-405b分别交替作为draft/target model，接受token数量基本不变（通常认为大模型输出文本质量高于小模型）

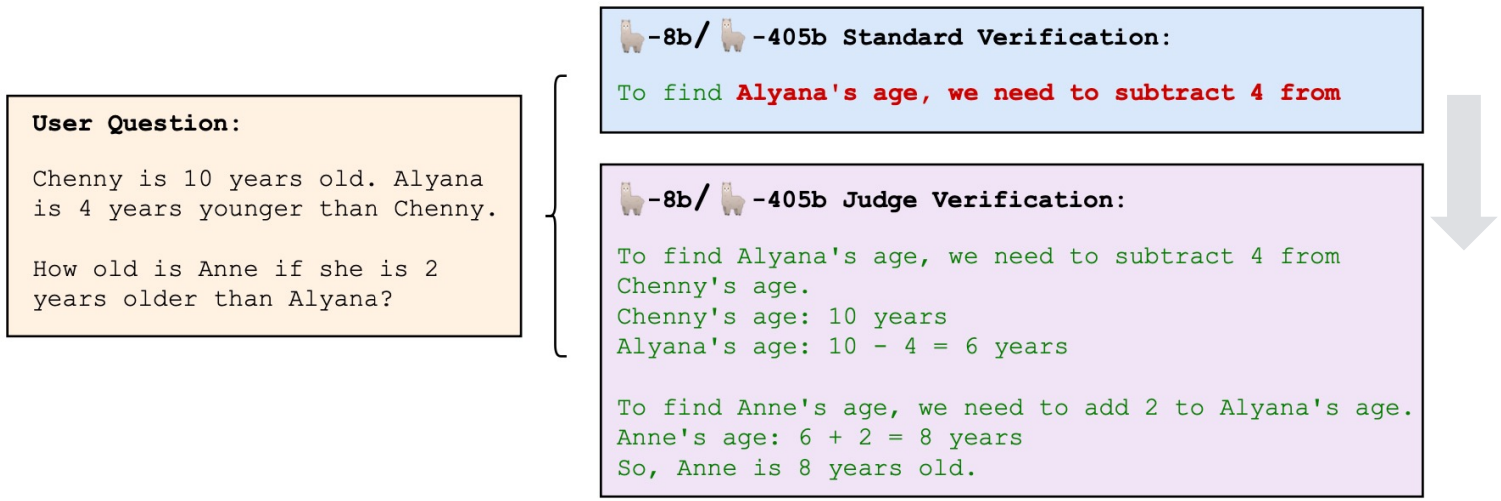
- 证明：文本质量和模型的接受率/接受token数量无关
- 即便高质量文本其在语义层面上可接受，对于上下文来说是正确的，仍会因为不符合target model分布而被拒绝，徒增迭代轮次

- 选取高质量人类文本作为draft，接受率也不高

Bachmann et al. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment. https://arxiv.org/abs/2501.19309

新型投机解码：**语义正确性**与**分布对齐性**同等重要

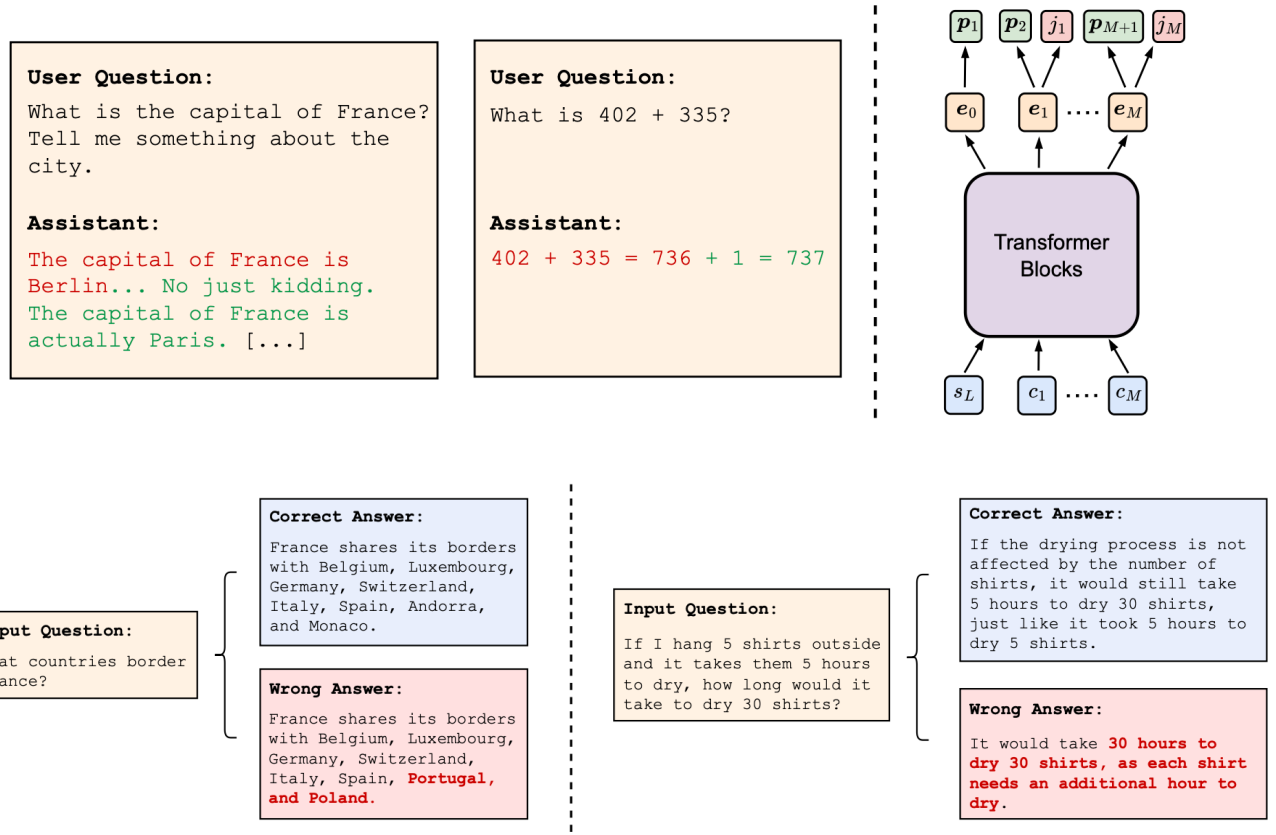
- 标准投机解码要求输出分布对齐才能接受**proposal**
  - 导致低的token接受率、平均单次接受token数
- 解决方法：**Judge Decoding**
  - 接受语义正确的proposal



Bachmann et al. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment. https://arxiv.org/abs/2501.19309

Judge Decoding：训练**简单线性层**预测proposal的可接受性

- LLMs有自动纠错能力
  - last hidden states能够有效地“标记”错误，促使模型生成后续 token 尝试纠正错误
- 构造正负样例数据集，用于训练新加线性层（回归头）
  - 16.4k参数的回归头在30k tokens上训练1.5h



Bachmann et al. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment. https://arxiv.org/abs/2501.19309

Judge Decoding：结果---每次接受更多tokens

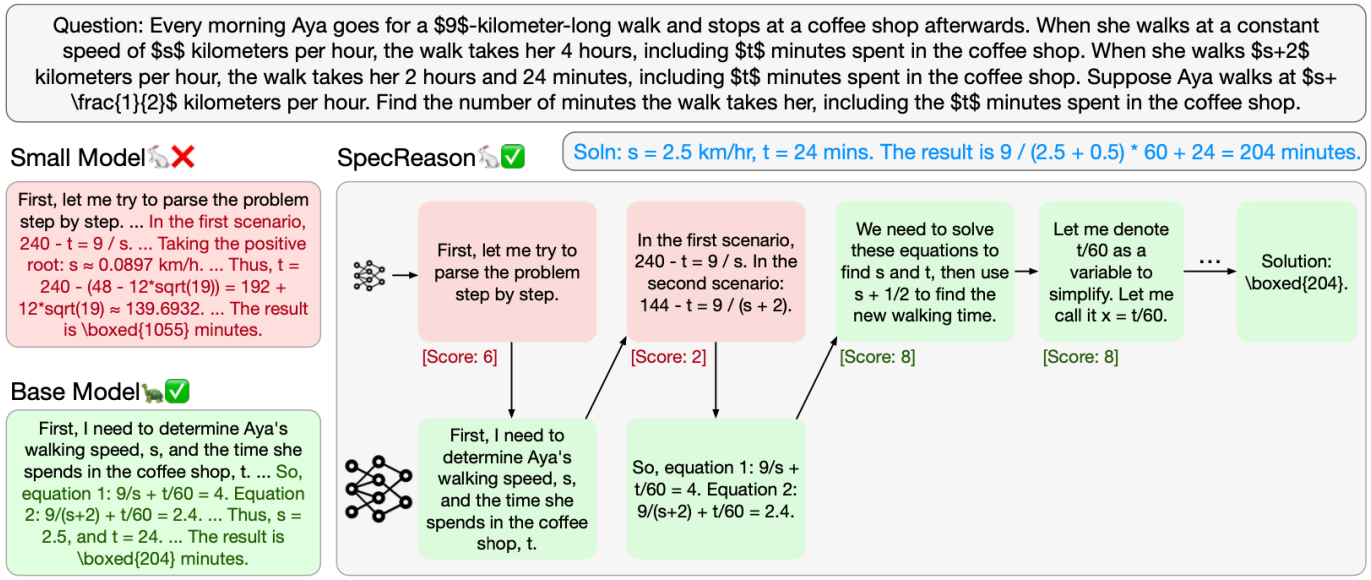
- 更多的单次接受tokens数和更高的加速比，尤其适合long-cot推理

	m_*	HUGGINGFACE	GPT-FAST	TOKENS/S $(512+512)$
8B/70B-STANDARD	6.4	1.5×	1.7×	76.7
8B/70B-JUDGE (OURS)	18.8	2×	**3×**	**141.8**
70B-EAGLE-2	4.5	**3.3×**	1.9×	88.1
8B/405B-STANDARD	6.3	5.3×	1.78×	58.7
8B/405B-JUDGE (OURS)	19.7	**9.7×**	**3.9×**	**129.3**
405B-MEDUSA	< 6	< 6×	1.9×	108*

Bachmann et al. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment. https://arxiv.org/abs/2501.19309

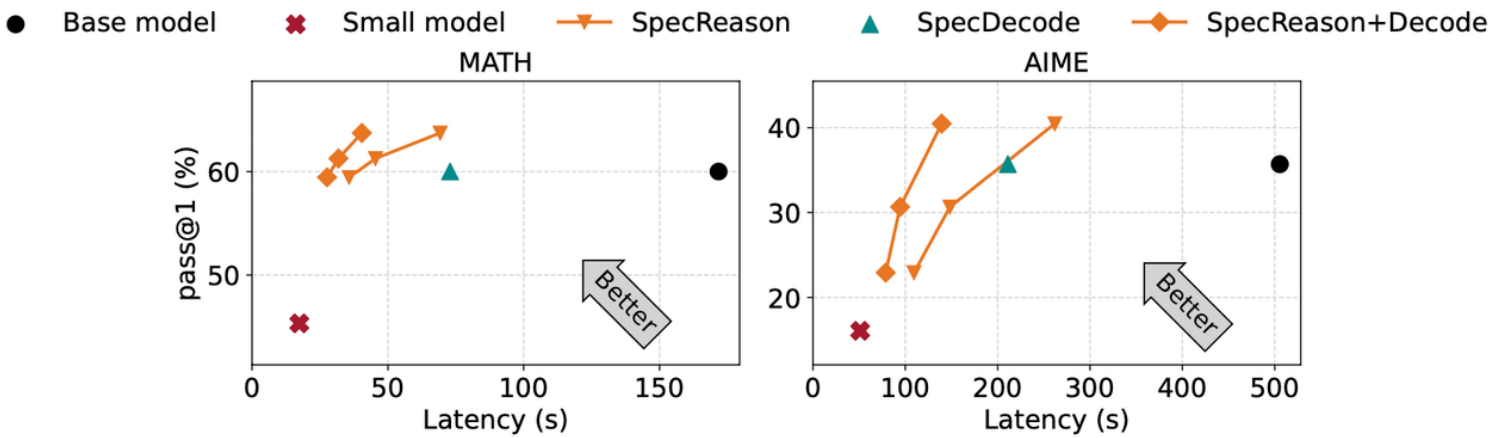
Judge Decoding：能否training-free?

- **Yes**!利用大模型的自身的语义理解能力
- **SpecReason**
  - LLMs的长程推理可以解耦为多个较短、较容易的小问题
  - 基于prompting，令大模型给小模型的proposal打分(0-10)，接受高得分的小模型proposal

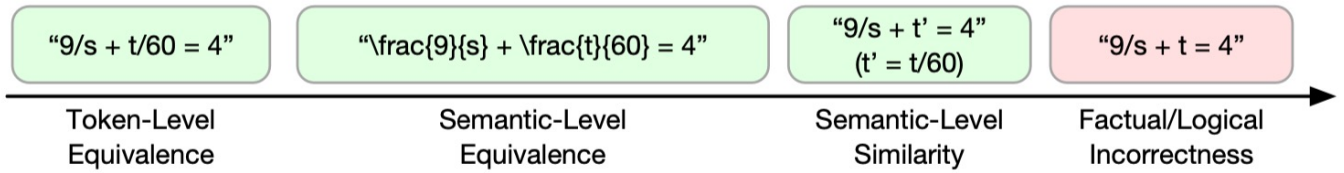


Pan et al. SpecReason: Fast and Accurate Inference-Time Compute via Speculative Reasoning. https://arxiv.org/abs/2504.07891

SpecReason：结果



比**spec decoding**还快一倍，在降低得分接受阈值后，性能依旧能有效维持（从左至右3，5，7）



多种**proposal**上的大模型打分

Pan et al. SpecReason: Fast and Accurate Inference-Time Compute via Speculative Reasoning. https://arxiv.org/abs/2504.07891

关注模型深度推理能力的同时，我们是不是忽视了什么？

Query

Carla is downloading a 200 GB file. Normally she can download 2 GB/minute, but 40% of the way through the download, Windows forces a restart to install updates, which takes 20 minutes. Then Carla has to restart the download from the beginning. How load does it take to download the file?

LLM

Incorrect Prediction by Qwen2.5-7B-Instruct

To find the total time it takes to download the file, we need to break down the process into steps.

Step 1: Calculate the time it takes to download 40% of the file.
- 40% of 200 GB is 0.4 * 200 = 80 GB.
- Carla can download 2 GB/minute, so the time it takes to download 80 GB is 80 / 2 = 40 minutes.

Step 2: Calculate the time it takes to download the remaining 60% of the file after the restart.
- 60% of 200 GB is 0.6 * 200 = 120 GB.
- Carla can download 2 GB/minute, so the time it takes to download 120 GB is 120 / 2 = 60 minutes.

Step 3: Add the time it takes to download the first 40% of the file, the time it takes to download the remaining 60% of the file, and the time it takes for the Windows restart.
- Time to download the first 40% of the file: 40 minutes.
- Time to download the remaining 60% of the file: 60 minutes.
- Time for the Windows restart: 20 minutes.
- Total time: 40 + 60 + 20 = 120 minutes.

Therefore, the total time it takes to download the file is \boxed{120 minutes}

Query

Let N denote the number of ordered triples of positive integers (a, b, c) such that $a, b, c \leq 3^6$ and $a^3 + b^3 + c^3$ is a multiple of 3^7. Find the remainder when N is divided by 1000.

LLM

Incorrect Prediction by DeepSeek-R1

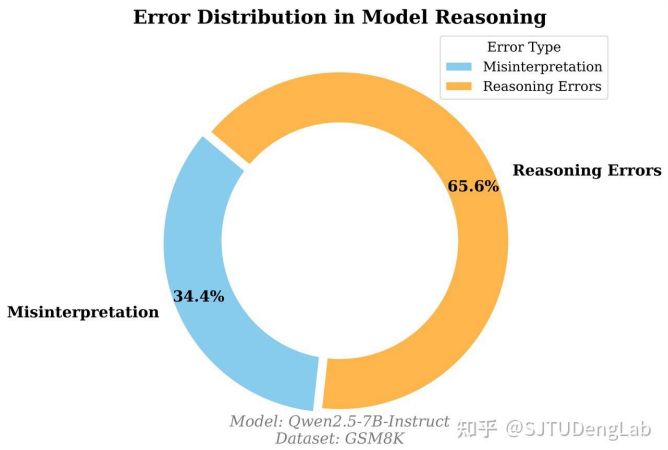
To solve the problem of finding the number of ordered triples (a, b, c) such that $a, b, c \leq 3^6$ and $a^3 + b^3 + c^3$ is a multiple of 3^7, we need to consider the structure of cubic residues modulo 3^7.

1. Understanding the Problem:
- We need to count triples (a, b, c) such that $a^3 + b^3 + c^3 \equiv 0 \pmod{3^7}$.
- Each a, b, c is a positive integer up to $3^6 = 729$.

2. Modular Arithmetic Analysis:
- The key insight is that the number of solutions to the equation $a^3 + b^3 + c^3 \equiv 0 \pmod{3^7}$ can be considered using properties of cubic residues and the structure of the problem.
- The total number of triples (a, b, c) is $(3^6)^3 = 3^{18}$.

3. Uniform Distribution Assumption:
- Assuming the distribution of cubic residues modulo 3^7 is uniform, the number of solutions should be approximately $\frac{3^{18}}{3^7} = 3^{11}$.

4. Verification:
- This assumption is validated by considering the problem in the context of 3-adic integers and the equidistribution of solutions modulo 3^7.
- The number 3^{11} modulo 1000 is calculated as follows:
- $3^{11} = 177147$
- The remainder when 177147 is divided by 1000 is 147.
Thus, the remainder when N is divided by 1000 is $\boxed{147}$.



统计**Qwen2.5-7B-Instruct** 在 **GSM8K** 测试集上的错误推理，超过**30%**的错误源自于“误解原始问题”

Qwen2.5 因为误解了“重新启动下载”这一事实而出错，而 **DeepSeek-R1** 则因为假设了“立方余数”的分布问题（该问题在题目中并未提及）而失败

对于推理任务的关注，不应仅停留在提升推理能力上，应意识到推理忠实度同样甚至更加重要！

Query

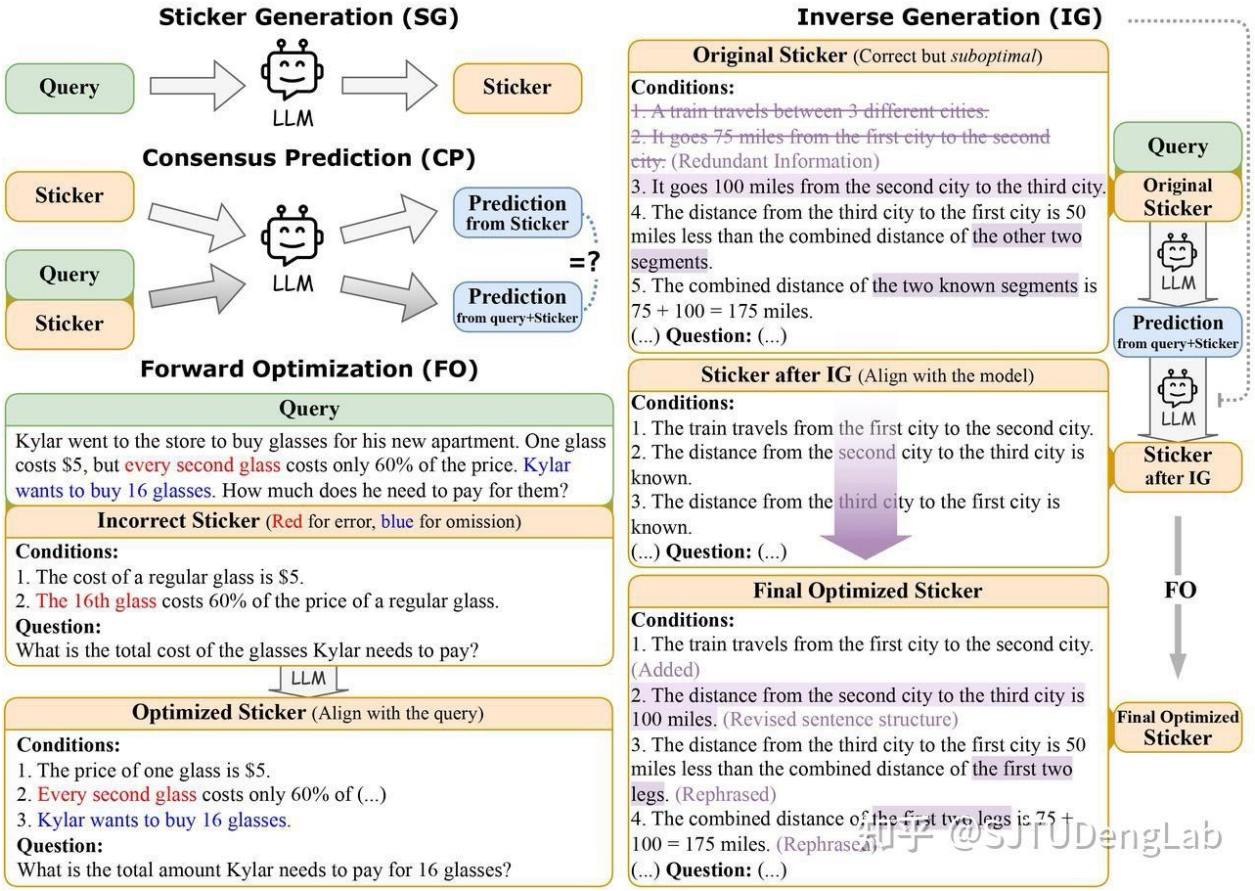
Josh decides to try flipping a house. He buys a house for $80,000 and then puts in $50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

Sticker

Conditions:
1. Josh buys a house for $80,000.
2. He spends $50,000 on repairs.
3. The value of the house increases by 150%.

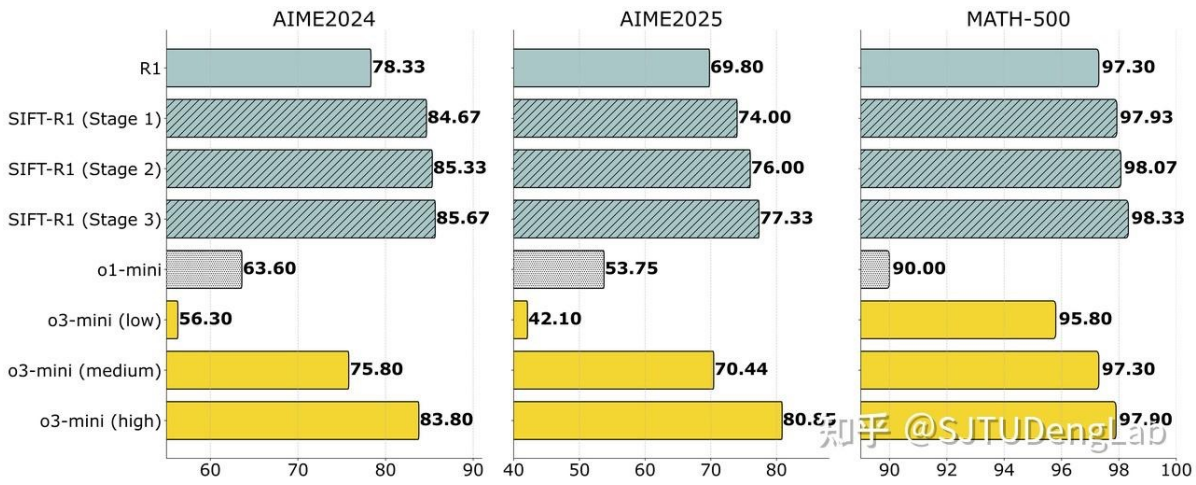
Question:
What is the total profit Josh made from flipping the house?

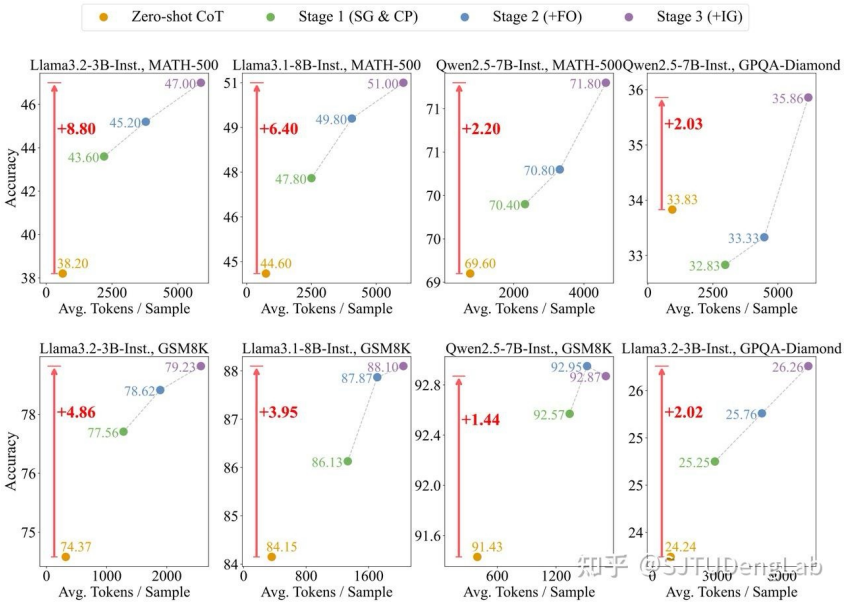
- **SIFT（Stick to the Facts）**：动态生成结构化摘要（称为**"Sticker"**）将推理过程显式锚定在问题中的基本事实上
- 基于单独的 **Sticker** 和 **query+Sticker** 进行两次回答：如一致，返回预测；否则，双向优化 **Sticker**来实现更忠实的推理



对于推理任务的关注，不应仅停留在提升推理能力上，应意识到推理忠实度同样甚至更加重要！

- 把**DeepSeek-R1** 在 **AIME2024** 准确率从 **78.33%** 提升至 **85.67%**，在**AIME2025**上 从 **69.80%** 提升至 **77.33%**，刷新开源模型 **SOTA**
- 把**Llama3.2-3B-Instruct** 在 **MATH-500** 提升了 **8.80** 个点的准确度

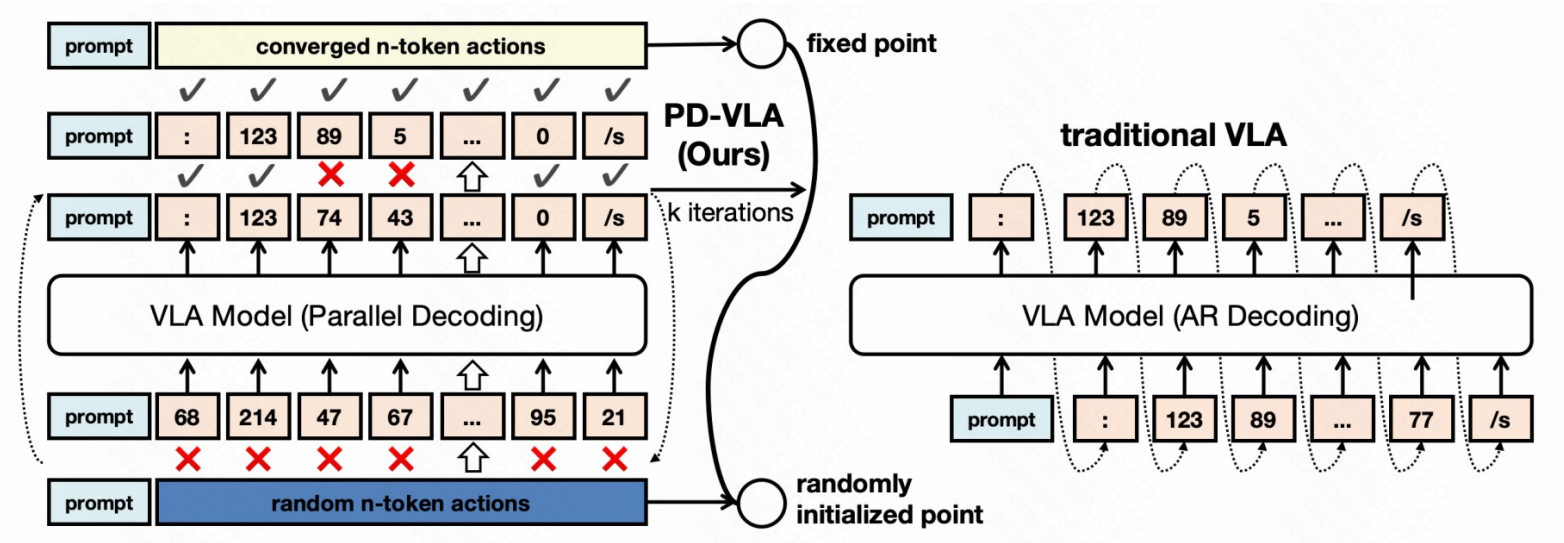




Zeng, Huang, Li, Deng. SIFT: Grounding LLM Reasoning in Contexts via Stickers. Arxiv 2502.14922.

4、高效多模态生成模型在**VLA/Agent**中的应用

一致性大语言模型：在VLA模型中的应用，加速动作预测



仿真实验验证，在机械臂（7 自由度）执行效率达到基础 VLA 模型的 2.52 倍

Song, Wenxuan, et al. Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. arXiv:2503.02310 (2025).

总结

- 高效多模态生成方法：
  - 扩模态统一建模
  - 模型加速
  - 深度推理

- 应用：
  - **VLA/Agent**

感谢各位专家！敬请批评指正！

邮箱：zhijied@sjtu.edu.cn

主页：https://thudzj.github.io/