清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

趋势解读：推理优化

Xuefei Ning

Department of Electronic Engineering
Tsinghua University

2025.01

foxdoraame@gmail.com

Disclaimer: The representative papers selected in this slide are not comprehensive and omit many influential works. These selected papers are intended to illustrate my perspective on the trend.

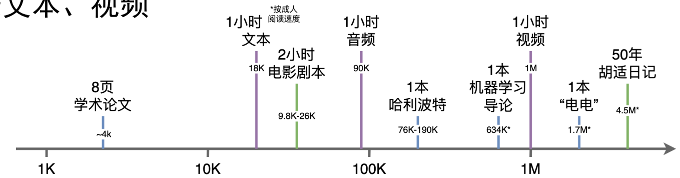
NICS EFC

应用需求与负载趋势

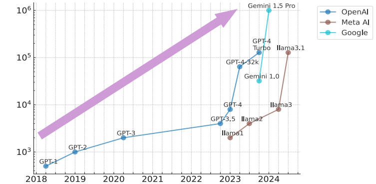
应用需求

负载特点

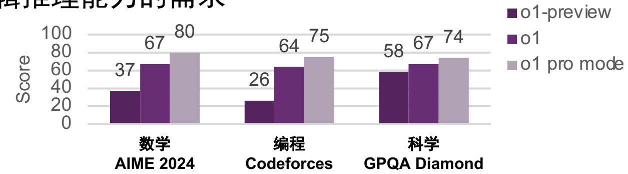
长文本、视频



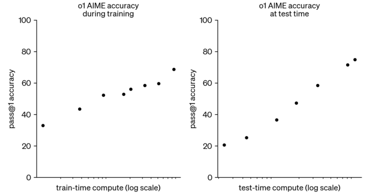
→ Scaling 输入序列长度



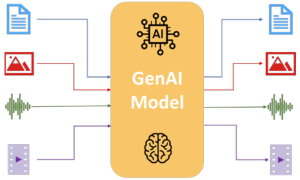
逻辑推理能力的需求



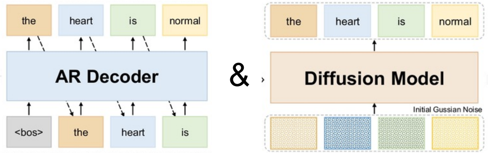
→ Scaling 推理时计算



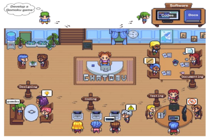
混合模态的理解和生成



→ 生成范式的融合或统一



Agentic流程



[1] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

[2] Reid, Machel, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." arXiv preprint arXiv:2403.05530 (2024).

[3] Dubey, Abhimanyu, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).

[4] OpenAI. (n.d.). Learning to reason with LLMs. Retrieved January 2, 2025, from https://openai.com/index/learning-to-reason-with-llms/

[5] OpenAI. (n.d.). Introducing ChatGPT Pro. Retrieved January 2, 2025, from https://openai.com/index/introducing-chatgpt-pro/

应用需求与资源限制

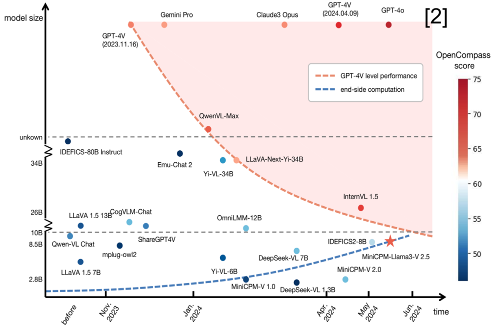
Generative AI
+
个人使用场景

On-Device | Server-Based | GPT-4o

Task Complexity

3B params | Unpublished

端侧应用



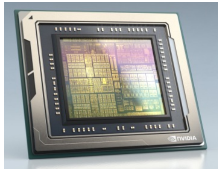
端侧设备
8GB 运行内存

支持部署
~3B Model

自动驾驶
L4/L5
>1000 TFLOPS[4]
[Jan Patnzar, VSORA]

无人机
路径规划
>100 frame/s[5]
[Nature 2023封面文章]

具身智能



Google DeepMind

NVIDIA DRIVE Orin
峰值算力

254 INT8 TOPS

RT-2-X-5B model [3]
1k prompt tokens

~ 12.7 token/s

[1] Gunter, Tom, et al. "Apple intelligence foundation language models." arXiv preprint arXiv:2407.21075 (2024).
[2] Yao, Yuan, et al. "Minicpm-v: A gpt-4v level mllm on your phone." arXiv preprint arXiv:2408.01800 (2024).
[3] O'Neill, Abby, et al. "Open x-embodiment: Robotic learning datasets and rt-x models." arXiv preprint arXiv:2310.08864 (2023).
[4] Jan Patnzar. "The Challenges to Achieve Level 4/Level 5 Autonomous Driving." from https://www.gsaglobal.org/forums/the-challenges-to-achieve-level-4-level-5-autonomous-driving/
[5] Kaufmann, E, et al. Champion-level drone racing using deep reinforcement learning. Nature 620, 982–987 (2023).

技术回顾：语言生成模型

算法层

算法优化

Speculative Decoding

Non-auto-regressive

Eagle PKU & Microsoft & UW… 2024.01 arxiv; ICML'24	**Eagle2** PKU & Microsoft & UW… 2024.01 arxiv; ICML'24	探索discrete diffusion做语言生成	采用 Jacobi 解码，同时生成多个 token
草稿模型针对最后特征而非输出token进行回归预测	根据草稿模型置信度动态调整草稿树的结构	***SEDD*** *Stanford & Pika Labs* *2023.10 arxiv; ICML'24*	**CLLMs** SJTU & UCSD 2024.02 arxiv; ICML'24

模型层-模型压缩

量化

PTQ

QAT

AWQ *MIT & SJTU & NVIDIA…* *2023.06 arxiv; MLSys'24*	**Quarot** ETH Zurich & EPFL… 2024.04 arxiv; NeurIPS'24	**SpinQuant** Meta 2024.05 arxiv	(QAT) W2 g128 -2.61% on LLaMA-2-7B	(QAT) W2 g128 -1.72% on LLaMA-2-7B
(PTQ) W4 g128 -0.13% on LLaMA-2-7B	(PTQ) W4 A4 KV4 -0.46% on LLaMA-2-7B	(PTQ) W4 A4 KV4 -0.40% on LLaMA-2-7B	**BitDistiller** HKUST & SJTU & MSRA 2024.02 arxiv	**EfficientQAT** HKU & Shanghai AI Lab 2024.07 arxiv

稀疏化

KV-Cache

Attention

StreamingLLM *MIT & Meta & CMU…* *2023.09 arxiv; ICLR'24*	**Quest** SJTU & MIT & UW… 2024.06 arxiv; ICML'24	**MoA** Tsinghua & Infinigence… 2024.06 arxiv	混合多种稀疏注意力模式，加速模型的 Prefill
对 LLM 静态 KV-Cache 稀疏的早期探索。实现流畅流式长文对话	根据输入的 query token，动态的取回需要使用的 KV-Cache	混合多种稀疏注意力模式和长度扩展模式，加速模型的 Decode	**MInference 1.0** Microsoft & Surrey 2024.07 arxiv; NeurIPS'24

技术回顾：语言生成模型

模型层-结构设计：小参数量模型

语言模型	语言模型	语言模型	图文多模态模型	图文音多模态模型
PanGu-Π-1B~7B *Huawei* *2023.12*	**MiniCPM-2B** OpenBMB 2024.02	**Llama3.2-1B/3B** Meta 2024.09	**MiniCPM-V-2.6-8B** OpenBMB 2024.08	**Megrez-3B-Omni** Infinigence 2024.12
配置合适的宽度/深度的宏观架构参数；FFN多分支非线性设计；进行多轮训练+数据精炼；简化词表	在公开评测集上与 Mistral-7B 表现相近，整体性能超越 Llama2-13B、MPT-30B、Falcon-40B	利用 Llama3.1 系列，8B 剪枝、从 8B 和 70B 蒸馏、从 405B 模型收集合成数据训练小模型	支持单图、多图和视频理解，官方宣传其取得了优于 GPT-4V 的表现	同时具备图片、文本、音频三种模态数据的理解分析能力

模型层-结构设计：低复杂度结构

Mamba *CMU & Princeton* *2023.12 arxiv; CoLM'25*	**Mamba-2** Princeton & CMU 2024.05 arxiv; ICML'24	**Jamba** AI21 Labs 2024.03 arxiv	**TTT** Stanford & UCSD & UCB… 2024.07 arxiv
提出 State Space Model，解决 attention 计算时随着输入长度平方增长的的复杂度	揭示了 Mamba 和传统 Transformer 之间的相关性，同时设计了新的 Mamba 架构，提供更高的加速比	首个混合SSM和transformer的工作，成功将混合模型scale-up至52B，显著提升在长文本任务上的推理效率 其他混合模型工作: **Zamba, Hymba, …**	修改 RNN layer，并且提出将隐藏状态变成模型，提出新的线性复杂度模型层：TTT

技术趋势：语言生成模型

算法层

在所有针对AR模型、利用“并行”这一思想提高计算利用率方法中，**Speculative Decoding**方法已有长足进展，被广泛实现入各大框架，在优化小batch场景的latency上非常有效

Jacobi decoding、**Agentic generation** (e.g., **Skeleton-of-Thoughts**) 等方法由于加速比相对受限或与应用场景相关，研究和应用相对少

使用Diffusion建模语言已有不少探索，很多工作围绕discrete token space handling，一些工作也探讨了token sequence handling，但尚未充分验证scalability

AR与Diffusion/Flow Matching的结合或为重要方向

模型层

针对大语言模型的**Training-free模型压缩**研究已相当充分；针对多模态理解大模型的Training-free模型压缩在这半年出现

针对大语言模型的**Training-based模型压缩**研究(e.g., QAT)已有长足进展。将各个维度配置好、模型训好有望在工程上继续推进“能力密度”提升，但是否有数量级提升需要考虑

设计少参数小模型、低复杂度结构为一关注重点
- 少参数小模型的“能力密度”持续提升
- 小复杂结构的scalability验证为难点；混合结构能取得有不错Trade-off

技术回顾：视觉生成模型

算法与应用

应用算法

- **Sora** — OpenAI — 2024.02
  商业模型，第一个Transfomer-Based的大规模视频生成模型，视频时长首次达到分钟级，分辨率达1k，生成内容具有一定程度物理特性
- **Open-Sora** — HPC-AI — 2024.03
  首个开源的类Sora模型，GitHub Star 达22.9k；基于3D VAE与Flow Matching训练；可生成15s, 720p视频
- **可灵** — 快手 — 2024.06
  商业模型，中文能力突出，长度达2min，分辨率达1k；能够生成大幅度的合理运动；能够模拟真实物理世界的特性
- **HunyuanVideo** — 腾讯 — 2024.12
  开源模型，指标优于闭源的Gen-3 (Runway)；参数量达13B，基于3D VAE，和Flow matching；可生成5s 720p的视频

生成算法探索

AR

Diffusion/ Flow Matching

- **SD3** — Stability AI — 2024.03 arxiv; ICML'24
  基于Flow Matching的大规模少步文生图模型
- **VAR** — PKU & ByteDance — 2024.04 arxiv; NeurIPS'24
  离散token space，提出Next Scale Prediction
- **LlamaGen** — HKU & ByteDance — 2024.06 arxiv
  离散token space，更大的codebook size (14bit)，基于Llama架构
- **MAR** — MIT & DeepMind & THU — 2024.06 arxiv
  首次在连续的token space做Masked AR生成，用Diffusion建模连续token
- **Transfusion** — Meta & Waymo & USC — 2024.08 arxiv
  AR+Diffusion，语言部分用AR生成，视觉部分用Diffusion生成

效率优化

模型层

算法层

- ***DMD*** — *MIT & Adobe* — *2023.11 arxiv; CVPR'24*
  少步模型蒸馏；匹配单步生成器生成数据的分布与教师Diffusion模型建模的数据分布
- **ViDiT-Q** — THU & Infinigence & MSR — 2024.05 arxiv
  PTQ；Token-wise量化，在不同时间步上动态地做通道均衡；2.5x 显存优化，1.7x端到端加速
- **DiTFastAttn** — THU & Infinigence & SJTU — 2024.06 arxiv; NeurIPS'24
  高效attention；Window & reused attention for DiT；1.6x端到端加速
- **SageAttention** — THU — 2024.10 arxiv
  PTQ；对K做smoothing，然后对Q/K做int8量化，在1.3x端到端加速
- **SANA** — NVIDIA & MIT & THU — 2024.10 arxiv
  高效架构设计，1024x压缩率VAE+线性Attention；SANA 0.6B效果与FLUX-Dev 12B相当
- **Distilled Decoding** — THU & MSR — 2024.12 arxiv
  AR+Flow Matching；首次将Pretrained AR模型压缩至1步

技术趋势：视觉生成模型

算法与应用

- **Flow Matching**作为有着更简洁和通用的理论、可兼容Diffusion的生成模型算法，逐渐成为主流，可帮助实现更少步数的高质量生成
- 针对统一多模态这一目标，**AR模型进行视觉生成**再次受到大量关注，生成质量开始与Diffusion/Flow Matching可比
- **AR与Diffusion/Flow Matching的结合**成为目前探索“统一多模态生成”的重要方向

效率优化

- 针对Diffusion/Flow Matching模型的算法层**Training-free时间步压缩**在2023年基本就已达到上限；**Training-based时间步压缩**仍在继续研究
- 出现大量针对Diffusion/Flow Matching**模型层效率优化**工作。随着长视频生成应用的火爆和模型的出现，**Attention优化**或将成为热点
- 开始探索**针对AR视觉生成模型的效率优化**，尤其是压缩AR生成的步数

技术回顾：云侧系统优化

特点

优化目标：
更注重 throughput
（latency限制下的
throughput优化）

特点：
核心是软件
软件上重 serving 系统

<2024 算子优化

FlashAttention	FlashDecoding	FlashDecoding++
Stanford & UBuffalo	Stanford	Tsinghua & SJTU & Infinigence
2022.05 arXiv; NeurIPS'22	2023.10	2023.11 arXiv; MLSys'24
从访存角度 优化 prefill 时的 attention 计算方式	从提高并行度的角度 优化 decode 时的延迟	通过优化算子实现细节， 进一步提高 decode 效率

~2x throughput

2022~2024 Serving系统优化

ORCA	VLLM	SGLang	DistServe	?
Seoul National University	UCB & Stanford & UCSD	Stanford & UCB & SJTU…	PKU & StepFun & UCSD	Infinigence
2022.06 OSDI	2023.09 arXiv; SOSP'23	2023.12 arXiv	2024.01 arXiv	2025
continuous batching	paged attention	CPU & GPU overlap	Disaggregated Prefill & Decoding	SLO-aware scheduling
~5x throughput	~4x throughput	~1.5x throughput	~1.5x input request rate	~1.5x input request rate

近2个数量级的throughput提升。针对现有模型结构，不考虑
具体下游应用推理流程，serving系统的提升空间基本被榨干

技术回顾：端侧系统优化

特点

优化目标：
更注重 latency
严格的peak memory & energy budget等资源限制

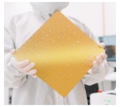
特点：
核心是硬件
软件上重部署工具链（深&多样的工具链栈）

芯片发展

2020

英伟达：A100
FP16（Tensor Core）：312 TFLOPS

cerebras
晶圆级芯片
400,000个计算单元



2022

AMD：MI100
FP 32：95.7 TFLOPS

英伟达H100
FP32：60 TFLOPS
FP16（Tensor Core）：1,000 TFLOPS

2023

2023年谷歌：TPU v5
393 TOPS（BF16/INT8）

cerebras
晶圆级芯片 WSE-2
850,000个AI优化核心。

AMD MI300

2024

英伟达B200/B100
FP16（Tensor Core）：2250 TFLOPS

端侧芯片投入持续加大

花费**40亿美元**研发AI推理芯片
支持特斯拉FSD自动驾驶

高通模型库提供**75+AI模型**
适配其端侧硬件4倍加速

苹果半导体年研发费用**300亿美元**
支撑最新手机产品搭载苹果端侧AI

英特尔发起AI PC加速项目
超**100家**合作企业参与

技术展望

应用

- 新负载特征：更长输入和输出；复杂&多模型协作流程 (e.g., Agentic pipelines)；具身智能相关负载 (e.g., VLA / 3DGS)
- 新场景特征=>资源限制：泛端侧 (e.g., 手机 / PC / 机器人)

已有长足进展的技术 → **潜力方向**

算法层

语言
- 针对AR算法(语言模态)的并行输出方法 (e.g., speculative decoding, agentic generation) 、输入压缩方法
- 针对AR算法(语言模态)的并行输出方法

视觉
- 针对Diffusion算法的时间步压缩 (Training-free)
- 针对Diffusion算法的时间步压缩 (Training-based)
- AR算法设计 (视觉模态) 和相应加速方法

→ 多模态统一/融合的生成模型算法

- 提升针对逻辑推理能力的 Test Time Scaling

模型层

- 大语言模型、多模态理解大模型、视觉生成模型的模型压缩方法
  quantization, weight pruning, sparse attention, token merging, weight/activation sharing
- 新一代架构设计
  - 少参数小模型 => 端侧场景
  - 低复杂度结构 => 长文本场景
  - ?

系统层

- 算子优化
- Serving 系统优化
- 考虑具体下游应用推理流程的Serving优化
- 模型-系统协同设计
- 芯片: 3D堆叠, 芯粒, 构建软件生态?

语言模型参考文献

语言生成模型

1. **[Eagle]** Li, Yuhui, et al. "Eagle: Speculative sampling requires rethinking feature uncertainty." *arXiv preprint arXiv:2401.15077* (2024).
2. **[Eagle2]** Li, Yuhui, et al. "Eagle-2: Faster inference of language models with dynamic draft trees." *arXiv preprint arXiv:2406.16858* (2024).
3. **[SEDD]** Lou, Aaron, Chenlin Meng, and Stefano Ermon. "Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution." *Forty-first International Conference on Machine Learning*.
4. **[CLLMs]** Kou, Siqi, et al. "Cllms: Consistency large language models." *arXiv preprint arXiv:2403.00835* (2024).
5. **[AWQ]** Lin, Ji, et al. "AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration." Proceedings of Machine Learning and Systems 6 (2024): 87-100.
6. **[Quarot]** Ashkboos, Saleh, et al. "Quarot: Outlier-free 4-bit inference in rotated llms." arXiv preprint arXiv:2404.00456 (2024).
7. **[SpinQuant]** Liu, Zechun, et al. "SpinQuant--LLM quantization with learned rotations." arXiv preprint arXiv:2405.16406 (2024).
8. **[BitDistiller]** Du, Dayou, et al. "Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation." arXiv preprint arXiv:2402.10631 (2024).
9. **[EfficientQAT]** Chen, Mengzhao, et al. "Efficientqat: Efficient quantization-aware training for large language models." arXiv preprint arXiv:2407.11062 (2024).
10. **[StreamingLLM]** Xiao Guangxuan, et al. "Efficient Streaming Language Models with Attention Sinks." ICLR 2024.
11. **[Quest]** Tang, Jiaming, et al. "Quest: Query-Aware Sparsity for Efficient Long-Context LLM Inference." https://arxiv.org/abs/2406.10774
12. **[MoA]** Fu, Tianyu, et al. "MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression." https://arxiv.org/abs/2406.14909
13. **[MInference1.0]** Jiang, Huiqiang, et al. "MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention." https://arxiv.org/abs/2407.02490
14. **[PanGu-π]** Wang, Yunhe et al. "PanGu-π: Enhancing Language Model Architectures via Nonlinearity Compensation." ArXiv abs/2312.17276 (2023)
15. **[MiniCPM]** Hu, Shengding, et al. "Minicpm: Unveiling the potential of small language models with scalable training strategies." arXiv preprint arXiv:2404.06395 (2024).
16. **[Mamba]** Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." *arXiv preprint arXiv:2312.00752* (2023).
17. **[Mamba-2]** Dao, Tri, and Albert Gu. "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality." *arXiv preprint arXiv:2405.21060* (2024).
18. **[Jamba]** Lieber, Opher, et al. "Jamba: A Hybrid Transformer-Mamba Language Model." https://arxiv.org/abs/2403.19887
19. **[Zamba]** Glorioso, Paolo, et al. "Zamba: A Compact 7B SSM Hybrid Model." *arXiv preprint arXiv:2405.16712* (2024).
20. **[Hymba]** Dong, Xin, et al. "Hymba: A Hybrid-head Architecture for Small Language Models." *arXiv preprint arXiv:2411.13676* (2024).
21. **[TTT]** Sun, Yu, et al. "Learning to (learn at test time): Rnns with expressive hidden states." *arXiv preprint arXiv:2407.04620* (2024).

视觉模型参考文献

视觉生成模型

1. **[Sora]** Brooks, Tim, et al. "Video generation models as world simulators. 2024." URL https://openai. com/research/video-generation-models-as-world-simulators 3 (2024).
2. **[Open Sora]** Zheng, Zangwei, et al. "Open-sora: Democratizing efficient video production for all." *arXiv preprint arXiv:2412.20404* (2024).
3. **[Kling]** Kuaishou. Kling. https://klingai.kuaishou.com/
4. **[Hunyuan]** Kong, Weijie, et al. "HunyuanVideo: A Systematic Framework For Large Video Generative Models." arXiv preprint arXiv:2412.03603 (2024).
5. **[SD3]** Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." Forty-first International Conference on Machine Learning. 2024.
6. **[VAR]** Tian, Keyu, et al. "Visual autoregressive modeling: Scalable image generation via next-scale prediction." arXiv preprint arXiv:2404.02905 (2024).
7. **[LlamaGen]** Sun, Peize, et al. "Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation." arXiv preprint arXiv:2406.06525 (2024).
8. **[MAR]** Li, Tianhong, et al. "Autoregressive Image Generation without Vector Quantization." arXiv preprint arXiv:2406.11838 (2024).
9. **[Transfusion]** Zhou, Chunting, et al. "Transfusion: Predict the next token and diffuse images with one multi-modal model." arXiv preprint arXiv:2408.11039 (2024).
10. **[DMD]** Yin, Tianwei, et al. "One-step diffusion with distribution matching distillation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
11. **[ViDiT-Q]** Zhao, Tianchen, et al. "Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation." arXiv preprint arXiv:2406.02540 (2024).
12. **[DiTFastAttn]** Yuan, Zhihang, et al. "Ditfastattn: Attention compression for diffusion transformer models." arXiv preprint arXiv:2406.08552 (2024).
13. **[SageAttention]** Zhang, Jintao, et al. "SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration." arXiv preprint arXiv:2410.02367 (2024).
14. **[SANA]** Xie, Enze, et al. "Sana: Efficient high-resolution image synthesis with linear diffusion transformers." arXiv preprint arXiv:2410.10629 (2024).
15. **[Distilled Decoding]** Liu, Enshu, et al. "Distilled Decoding 1: One-step Sampling of Image Auto-regressive Models with Flow Matching." https://arxiv.org/abs/2412.17153

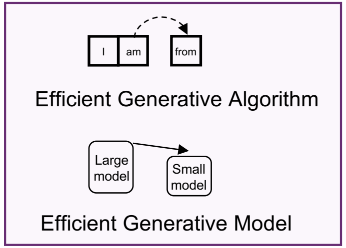
云侧优化参考文献

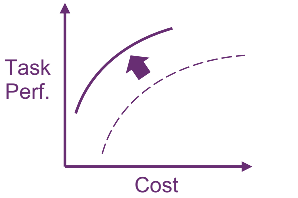
云侧系统优化

1. **[FlashAttention]** Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." Advances in Neural Information Processing Systems 35 (2022): 16344-16359.
2. **[FlashDecoding++]** Hong, Ke, et al. "FlashDecoding++: Faster Large Language Model Inference with Asynchronization, Flat GEMM Optimization, and Heuristics." *Proceedings of Machine Learning and Systems* 6 (2024): 148-161.
3. **[FlashDecoding]** Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. Flash-decoding for long-context inference. from https://crfm.stanford.edu/2023/10/12/flashdecoding.html.
4. **[ORCA]** Yu, Gyeong-In, et al. "Orca: A distributed serving system for {Transformer-Based} generative models." *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. 2022.
5. **[VLLM]** Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." *Proceedings of the 29th Symposium on Operating Systems Principles*. 2023.
6. **[SGLang]** Zheng, Lianmin, et al. "Sglang: Efficient execution of structured language model programs." *arXiv preprint arXiv:2312.07104* (2024).
7. **[DistServe]** Zhong, Yinmin, et al. "Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving." *arXiv preprint arXiv:2401.09670* (2024).

Research Goal

Develop efficient algorithms and models





Improve the perf.-cost trade-off

Team Website	Bilibili	GitHub Org.
https://nics-effalg.com/	https://space.bilibili.com/642618077	https://github.com/thu-nics

感谢聆听！欢迎讨论！

宁雪妃

2025.01

foxdoraame@gmail.com

Efficient AIGC工作介绍

https://www.bilibili.com/video/BV1AWCJY3EB8/