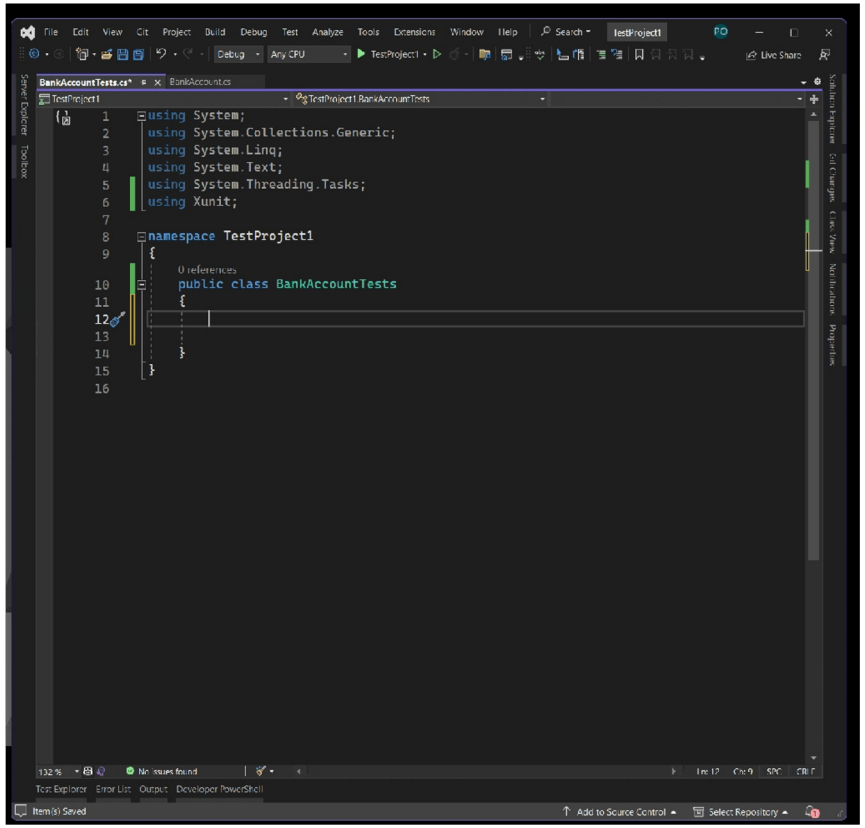
AIGC大模型的高效推理方法

邓志杰

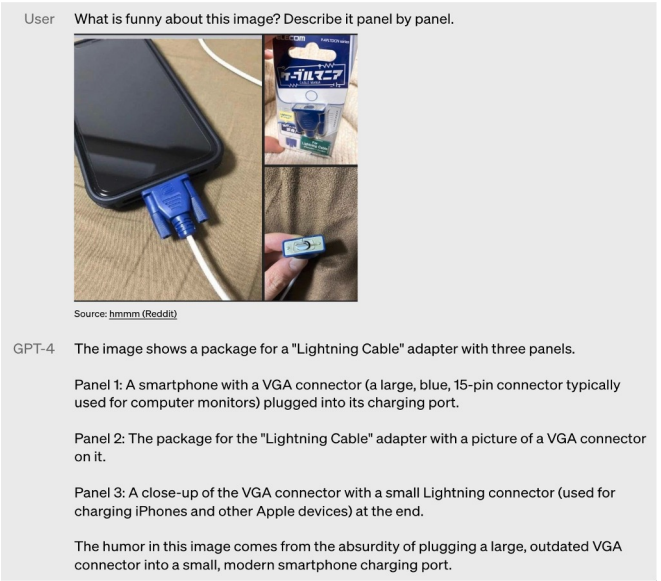
上海交通大学 清源研究院

背景：大语言模型成为广泛下游任务的基石

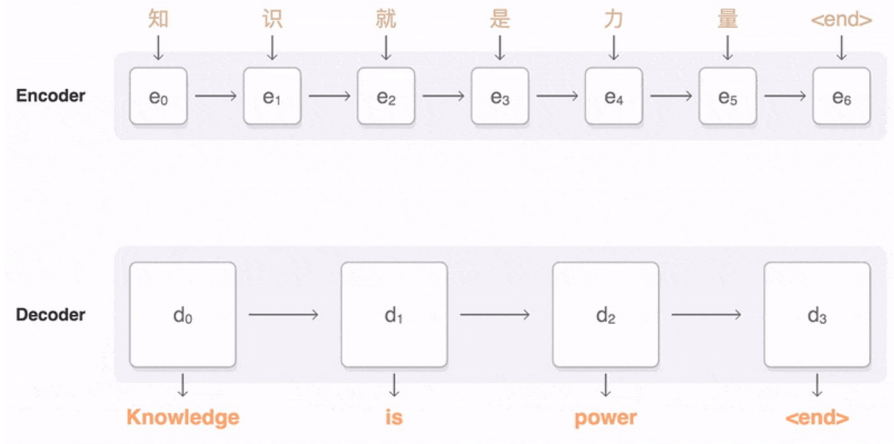


Code generation

Agents

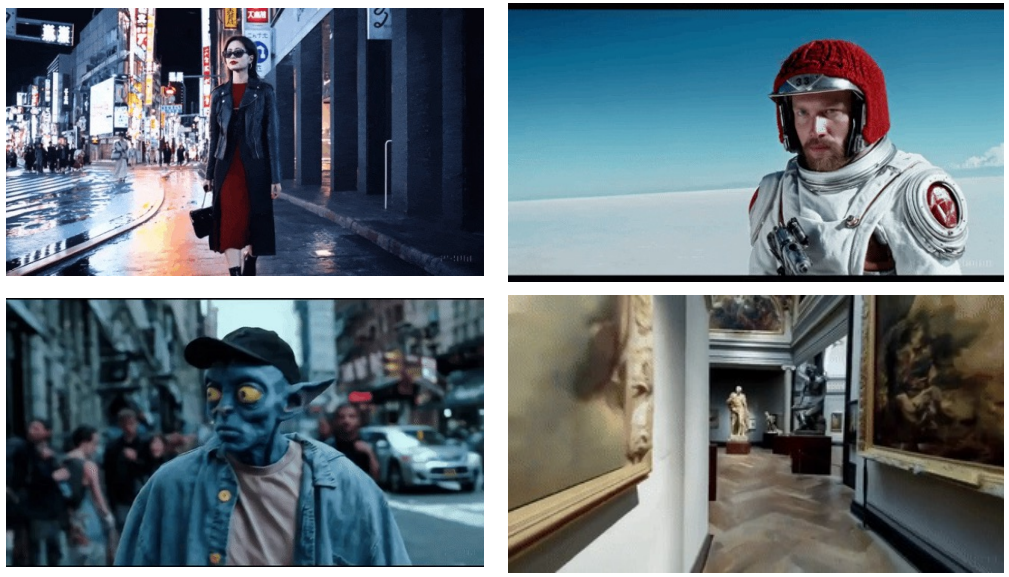


(Multimodal) Chatbot



Machine translation

背景：大扩散模型正在形成“世界模拟器”



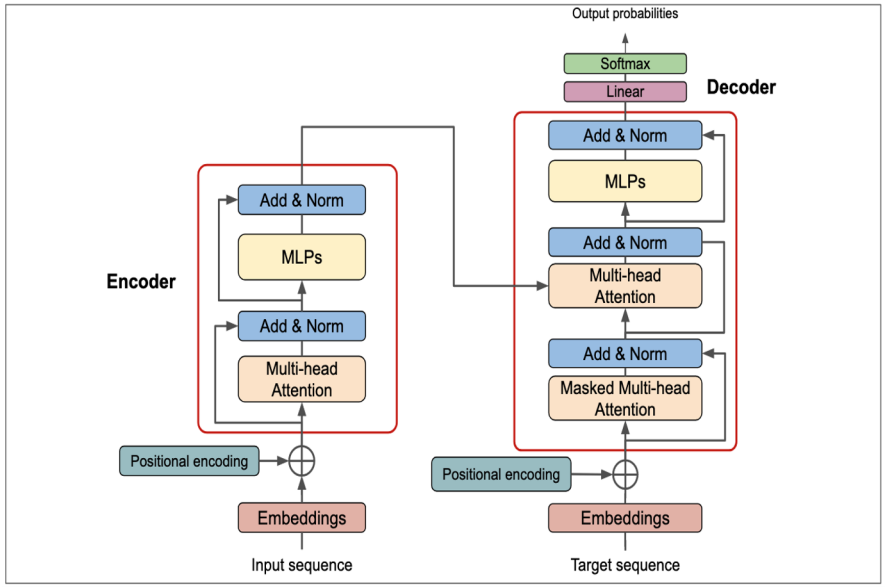
Sora by OpenAI

Vidu by ShengShu

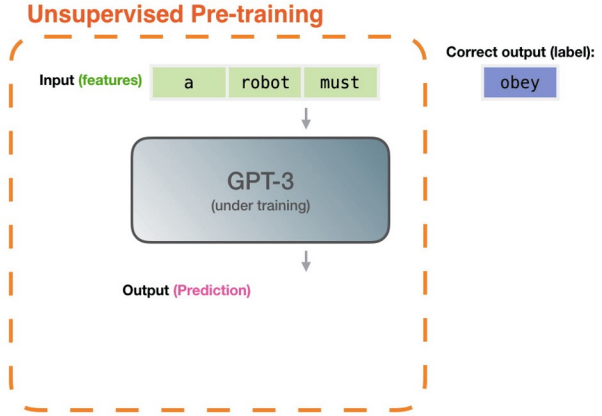
W.A.L.T

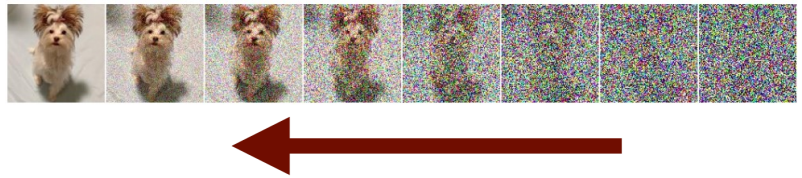
https://openai.com/index/sora/; https://www.vidu.io/text-to-video-ai; https://walt-video-diffusion.github.io/

AIGC大模型的趋势

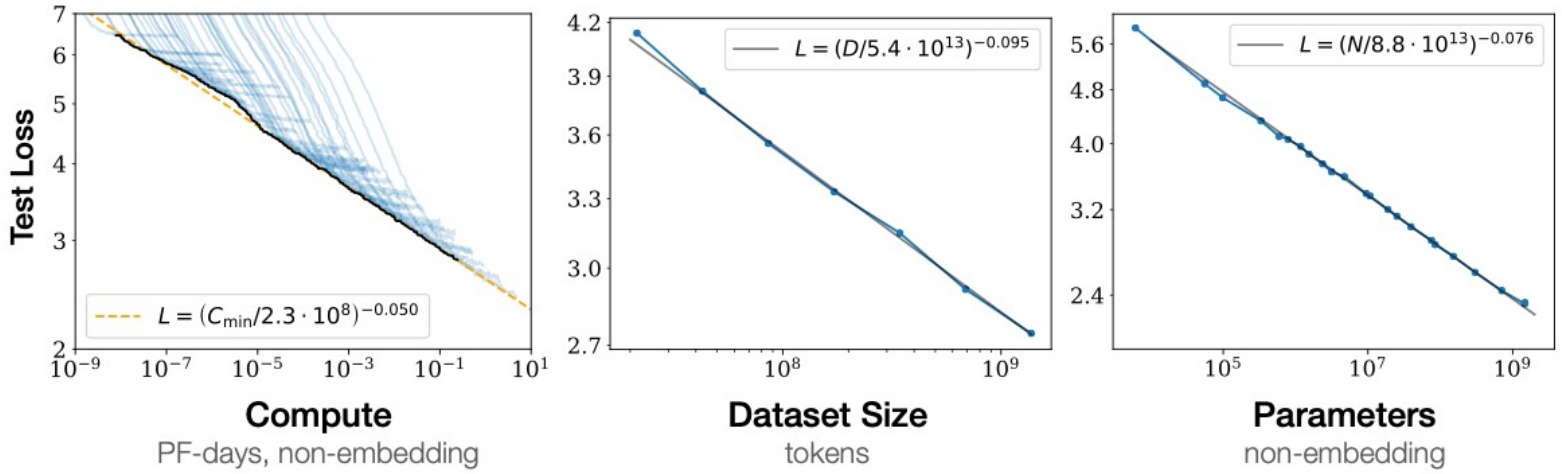


架构趋于统一





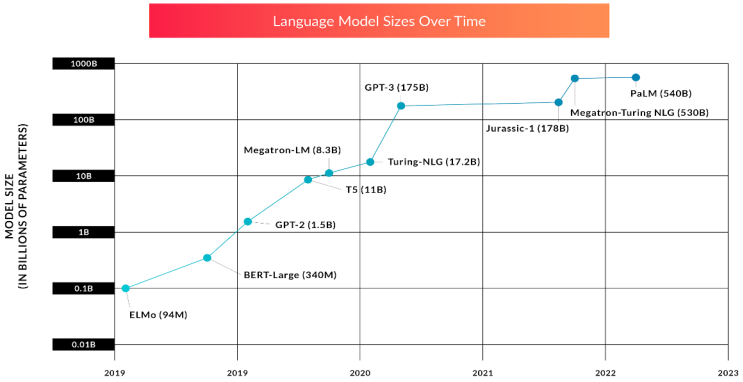
学习方式存在两个主流



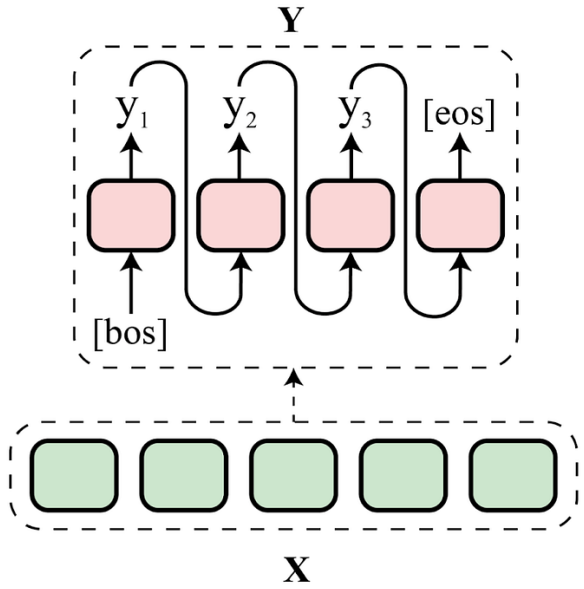
Scaling law持续发挥作用

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

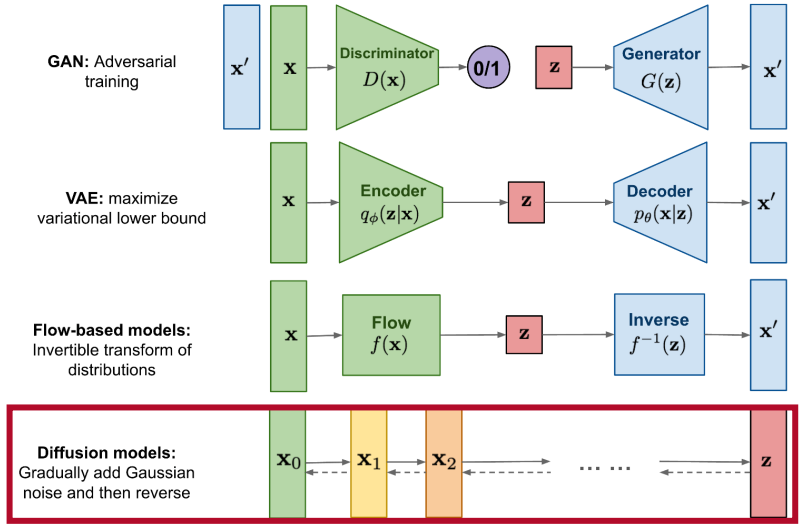
AIGC大模型推理低效，导致高部署成本、差用户体验



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

低效的顺序推理过程

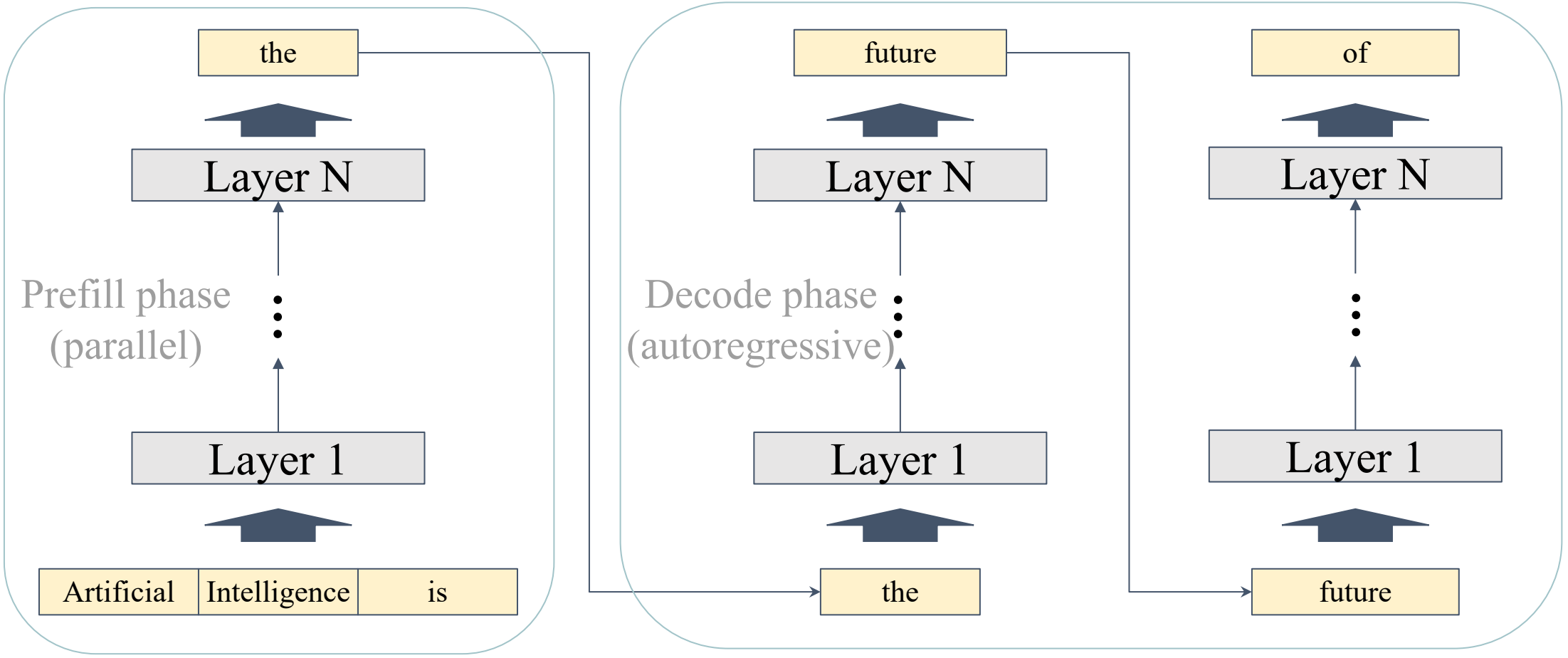
AIGC大模型的高效推理方法

- 大语言模型的并行推理方法（**15min**）
  - 投机解码
  - 一致性蒸馏
- 大扩散模型的低步数推理方法（**15min**）
  - 采样器设计
  - 一致性蒸馏
- 大模型架构、序列状态、缓存等方面的优化方法（**5min**）

AIGC大模型的高效推理方法

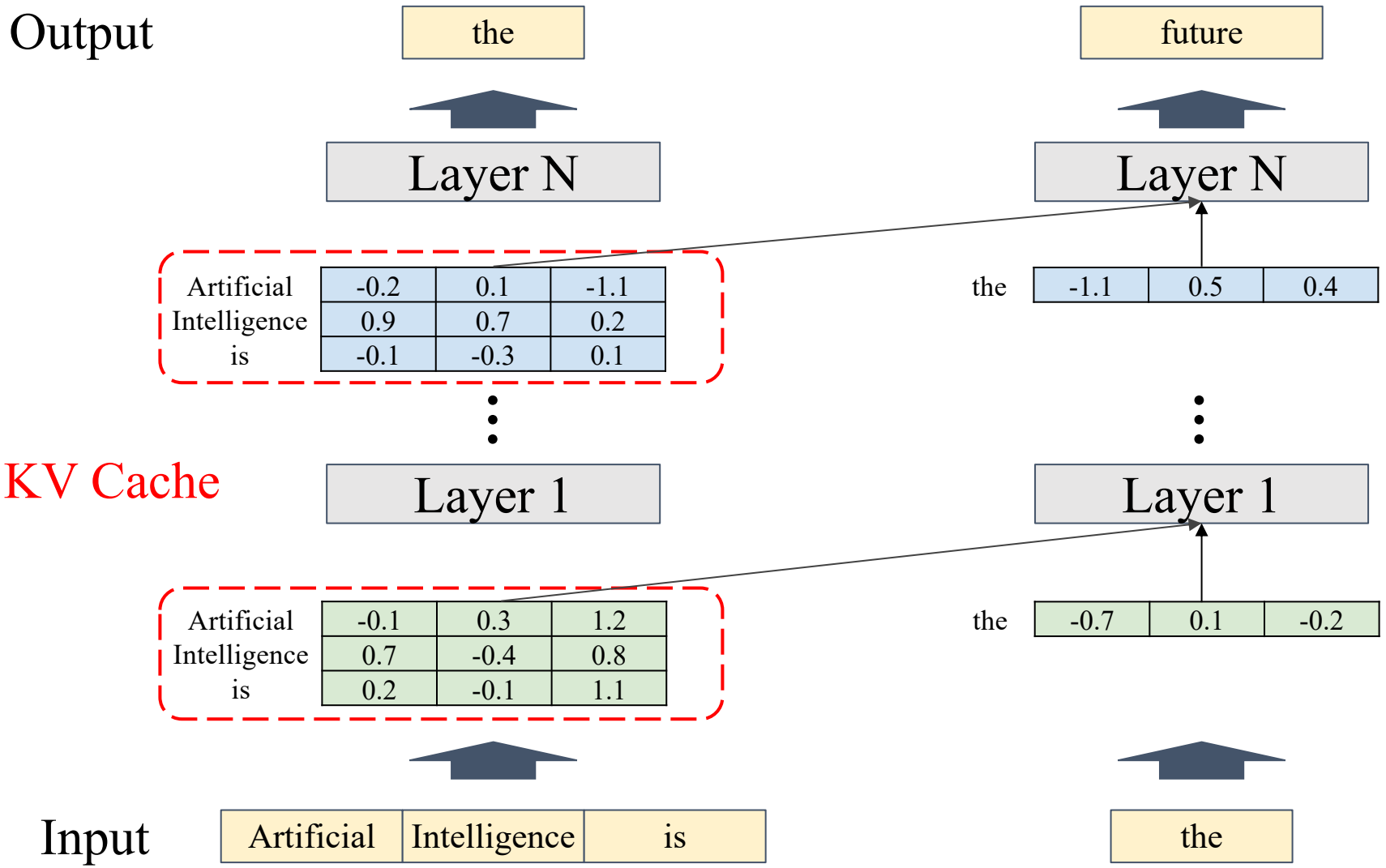
- 大语言模型的并行推理方法（**15min**）
  - 投机解码
  - 一致性蒸馏
- 大扩散模型的低步数推理方法（**15min**）
  - 采样器设计
  - 一致性蒸馏
- 大模型架构、序列状态、缓存等方面的优化方法（**5min**）

大语言模型的推理过程



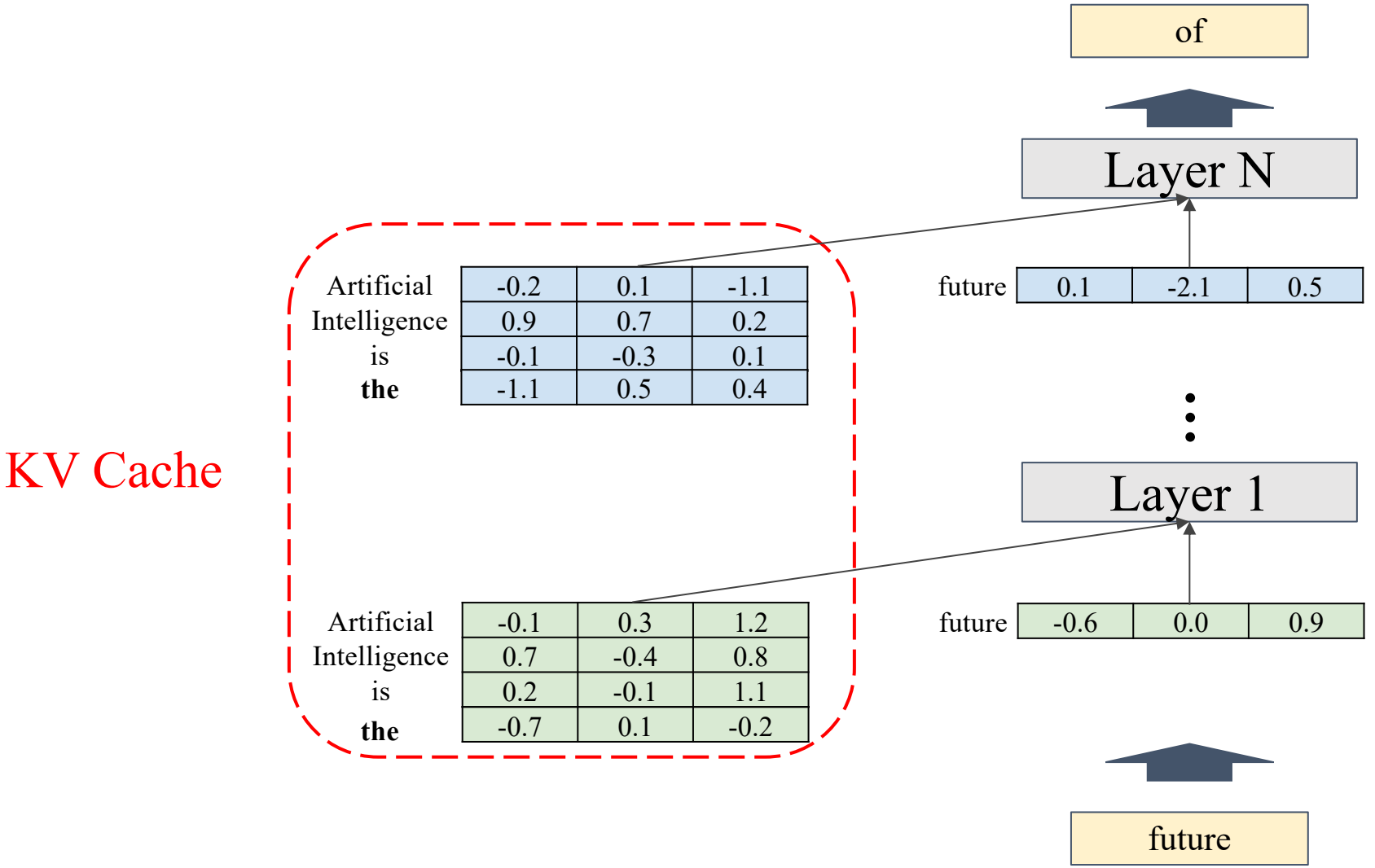
重复生成过程直到达到最大生成长度(e.g., 2048 tokens) 或生成出结束符(e.g., “<|end of sequence|>”)

KV Cache


Output
the
future
Layer N
Layer N
Artificial
Intelligence
is
-0.2
0.1
-1.1
0.9
0.7
0.2
-0.1
-0.3
0.1
the
-1.1
0.5
0.4
KV Cache
Layer 1
Layer 1
Artificial
Intelligence
is
-0.1
0.3
1.2
0.7
-0.4
0.8
0.2
-0.1
1.1
the
-0.7
0.1
-0.2
Input
Artificial
Intelligence
is
the

KV Cache

Output



Input

大语言模型推理过程的两个特点

- **Prefill phase计算密集，decode phase是memory-IO bound**
  - between High Bandwidth Memory (HBM) and Static Random Access Memory (SRAM)
- **得益于高效的并行计算，prefill几个tokens的时间与后续每个token生成时间接近**

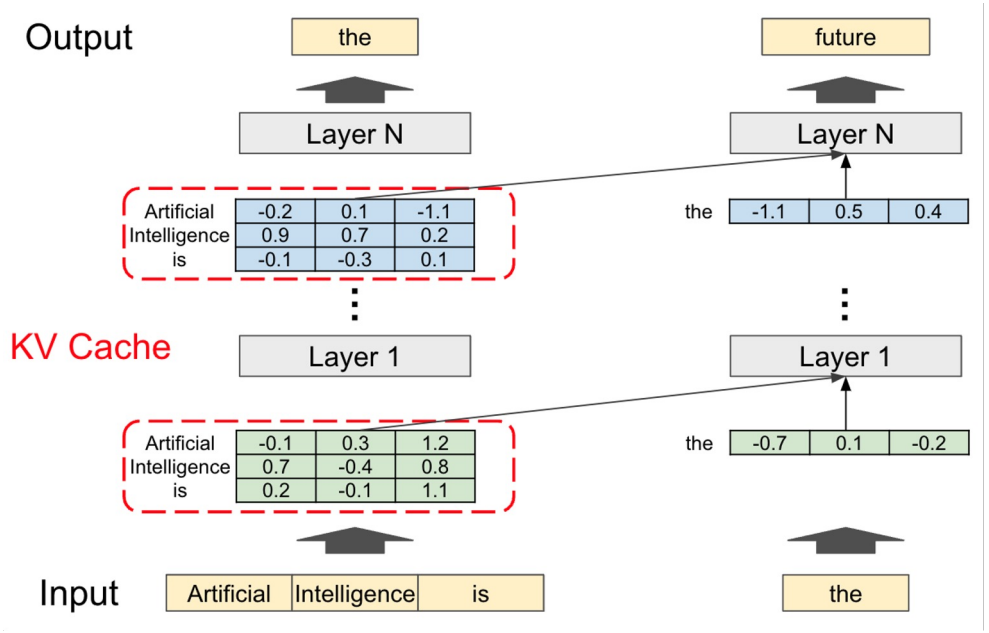


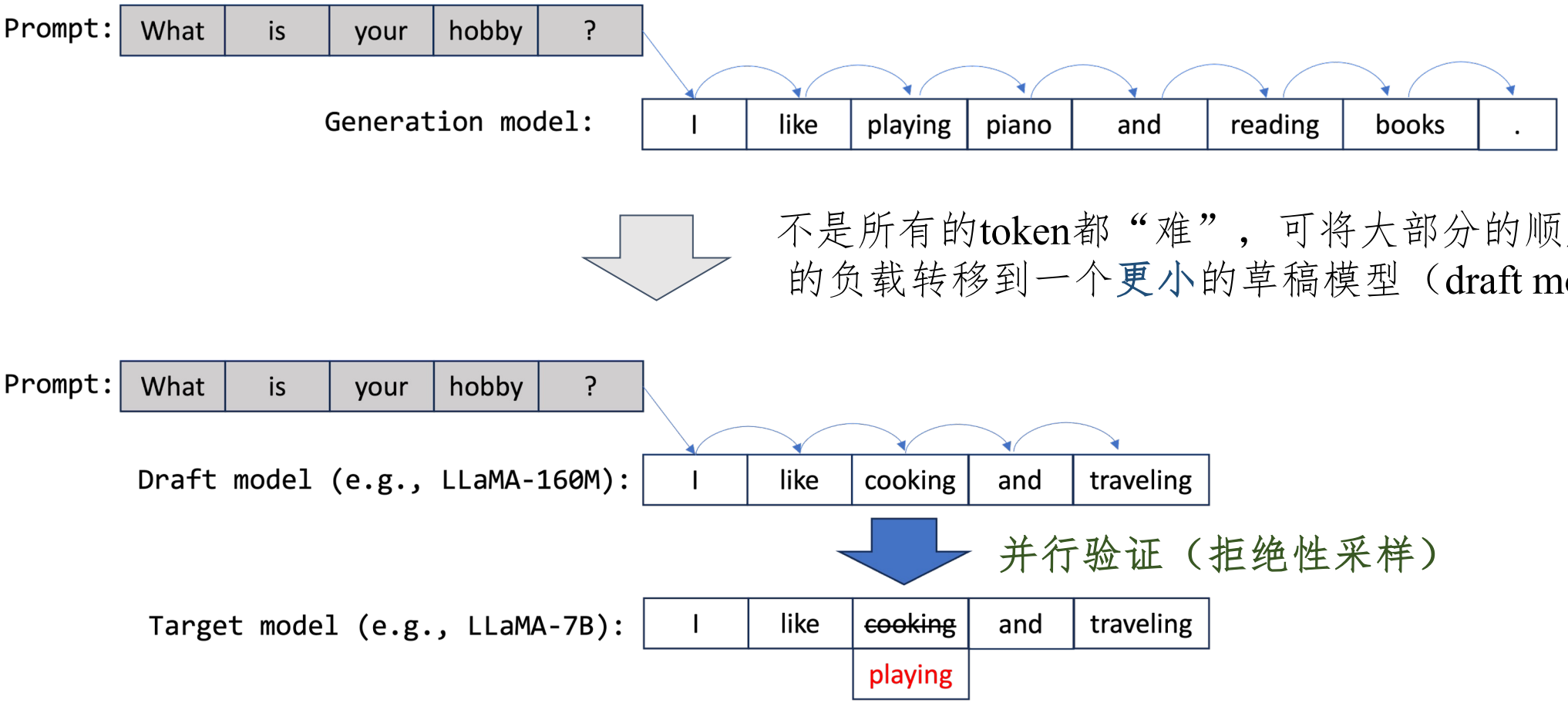
Table 1. Examples of neural network operations with their arithmetic intensities. Limiters assume FP16 data and an NVIDIA V100 GPU.

Operation	Arithmetic Intensity	Usually limited by...
Linear layer (4096 outputs, 1024 inputs, batch size 512)	315 FLOPS/B	arithmetic
Linear layer (4096 outputs, 1024 inputs, batch size 1)	1 FLOPS/B	memory
Max pooling with 3x3 window and unit stride	2.25 FLOPS/B	memory
ReLU activation	0.25 FLOPS/B	memory
Layer normalization	< 10 FLOPS/B	memory

https://docs.nvidia.com/deeplearning/performance/dl-performance-gpu-background/index.html#understand-perf

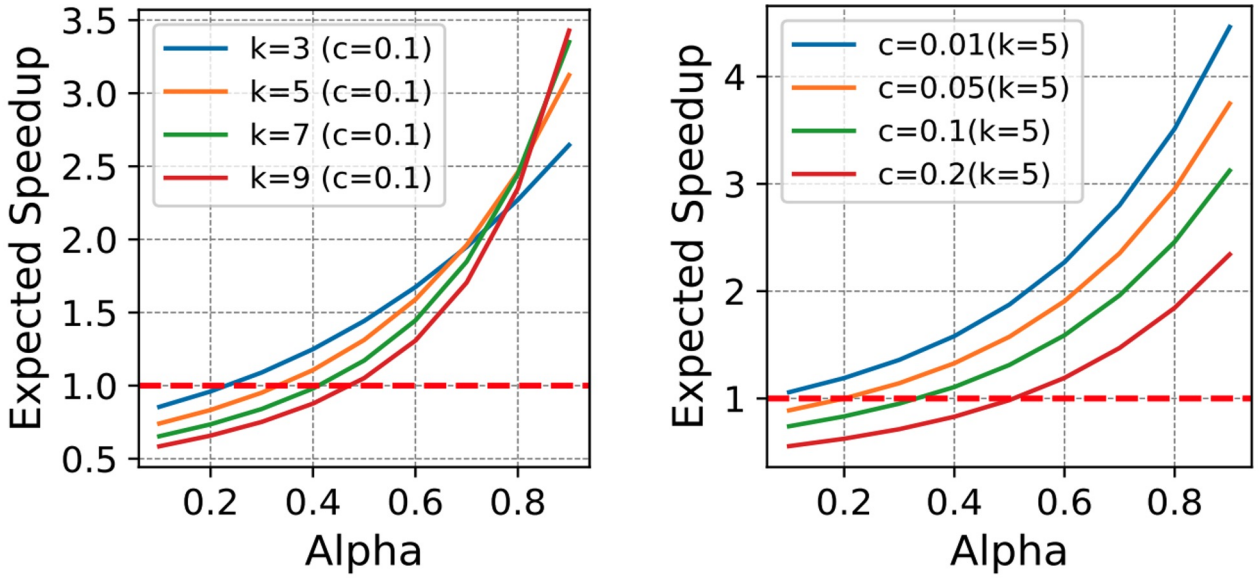
并行推理方法：投机解码（Speculative decoding，SD）

KV cache和模型权重的加载开销成倍下降



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

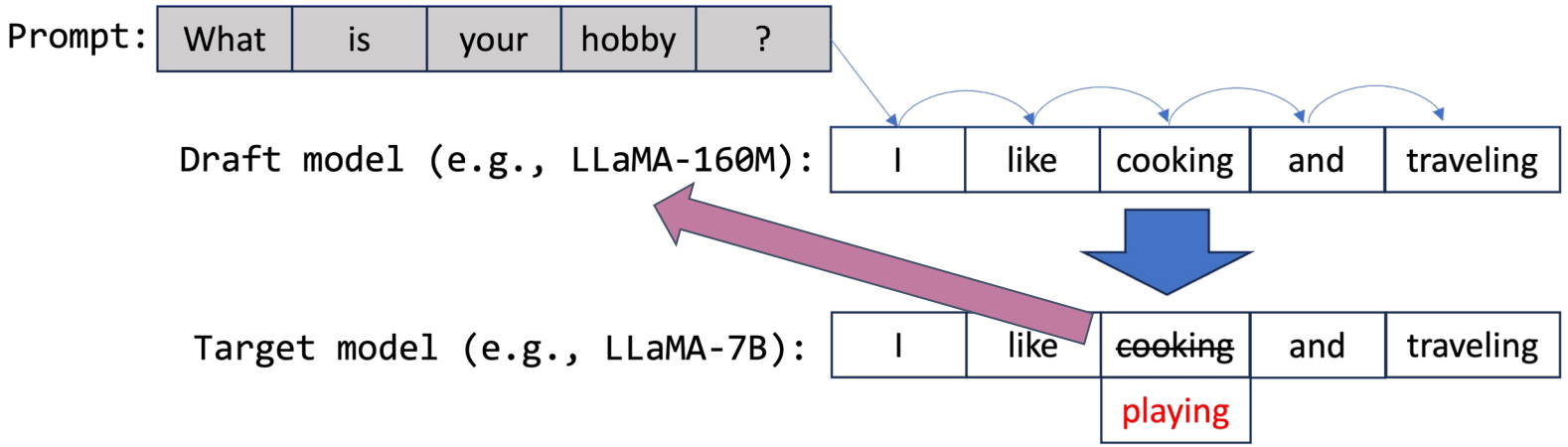
投机解码的工作条件



c: 草稿和目标模型的速度比；k: 每一步提议的token数； Alpha: token acceptance rate

- **结论：更高的token acceptance rate带来显著提高的加速效果**
  - 草稿模型必须足够逼近目标模型，同时保持较小的规模

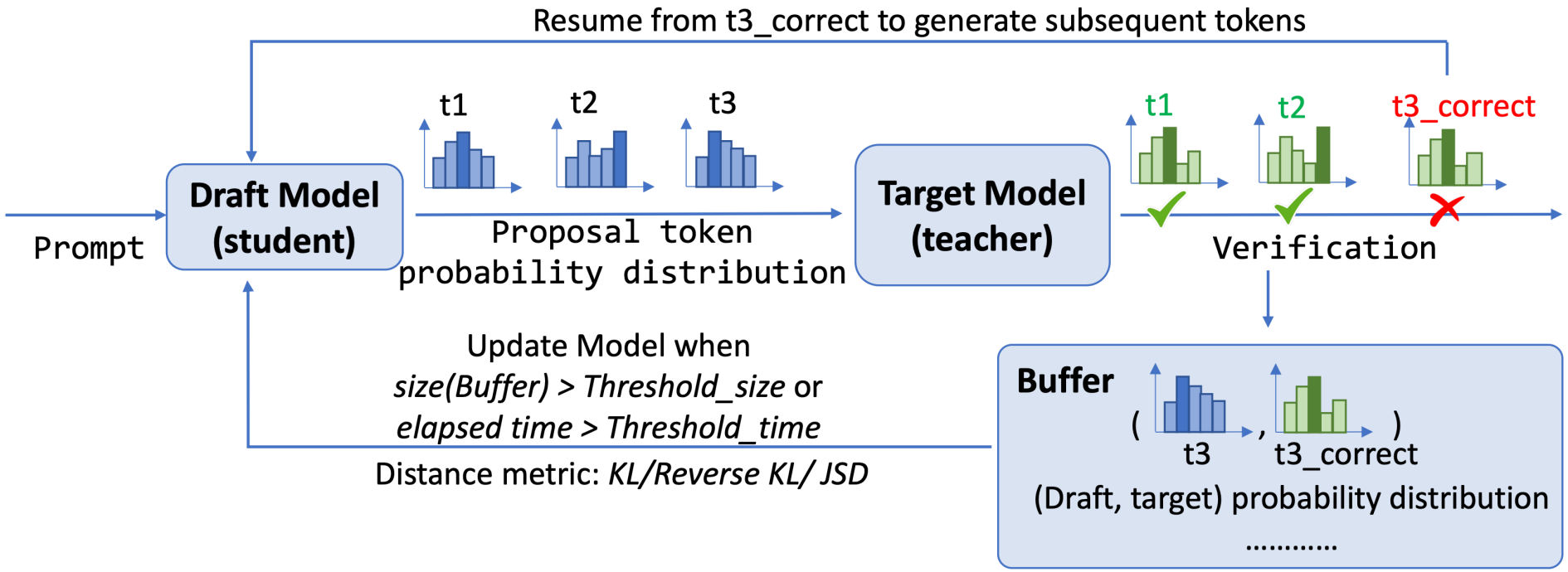
提高token acceptance rate的两个机会



- 投机解码可以检测草稿模型的错误并提供修正
  - 这些信息可以“免费”改进草稿模型，提高tokens acceptance rate，无需额外的标注成本
- 基于投机解码的**LLM推理系统中的空闲算力（“spare” FLOPs）**

在线的投机解码（OSD）：在线蒸馏+投机解码

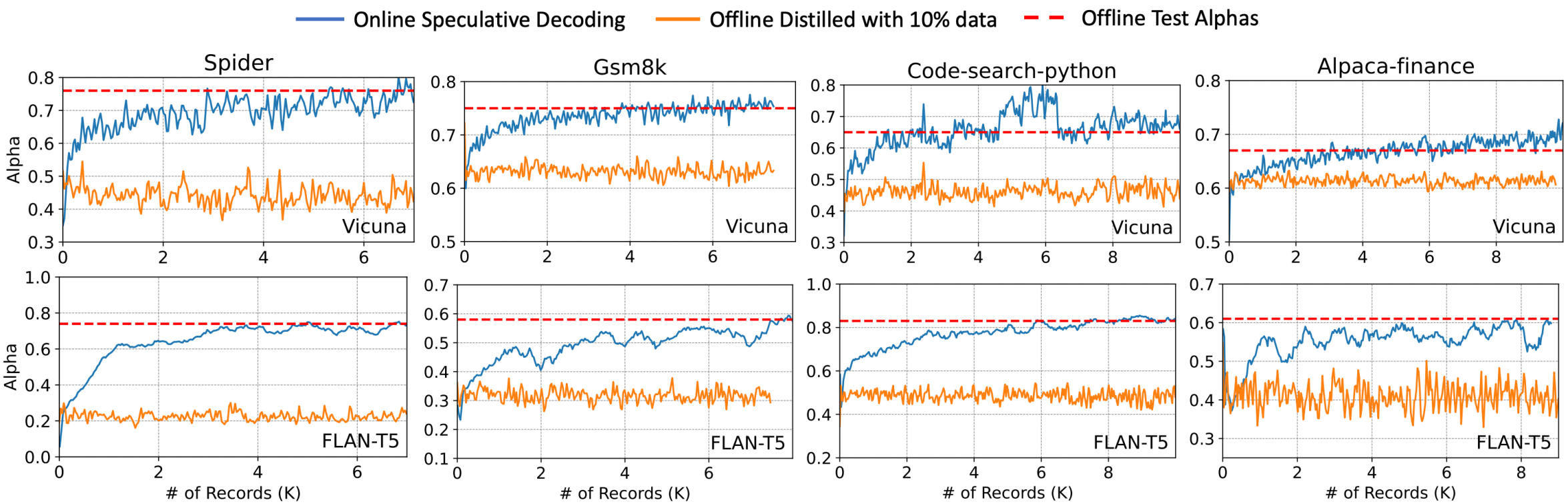
Open domain情况下，draft model快速适配query dist.



- 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.

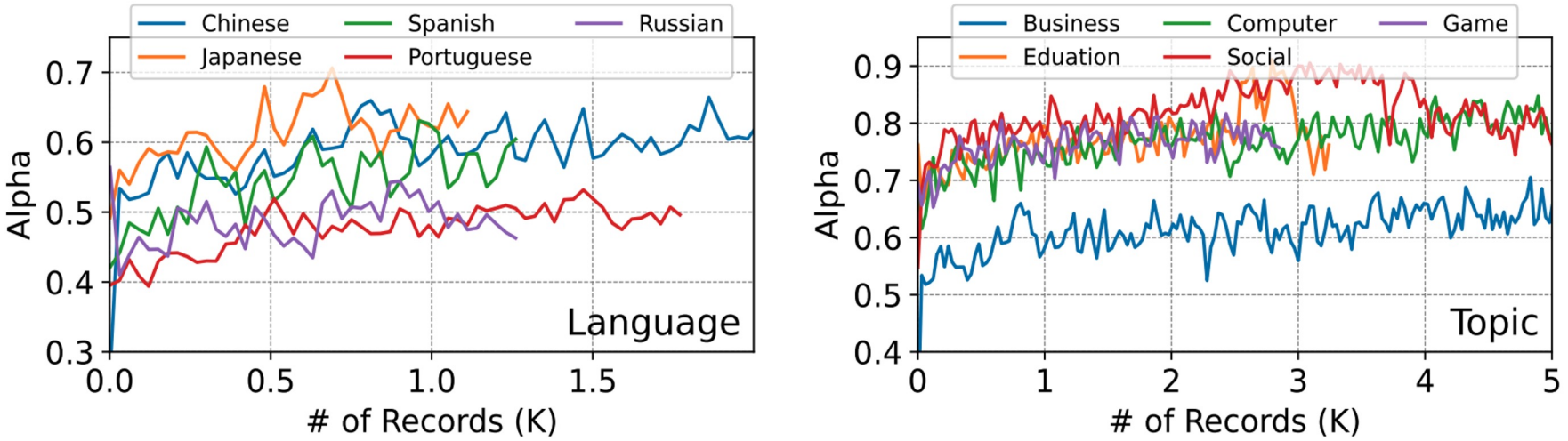
在线的投机解码：快速提高草稿模型的token acceptance rate



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.

在线的投机解码：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.

在线的投机解码：其他结果

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

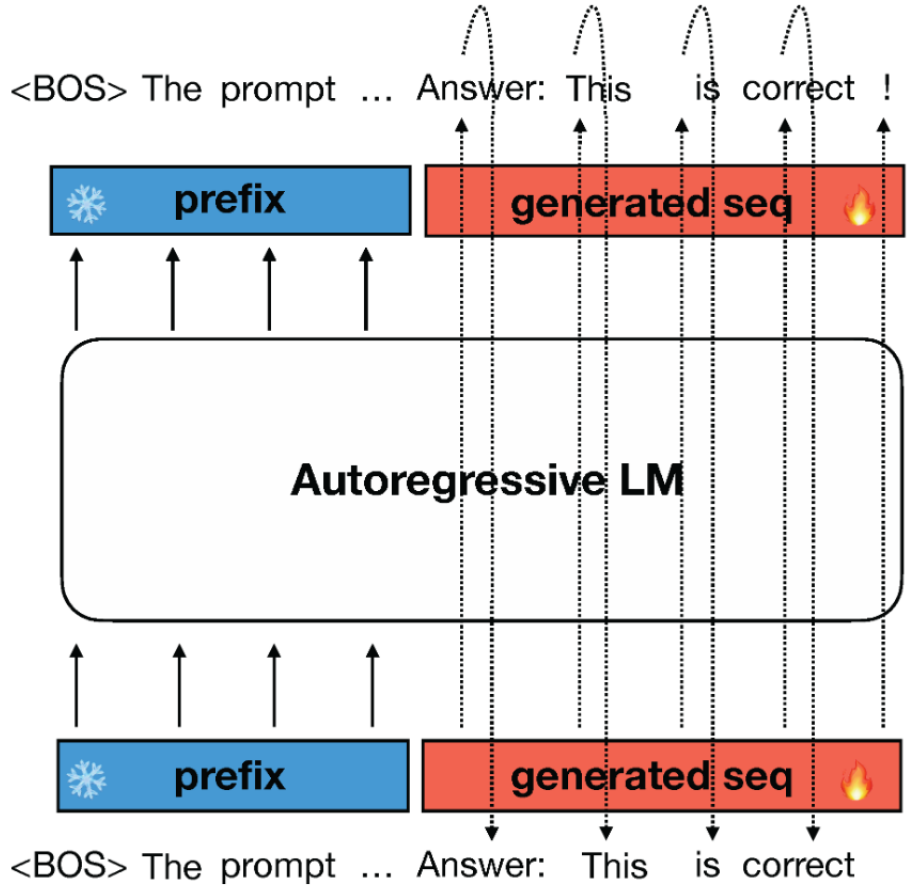
Dataset	**Spider**	**Gsm8k**	**Alpaca-Finance**	**Code-Python**
Tokens with the greatest precision increase	AV, SELECT, first, ⟨EOS⟩, template, SUM, G, COUNT, \n, city, WHERE, ';, (, IST, id	⟨EOS⟩, >>, +, To, <<, this, =, %, know, are, We, calculate, be, The, have	1, Here, (, :, provide, depends, However, 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

语言模型自身仍是顺序解码器

- 自回归的学习方式

- 语言模型可以一次预测出多个**tokens**吗？



Jacobi decoding：将大语言模型变为并行解码器的初试

- 给定大语言模型p, 同时预测 n个token等价于求解:

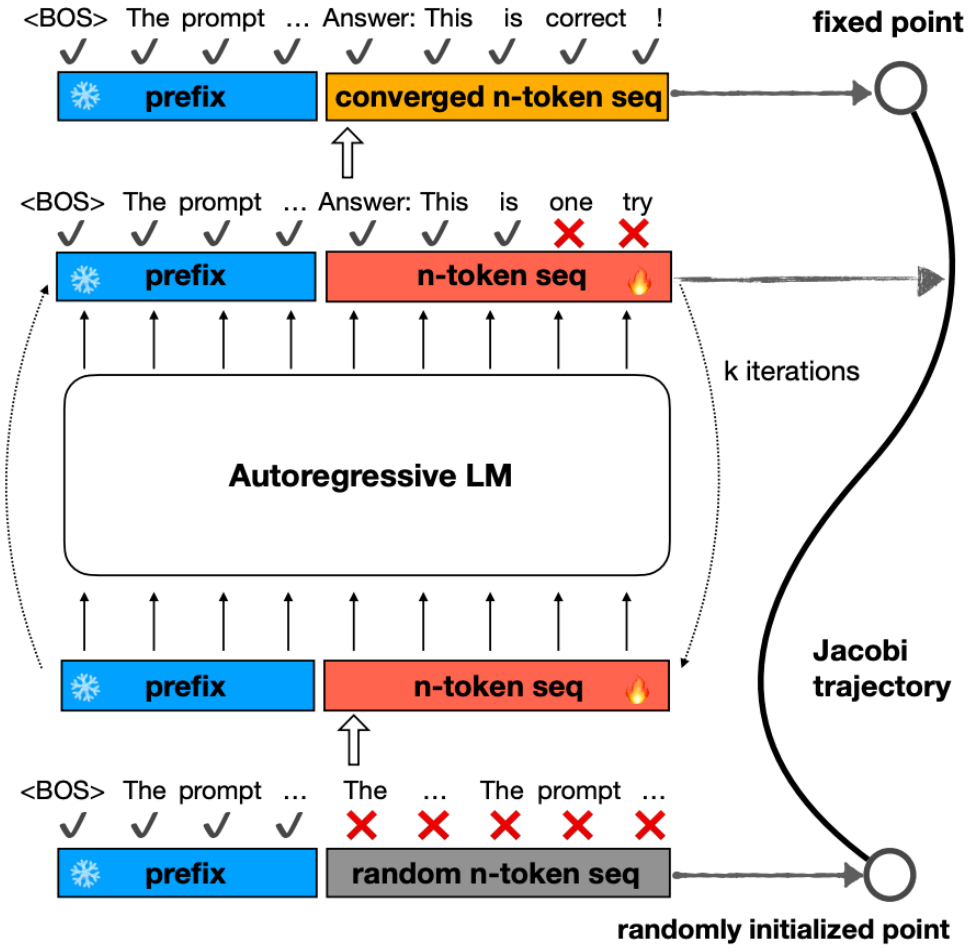
$$f(y_i, \boldsymbol{y}_{<i}, \boldsymbol{x}) = 0 \text{ for } i = 1, \ldots, n.$$

其中

$$f(y_i, \boldsymbol{y}_{<i}, \boldsymbol{x}) := y_i - \arg\max_y p(y|\boldsymbol{y}_{<i}, \boldsymbol{x})$$

- **这n个方程虽存在依赖，但可被Jacobi 不动点迭代法并行求解，步数不超过n，生成质量可保证**

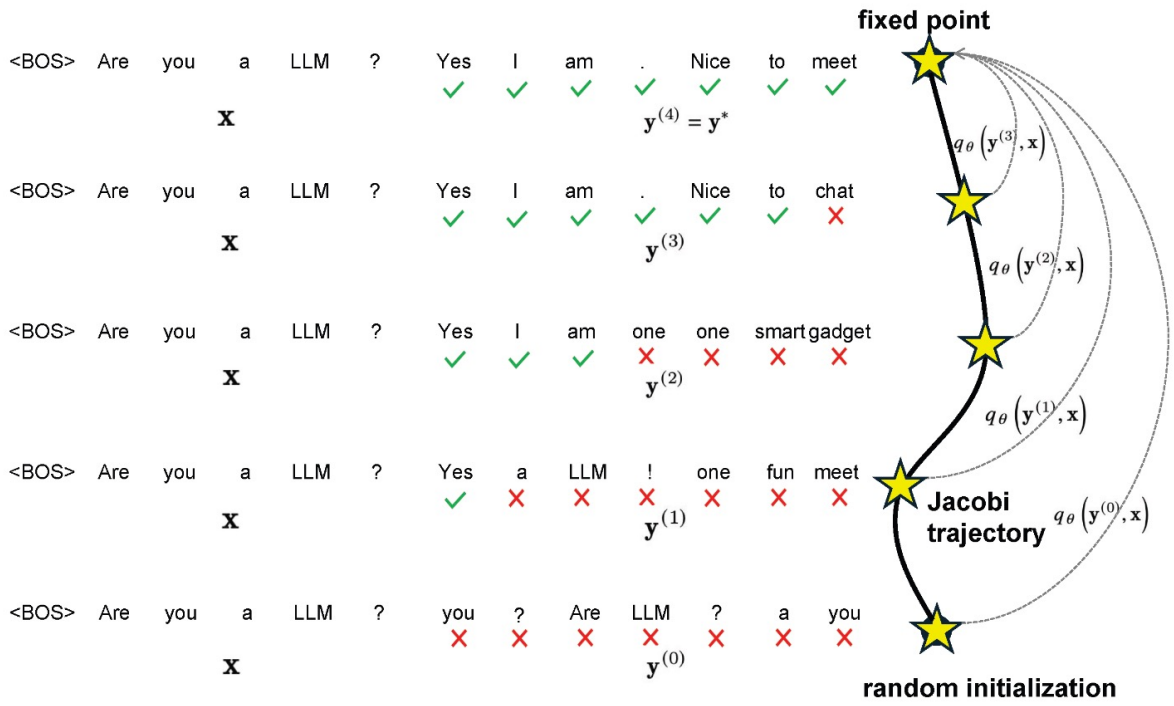
- 但实际效果差（如：仅1.05倍提升）
  - **原因：模型训练时未学过如何预测多个tokens**



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.

并行推理方法：一致性大语言模型（Consistency LLMs, CLLMs）

- 如何习得预测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：一致性损失函数

fixed point

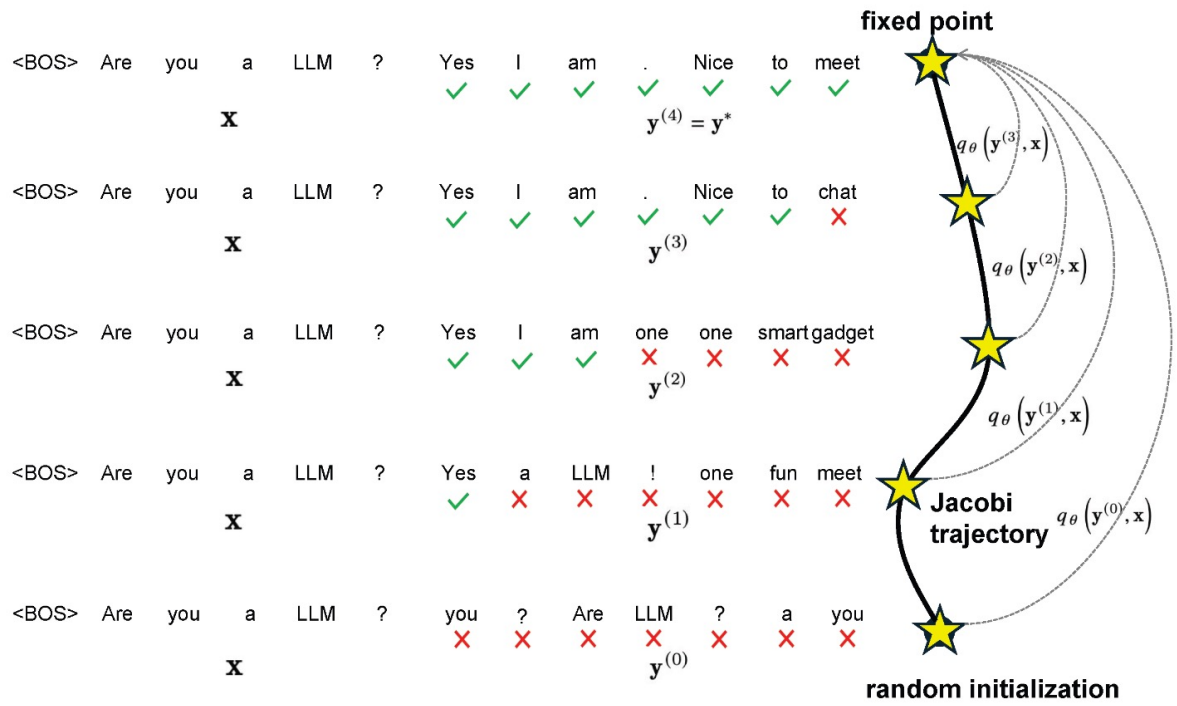
$$\mathcal{L}_{\mathrm{GC}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathcal{J})\sim\mathcal{D},\mathbf{y}\sim\mathcal{J}}\left[\sum_{i=1}^{n} D(q_{\theta^-}(\cdot|\mathbf{y}^*_{:i},\mathbf{x}))||q_\theta(\cdot|\mathbf{y}_{:i},\mathbf{x})\right]$$

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

$D(\cdot||\cdot)$是分布间距离度量

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

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：加速效果

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：系统性比较

- 至多3.6倍加速
- 相比于Medusa2, 不需要模型架构上的改变
- 有保证的生成质量

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

一致性大语言模型：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

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

一致性大语言模型：加速根源

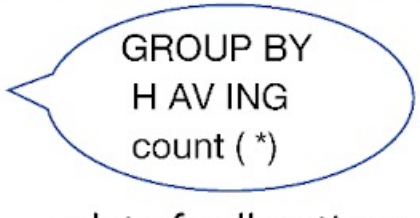
Target LLM

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

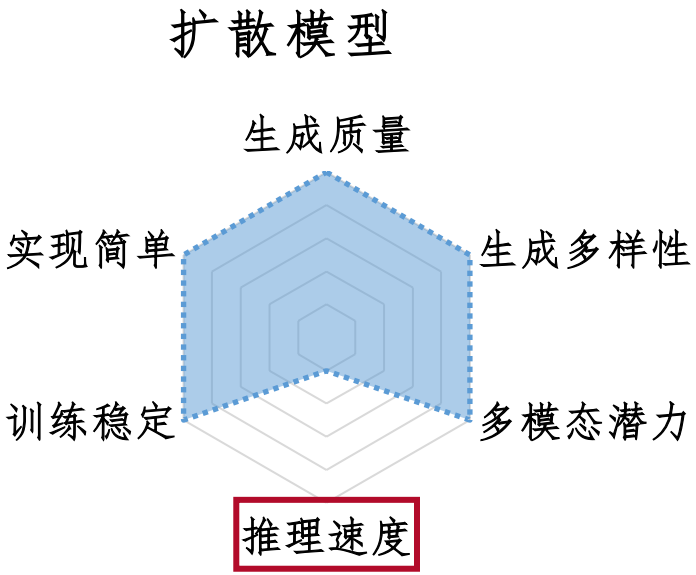
- **Fast forwarding**: 多个连续的tokens在单次网络传播中被正确预测
- **Stationary tokens**: 提前被正确预测，并在后续迭代中保持不变，即使之前的tokens存在错误

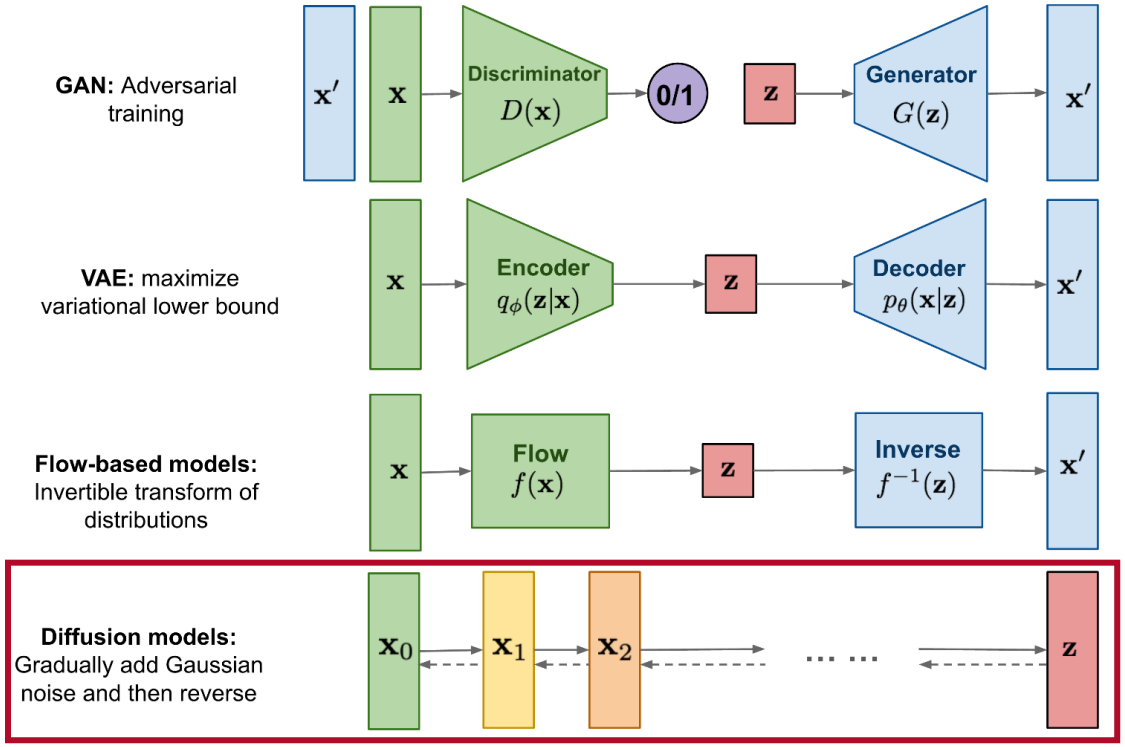
Liu, Hu, Bailis, Cheung, Deng, Stoica & Zhang. Online Speculative Decoding. ICML 2024.

AIGC大模型的高效推理方法

- 大语言模型的并行推理方法（15min）
  - 投机解码
  - 一致性蒸馏
- 大扩散模型的低步数推理方法（15min）
  - 采样器设计
  - 一致性蒸馏
- 大模型架构、序列状态、缓存等方面的优化方法（5min）

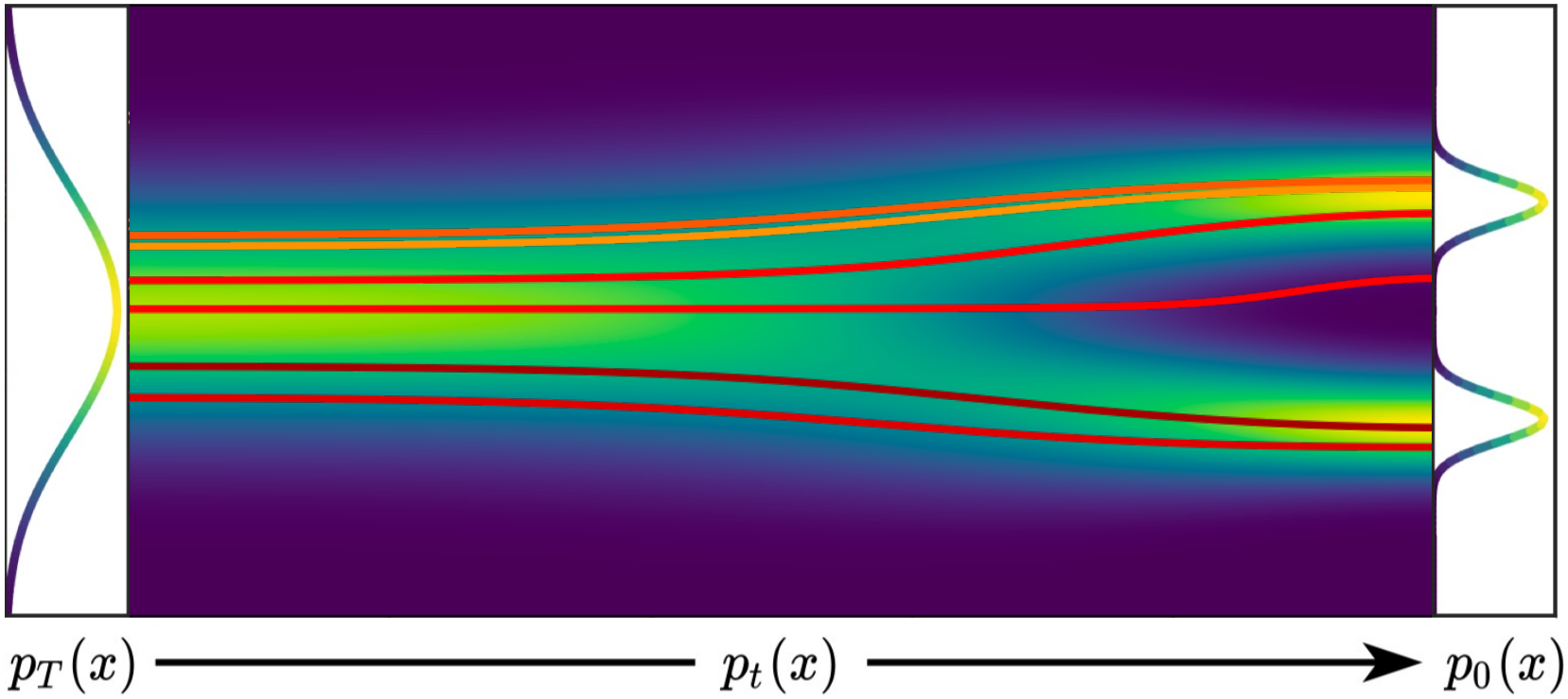
采样效率是扩散模型的瓶颈问题：每个去噪步都需要一次模型前传





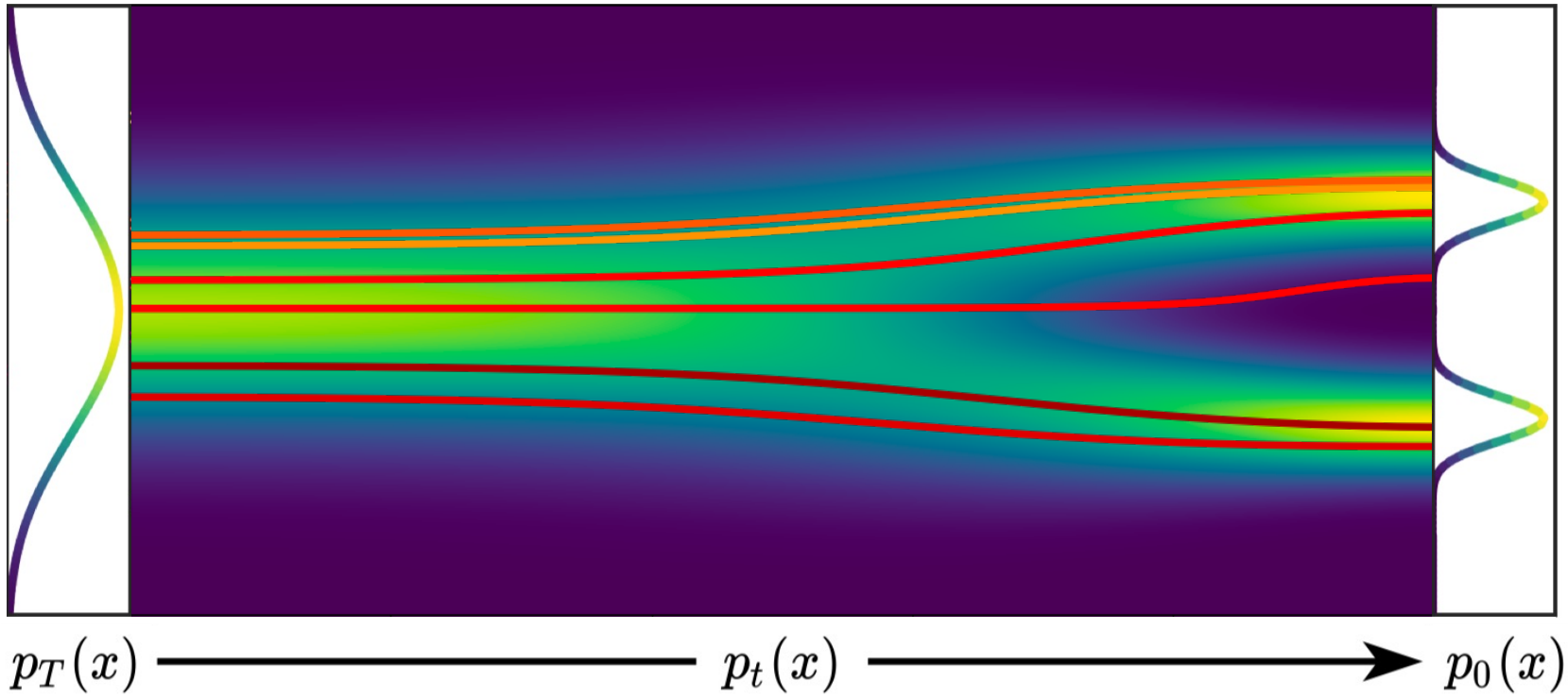
原则上需要迭代很多次，每次需调用神经网络

出发点：常微分方程的视角



时时刻刻边缘分布相同，但是路径上没有噪声！

扩散模型采样等价常微分方程离散化



Sampling method	Steps to converge
Traditional **SDE** Solvers	~200
Traditional **ODE** Solvers	~100

Checkpoint: CIFAR-10, VP

[Song et al., ICLR'21]

DPM-Solver：面向扩散概率模型的常微分方程离散化

- 针对扩散概率模型半线性等特点，设计等价常微分方程的离散化解析形式

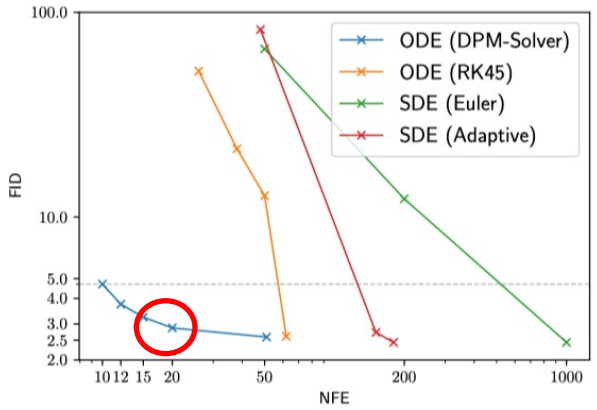
经典龙格库塔法 $$\boldsymbol{x}_t = \boldsymbol{x}_s + \int_s^t \left(f(\tau)\boldsymbol{x}_\tau + \frac{g^2(\tau)}{2\sigma_\tau}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_\tau, \tau) \right) \mathrm{d}\tau$$ 整体黑盒泰勒展开并做差分近似

所提 **DPM-Solver** $$\boldsymbol{x}_{t_{i-1}\to t_i} = \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}\tilde{\boldsymbol{x}}_{t_{i-1}} - \alpha_{t_i}\sum_{n=0}^{k-1}\hat{\boldsymbol{\epsilon}}_\theta^{(n)}(\hat{\boldsymbol{x}}_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}})\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{-\lambda}\frac{(\lambda-\lambda_{t_{i-1}})^n}{n!}\mathrm{d}\lambda + \mathcal{O}(h_i^{k+1})$$

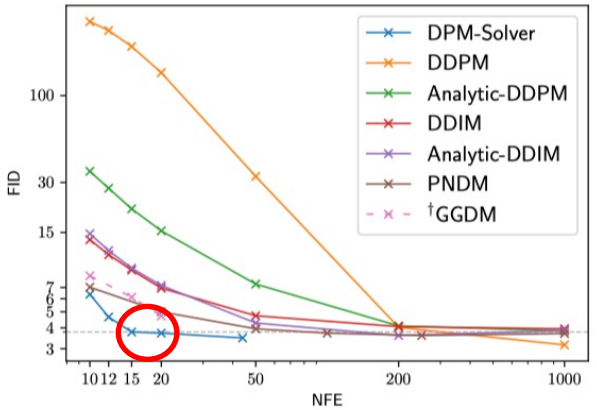
解析形式　　神经网络部分差分近似　　解析形式　　高阶小量

Lu C, et al. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. NeurIPS 2022.

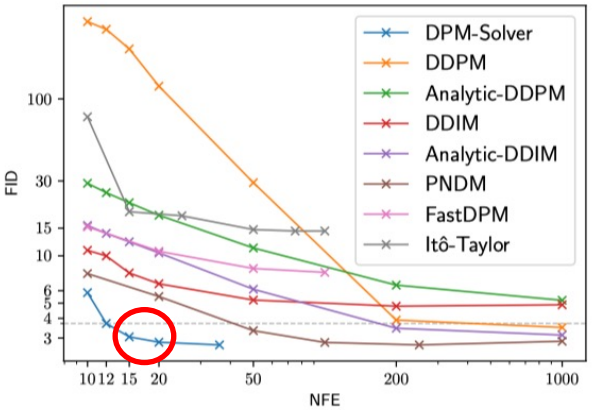
DPM-Solver：结果



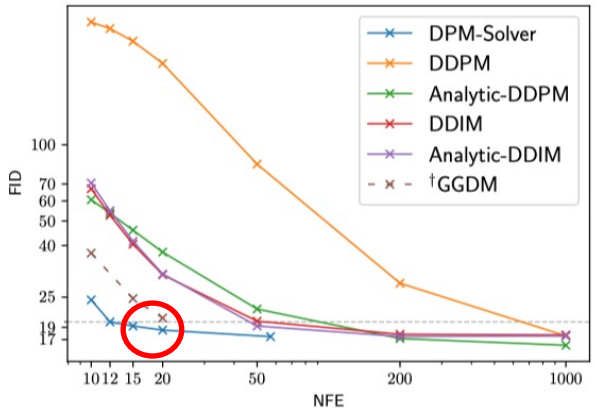
(a) CIFAR-10 (continuous)



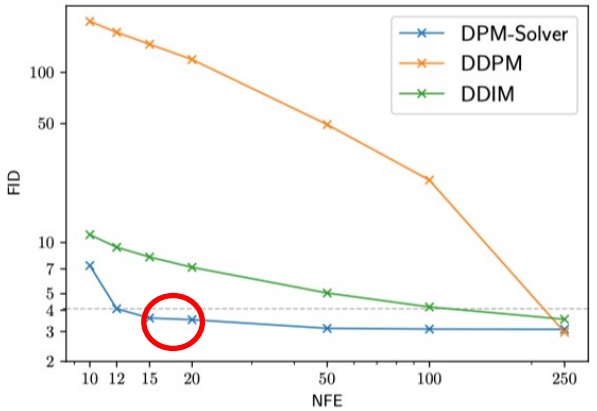
(b) CIFAR-10 (discrete)



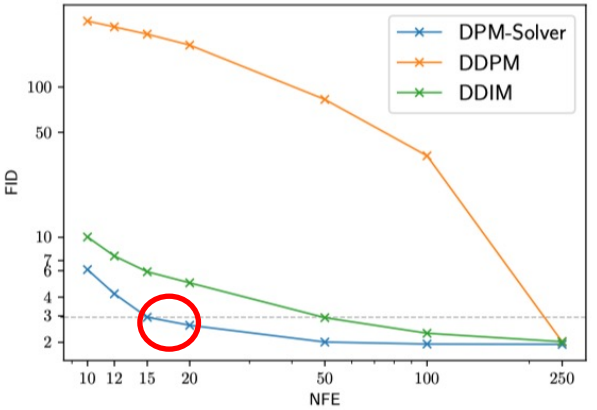
(c) CelebA 64x64 (discrete)



(d) ImageNet 64x64 (discrete)



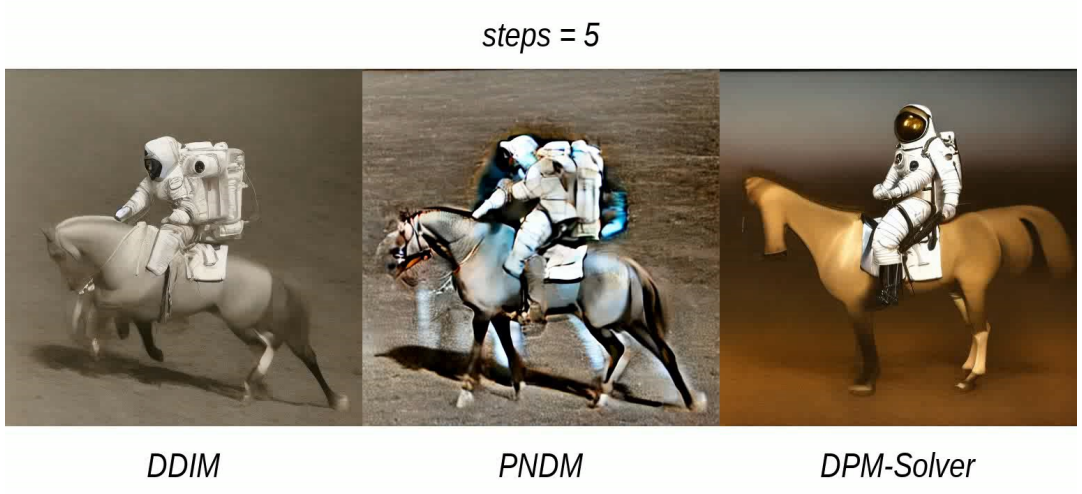
(e) ImageNet 128x128 (discrete)



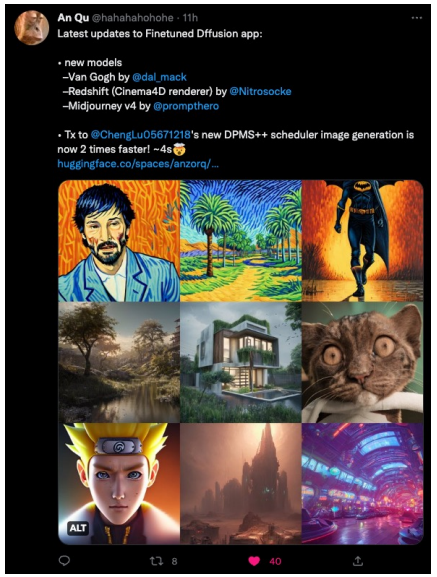
(f) LSUN bedroom 256x256 (discrete)

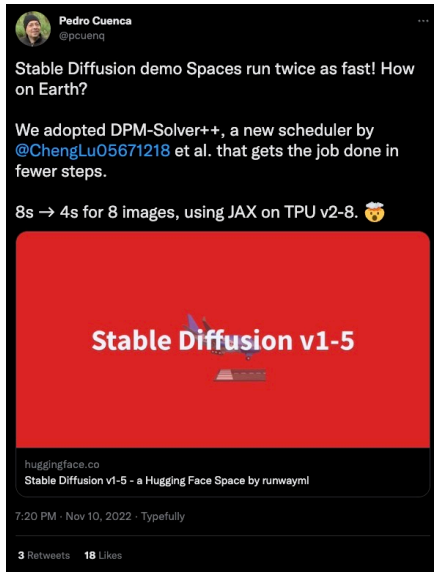
DPM-Solver：结果

- 是当前最快的无需额外学习的扩散概率模型采样算法，**15 步生成高清图像**
- 被多个主流开源社区（**Github 累计星标 6 万余次**）支持/设为默认算法



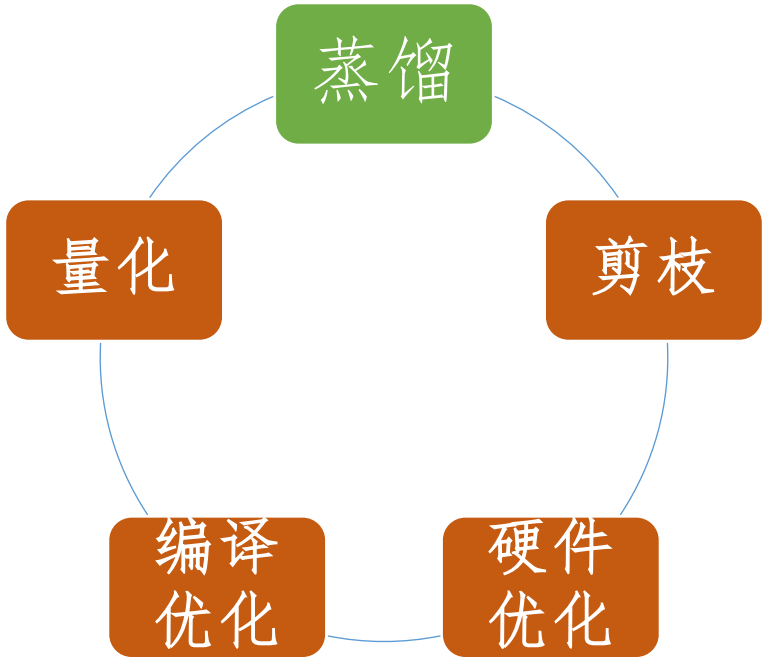
根据文本输入 **15 步生成 512x512 高清图像**



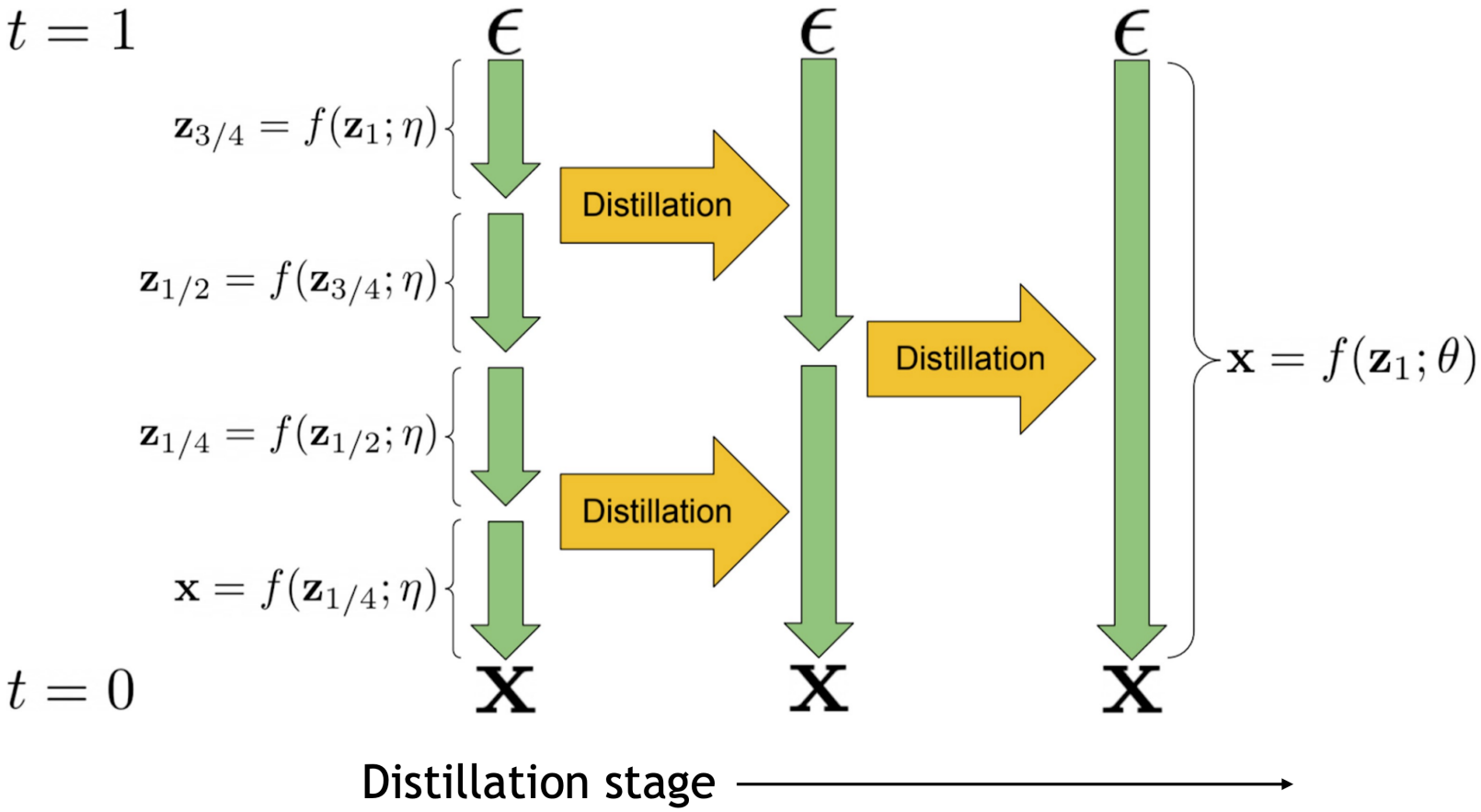


著名开源模型 **Stable Diffusion** 等官方宣传

从模型蒸馏的角度出发实现低步数推理



渐进蒸馏



- 提前确定noise schedule
- 基于预训练模型，以**确定性ODE sampler作为教师**
- **2步蒸馏为1步**：学生只需要教师的一半采样步
- 循环往复

Salimans & Ho. Progressive Distillation for Fast Sampling of Diffusion Models. ICLR 2022

改进：On Distillation of Guided Diffusion Models (Meng et al., CVPR 2023)

- 带来了：
  - CF-Guidance
  - Stochastic sampling
  - Text-to-image
  - Image-to-image
  - Inpainting
  - Latent Diffusion

Text-guided generation (1 step)

Text-guided generation (2 steps)

Class-conditional generation (1 step)

Input Mask Result 1 Result 2

Image inpainting (2 steps)

Text-guided generation (4 steps)

Input Output (different styles)

Image to image translation (3 steps)

问题：蒸馏多个模型，训练成本高

1-4步即可生成高质量大图

一致性蒸馏

- 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

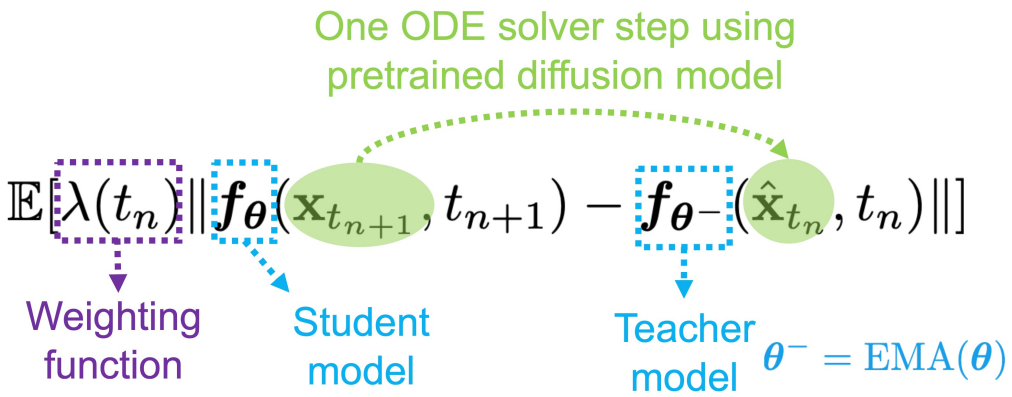
$$\forall t \in [0, T] : \boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \mathbf{x}_0$$

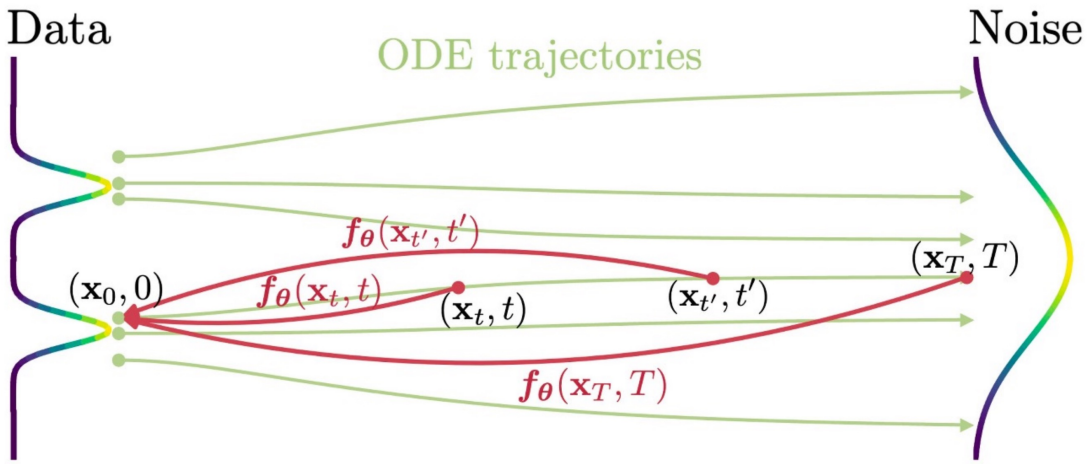
- 模型参数化：需保证边界条件（t=0）

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

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

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

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

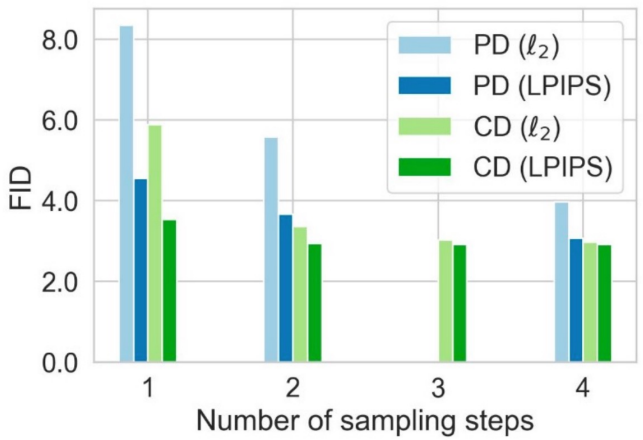
  $\mathbf{x} \leftarrow \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{\tau_n}, \tau_n)$

end for

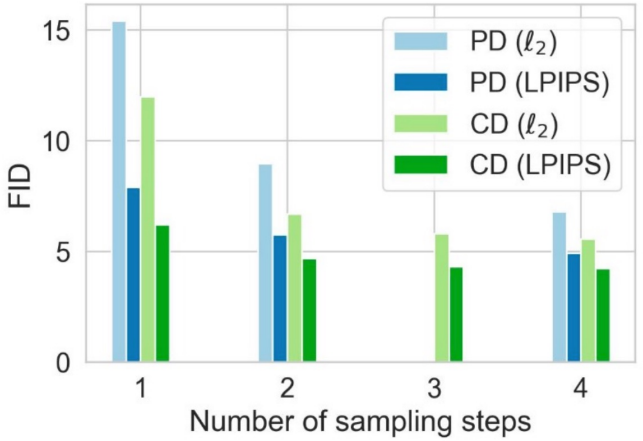
Output: $\mathbf{x}$

Song et al. Consistency Models. ICML 2023

一致性蒸馏：结果



(a) CIFAR-10

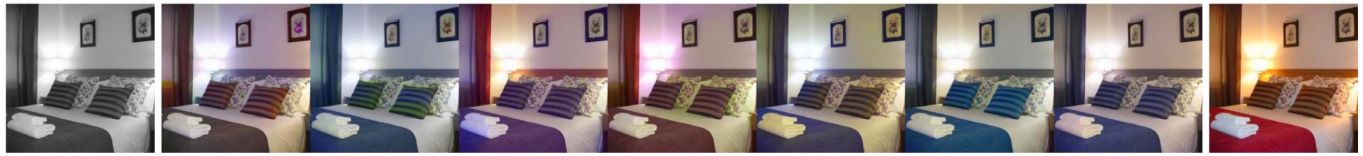


(b) ImageNet 64 × 64

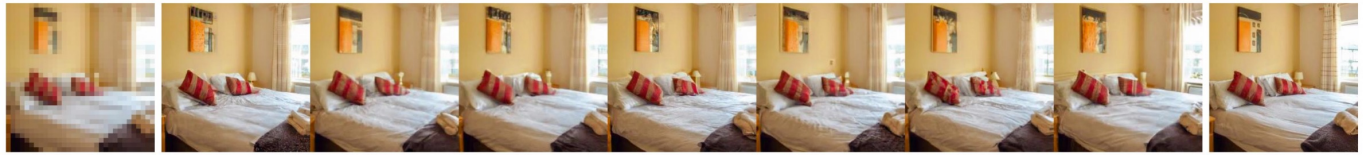
一步生成的**SOTA** FID:

- 3.55 on CIFAR-10
- 6.20 on ImageNet 64

可用于**零样本图像编辑**等应用



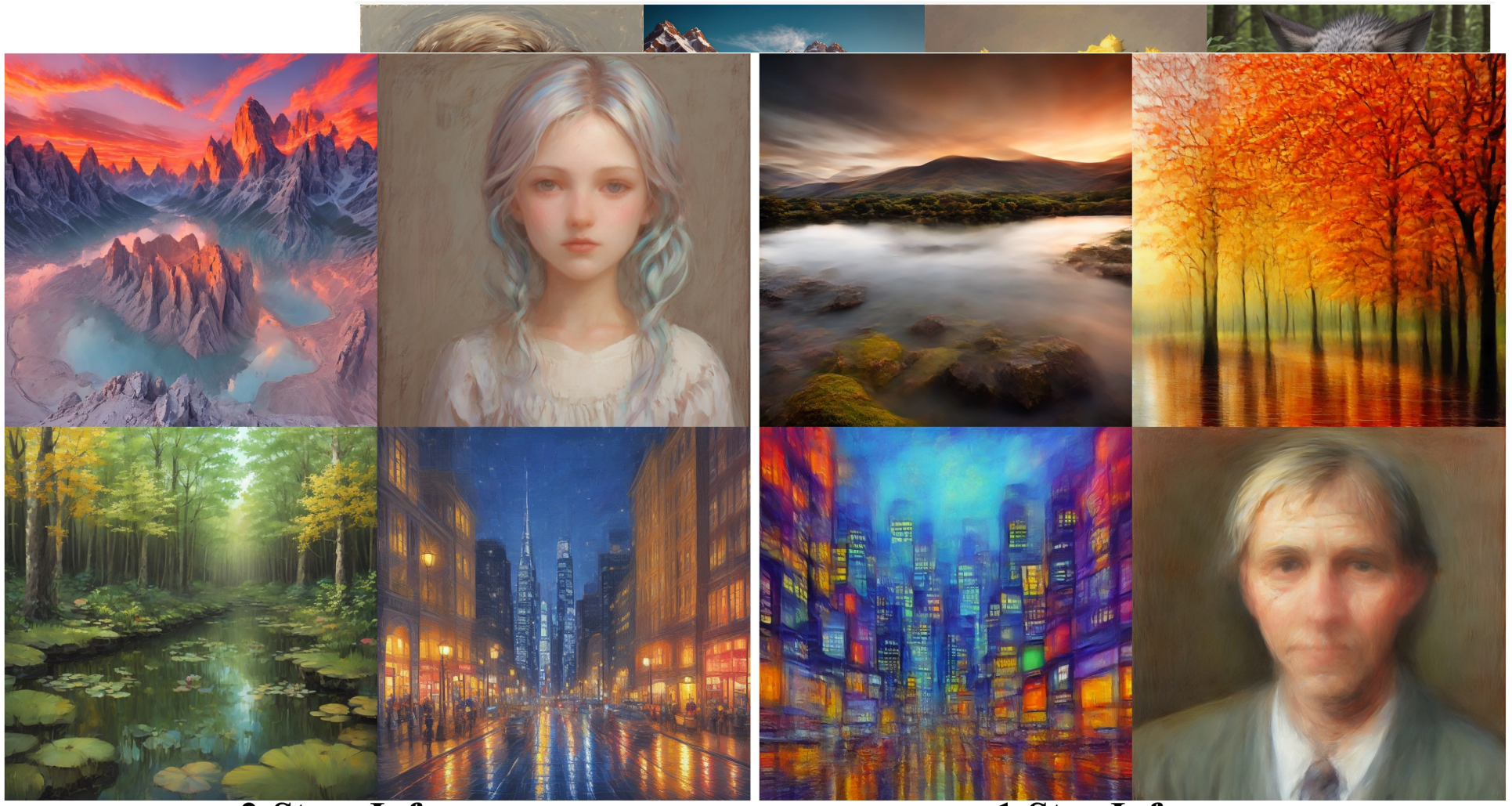
(a) *Left*: The gray-scale image. *Middle*: Colorized images. *Right*: The ground-truth image.



(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.

隐空间一致性蒸馏



从预训练SD蒸馏：

- 4,000个训练步（约32个A100 GPU hour）
- 生成高质量的768×768分辨率图像

Luo et al. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference.

多步隐空间一致性蒸馏

- 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

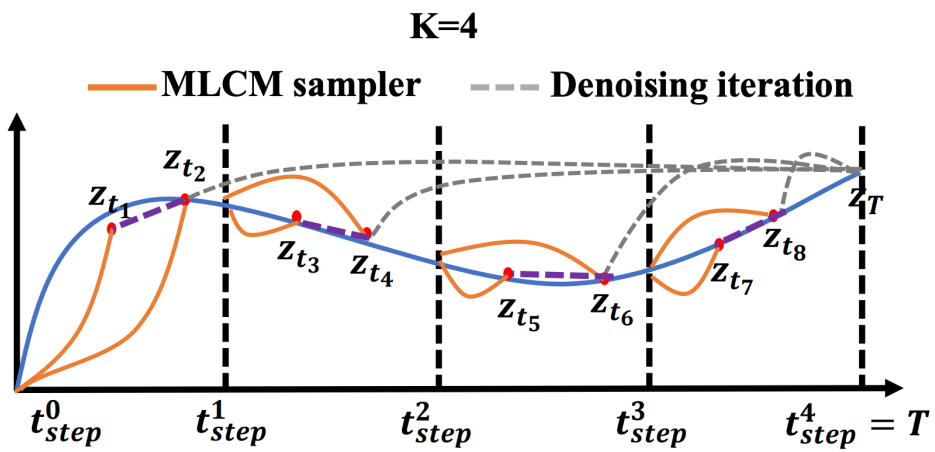


Image-free MLCD in the first stage

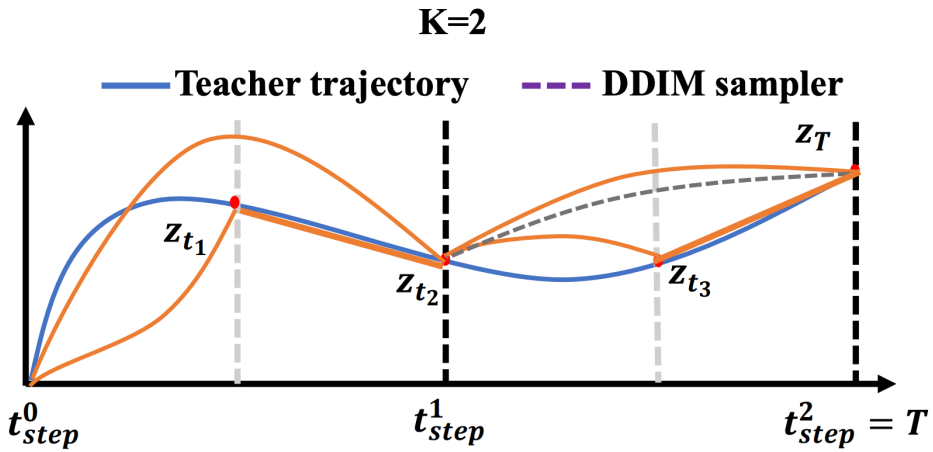


Image-free MLCD in the progressive training stage

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

多步隐空间一致性蒸馏：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1024^2的图像生成，远超SDXL, HyperSD等开源模型
- **赋能OPPO产品**

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

多步隐空间一致性蒸馏：结果

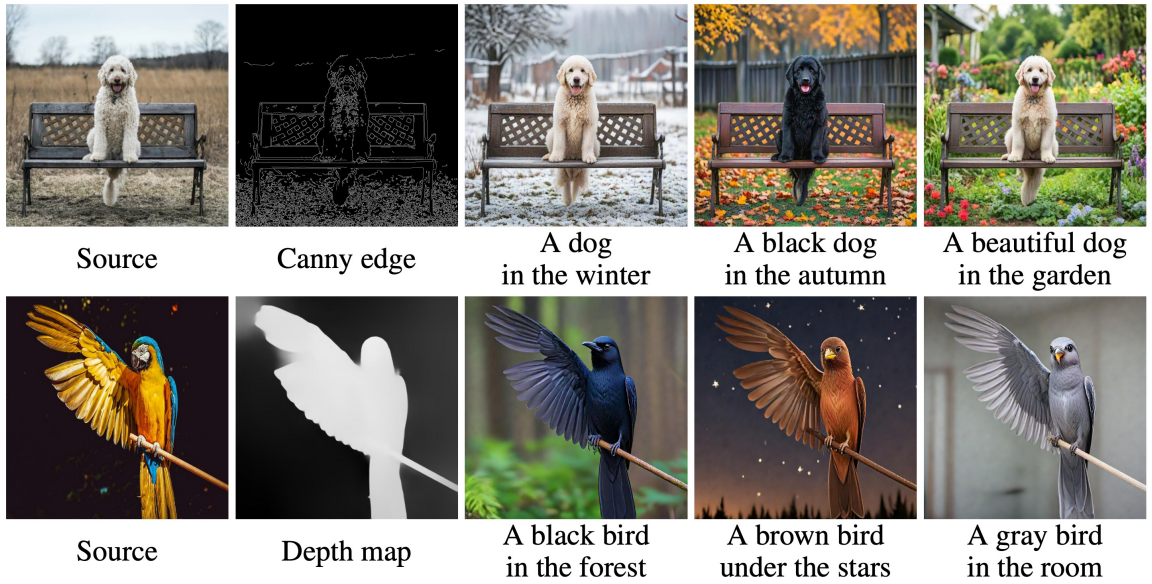


Figure 4: MLCM with ControlNet. Our MLCM can be incorporated into ControlNet pipeline and produce satisfactory results with 2 steps sampling.

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.

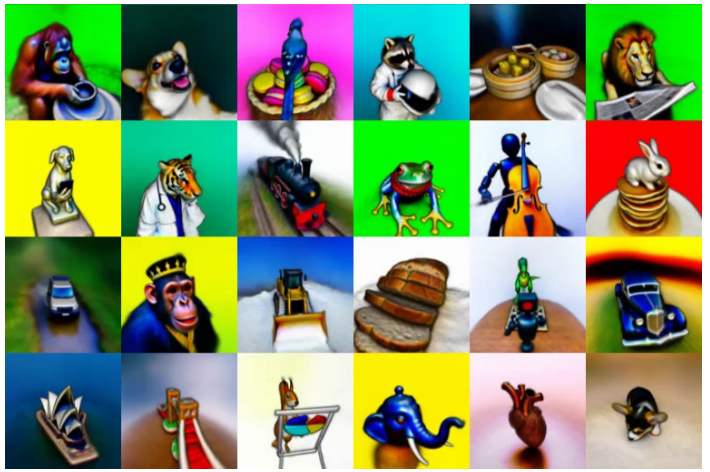
赋能中文-图像生成：3步采样结果

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

得分蒸馏（score distillation）和变分得分蒸馏（variational score distillation，VSD）

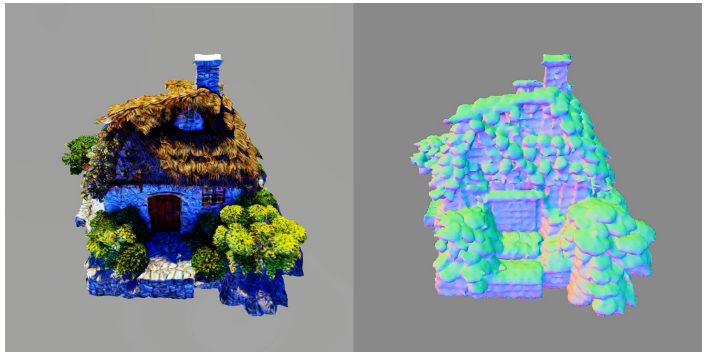
- 得分蒸馏：基础扩散模型定义蒸馏的先验分布

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) \triangleq \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)(\boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon})\frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta}\right]$$



- 变分得分蒸馏基于粒子的变分推断，缓解过饱和问题

$$\frac{\mathrm{d}\theta_\tau}{\mathrm{d}\tau} = -\mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)\Big(\underbrace{-\sigma_t \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|y^c)}_{\text{score of noisy real images}} - \underbrace{(-\sigma_t \nabla_{\boldsymbol{x}_t} \log q_t^{\mu_\tau}(\boldsymbol{x}_t|c,y))}_{\text{score of noisy rendered images}}\Big)\frac{\partial \boldsymbol{g}(\theta_\tau, c)}{\partial \theta_\tau}\right]$$



已知基础扩散模型分布　未知渲染图像分布，采用 **LoRA** 估计

问题：收敛慢，需要多阶段（**NeRF generation, geometry refinement, and texture refinement**）训练

Po

Wang et al. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. NeurIPS 2023.

变分得分蒸馏的实际改进：linearization+lookahead （L^2）

- 根据当前**3D**状态，**LoRA**模型向前看**1**步，保证对渲染数据分布的拟合

$$\nabla_{\theta_i}\mathcal{L}_{L-VSD}(\theta_i) = \mathbb{E}_{t,\epsilon,c}\left[w(t)(\epsilon_{pretrain}(x_t,t,y) - \epsilon_{\phi_{i+1}}(x_t,t,c,y))\frac{\partial g(\theta,c)}{\partial\theta}\Big|_{\theta=\theta_i}\right]$$

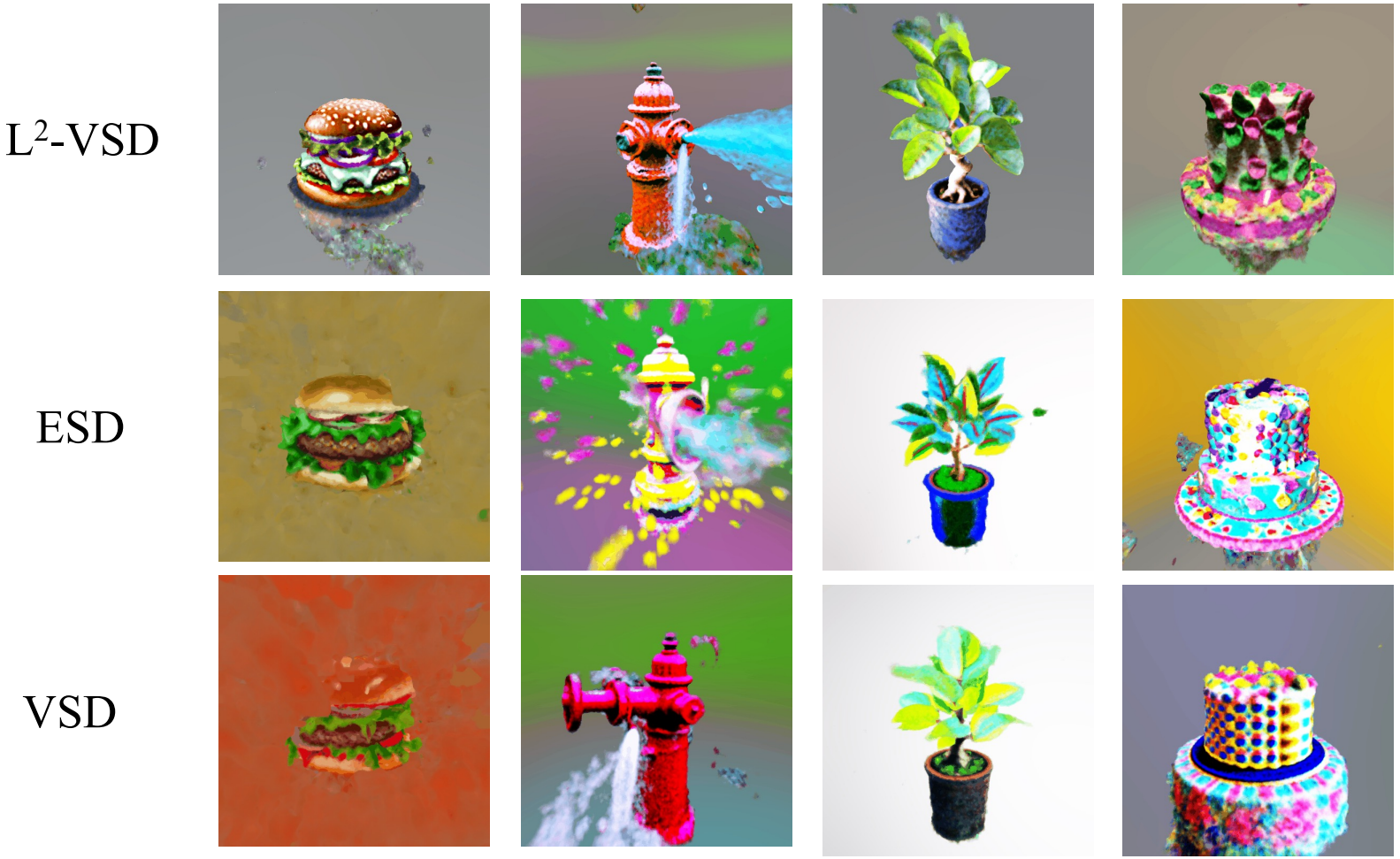
- 线性化**LoRA**模型防止过拟合

$$\epsilon_\phi^{\text{lin}}(x_t,t,c,y) = \epsilon_{\phi_i}(x_t,t,c,y) + (\phi-\phi_i)J_{\phi_i}^T(x_t,t,c,y)$$

- 基于**forward-mode autodiff**实现**VJP**

Lei, Liu, Xie & Deng. Advancing Text-to-3D Generation with Linearized Lookahead Variational Score Distillation.

L^2-VSD：结果



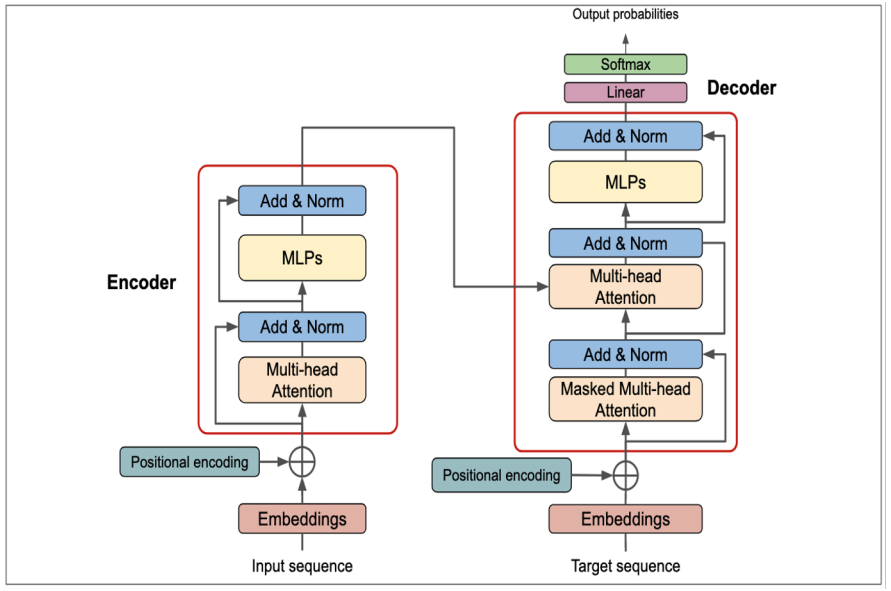
一个训练阶段即可
生成高质量的结果

Lei, Liu, Xie & Deng. Advancing Text-to-3D Generation with Linearized Lookahead Variational Score Distillation.

AIGC大模型的高效推理方法

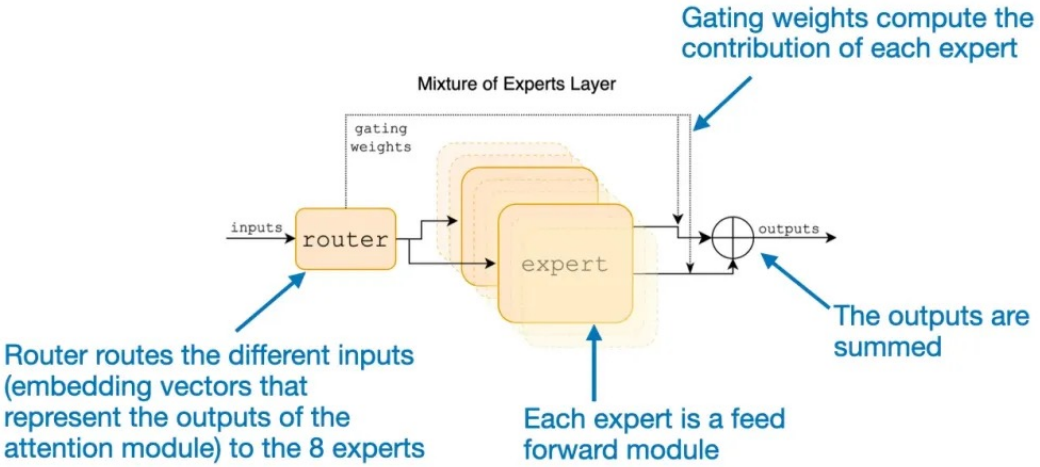
- 大语言模型的并行推理方法（15min）
  - 投机解码
  - 一致性蒸馏
- 大扩散模型的低步数推理方法（15min）
  - 采样器设计
  - 一致性蒸馏
- 大模型架构、序列状态、缓存等方面的优化方法（5min）

模型架构上的优化

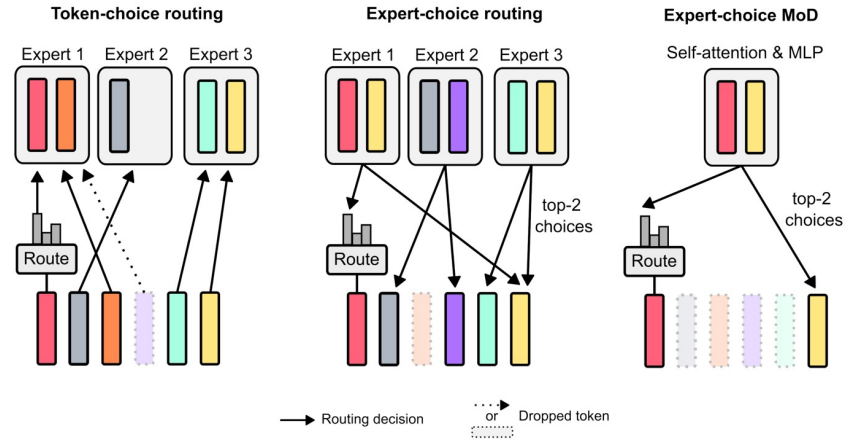


架构趋于统一：**Transformer**

问题：如何进一步高效提升模型表达能力？



专家混合（**MoE**）：稀疏激活专家，保证**model capacity**，同时降低计算开销。代表：**Mixtral-8*7B**



深度混合（**MoD**）：从**token routing**到**expert routing**，保障负载平衡，但难进行**casual modeling**

Vaswani, Ashish, et al. Attention is all you need. NeurIPS 2017.
Jiang et al. Mixtral of Experts. 2024
Raposo et al. Mixture-of-Depths: Dynamically allocating compute in transformer-based language models. 2024.

模型架构上的优化

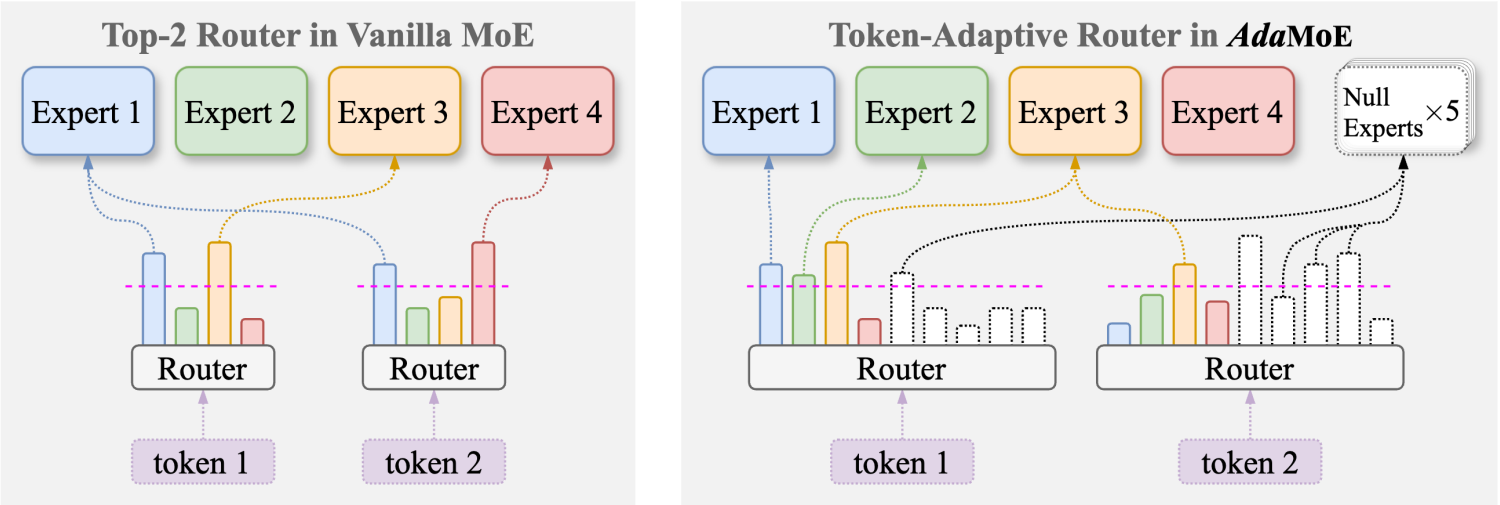


Figure 2: Comparison of Routing Mechanisms: Vanilla MoE vs. *Ada*MoE. **Left:** In the Vanilla MoE model, each token selects the top-2 experts based on the highest router probabilities. **Right:** The *Ada*MoE model introduces an additional set of *null experts*. Each token selects the top-4 experts, which can include both true and null experts. For example, token 1 selects three true experts, while token 2 selects only one true expert. Despite this variation, the average number of true experts selected per token remains two, maintaining parity with the vanilla method.

AdaMoE：引入空专家，增加灵活度，不同token可以选0-4个真专家，负载均衡损失保证空专家使用率，建模灵活性优于MoD

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

模型架构上的优化

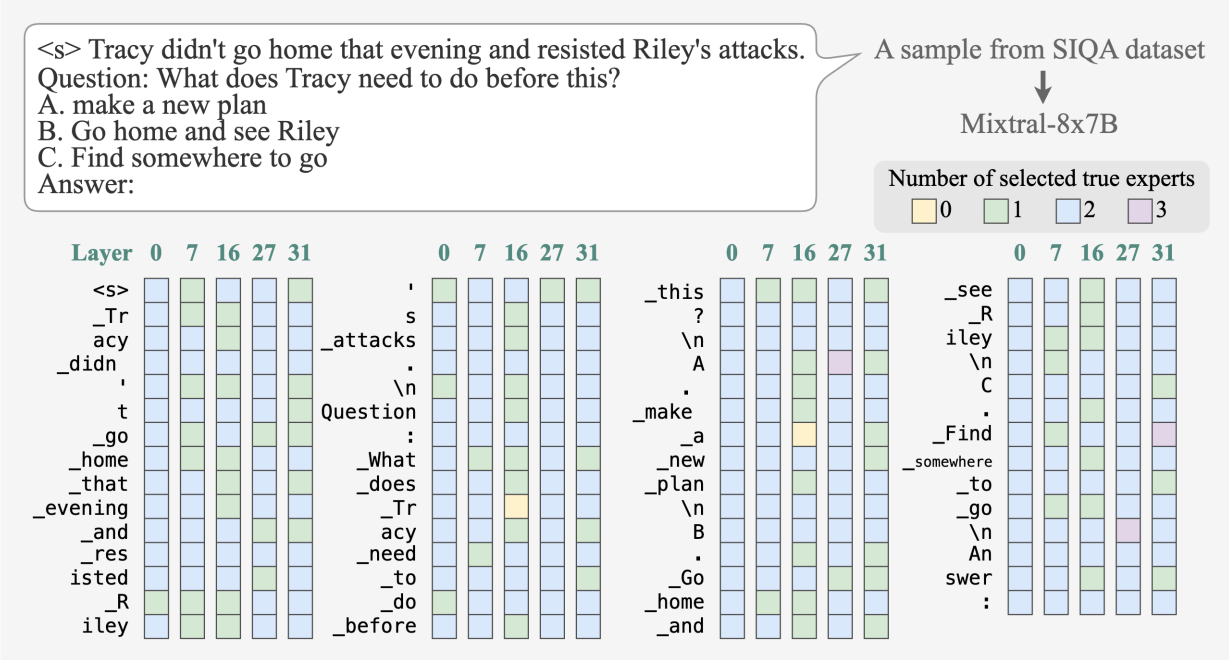
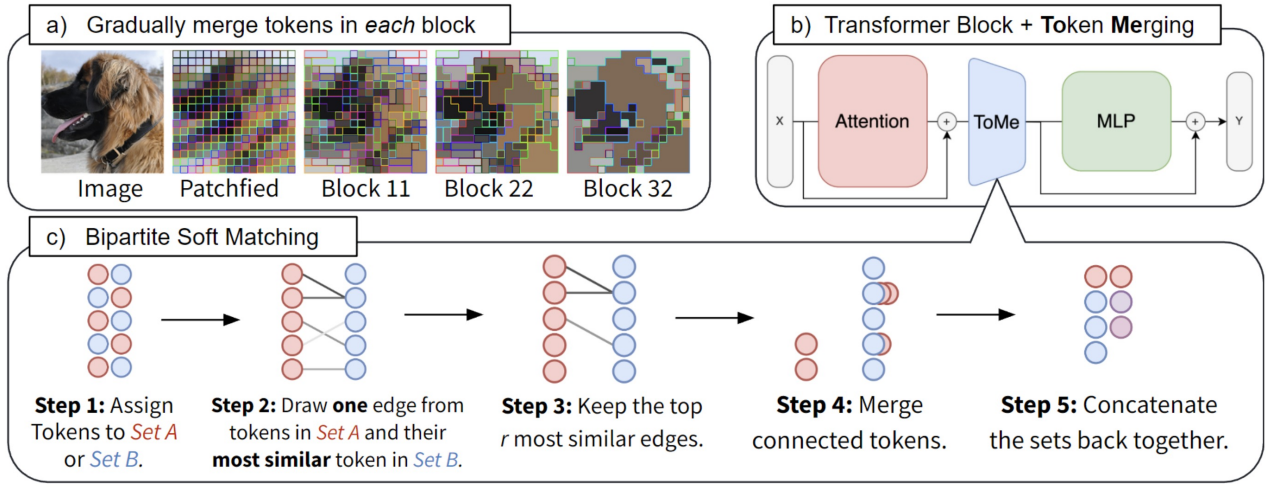


Figure 1: Inference on a sample from SIQA dataset after fine-tuning Mixtral-8x7b with *Ada*MoE. The image shows the number of true experts selected for each token in different layers.

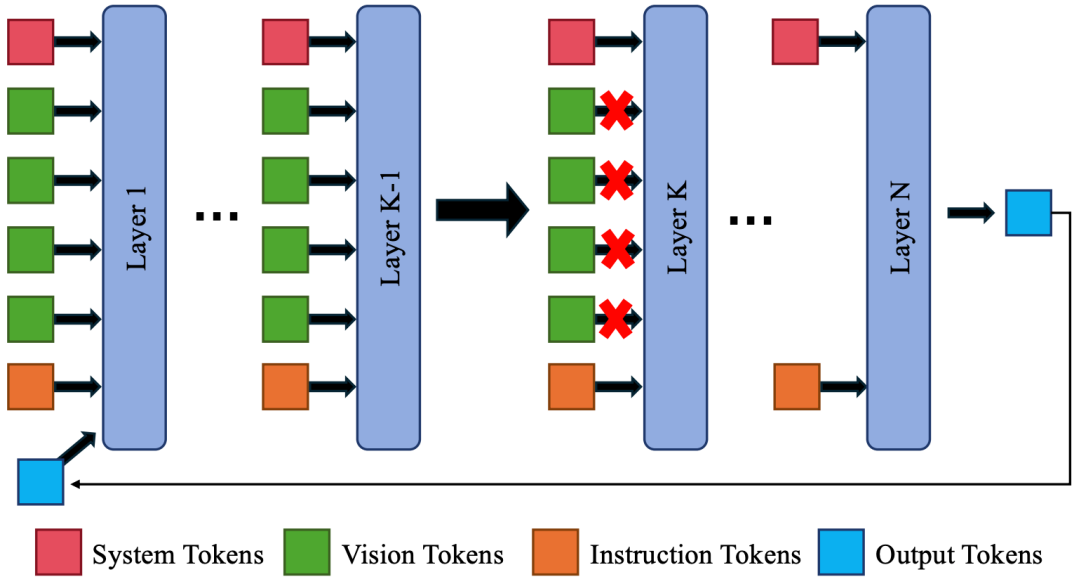
对Mixtral-8*7B简单微调，即可收获大量选择1个真专家的token，性能不减的情况下降低至多20%的FLOPs

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

Token序列上的优化



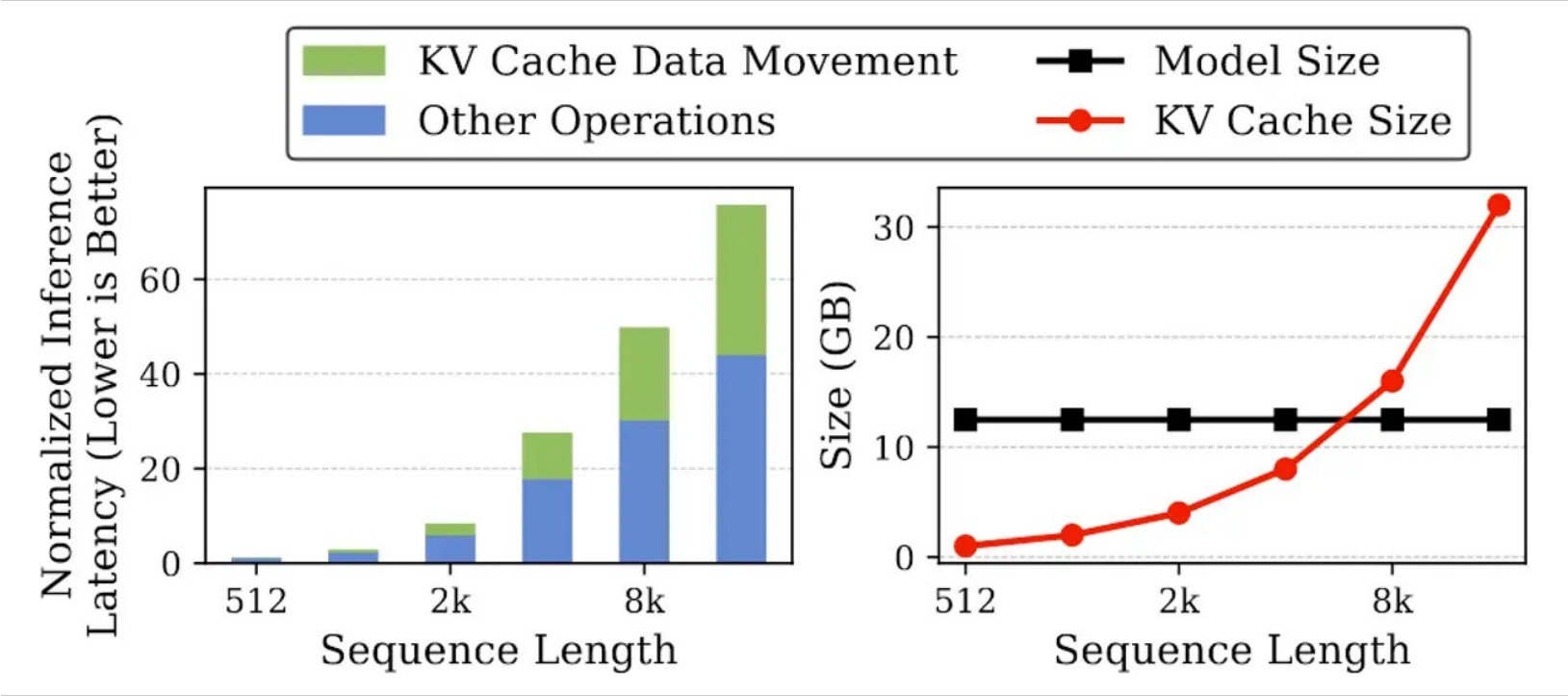
基于相似度对**token**序列进行合并，缩减序列长度，降低计算开销，可用于图像和视频生成



Multi-modal大模型中，中间视觉**tokens**在深层中获得的注意力极少，其信息转移到文本**tokens**中，因此若干层（如：**16**）之后，可直接丢弃视觉**tokens**

Bolya et al. TOKEN MERGING: YOUR VIT BUT FASTER. ICLR 2023.
Lin et al. BOOSTING MULTIMODAL LARGE LANGUAGE MODELS WITH VISUAL TOKENS WITHDRAWAL FOR RAPID INFERENCE. 2024.

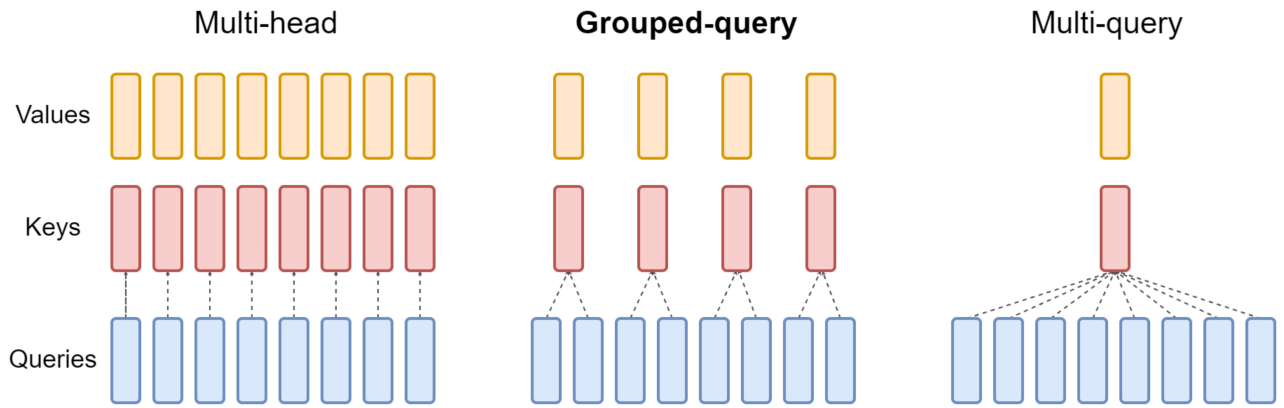
缓存（KV Cache）上的优化



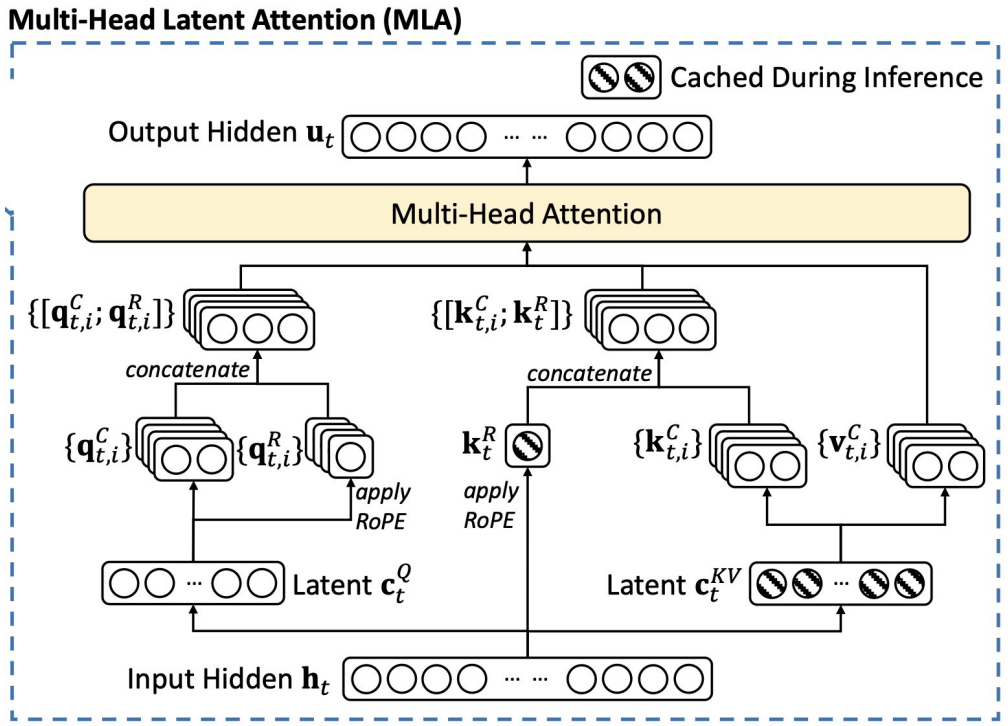
长序列场景下，访存（KV Cache）成为瓶颈（MPT-7B ）

Adnan et al. Keyformer: KV Cache Reduction through Key Tokens Selection for Efficient Generative Inference. MLSys 2024

缓存（KV Cache）上的优化



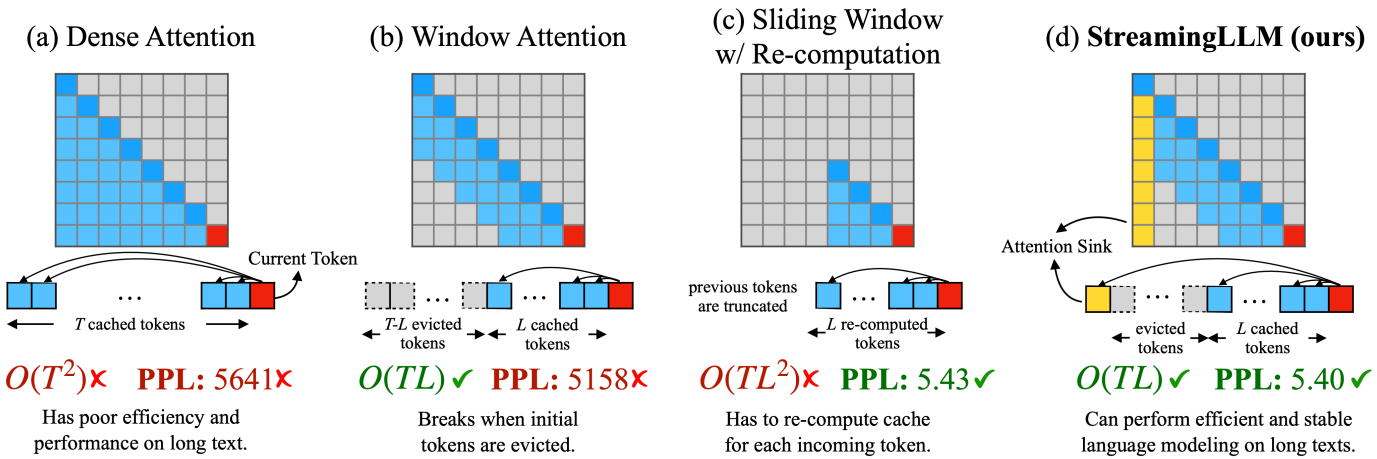
Grouped-Query Attention：将查询向量分组处理，共享KV



Multi-head Latent Attention：将KV cache投影到低维隐向量，用于DeepSeek v2

Ainslie et al.. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. EMNLP 2023.
DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. 2024.

缓存（KV Cache）上的优化



StreamingLLM： window attention+起始位置attention，弥补的window attention精度损失

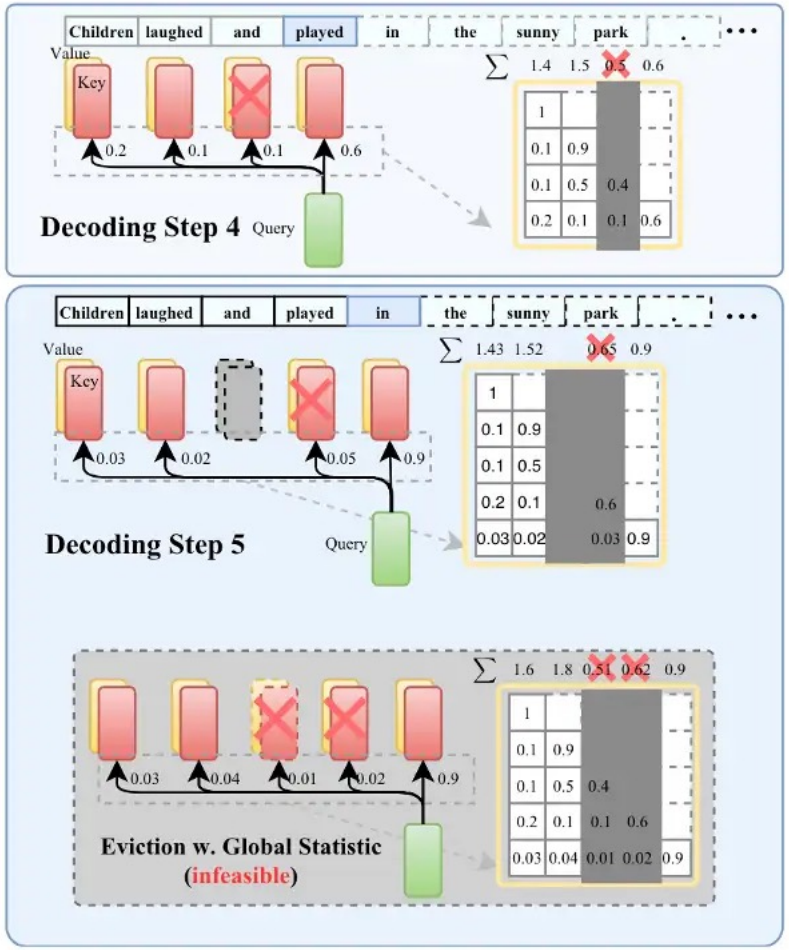


Figure 3: Illustration of Algorithm 1 during two consecutive decoding steps.

H2O：贪心地丢掉当前最不被注意的token

Xiao et al. EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS. ICLR 2024.
Zhang et al. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. NeurIPS 2023.

AIGC大模型的高效推理方法：未来研究方向

- 基础：
  - 模型架构、学习方法、采样算法

- 场景：
  - 多模态大模型
  - 视频生成模型（世界模型）

感谢各位专家！敬请批评指正！

邮箱： zhijied@sjtu.edu.cn

主页： https://thudzj.github.io/