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

Introduction to NICS-EFC Lab

Efficient Algorithm Team

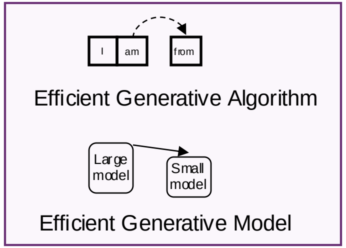
Xuefei Ning (宁雪妃)
Department of Electronic Engineering, Tsinghua University
foxdoraame@gmail.com

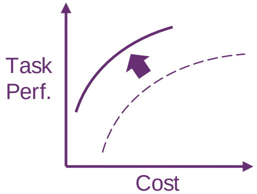
2024/11/24

NICS EFC

Team Overview

Research Goal
Develop **efficient algorithms and models**





Improve the perf.-cost trade-off

Team Website

https://nics-effalg.com/

Bilibili

https://space.bilibili.com/642618077

清华大学NICS-EFC实验室

GitHub Org.

https://github.com/thu-nics

Professor Yu Wang is the leader of the *Nanoscale Integrated Circuits and System - Energy Efficient Computing Lab* (***NICS-EFC***) in the Dept. EE at Tsinghua.

Research Assistant Professor Xuefei Ning is the leader of the *Efficient Algorithm Team* (***EffAlg***) in the ***NICS-EFC*** lab.

Team Overview

3 Ph.D. Students

Tianchen Zhao

Shiyao Li

Tianyu Fu

4 Master Students

Pu Lu

Enshu Liu

Jeff Chen

Tongcheng Fang

5 Graduate Student Interns

Yingchun Hu (Beihang), Hanlin Zhang (CMU), Rui Xie (SJTU), Jiayi Yang (Columbia), Songsheng Wang (UMacau)

10 Senior Undergraduate Student Interns

- **4th grade**: Tengxuan Liu, Yiran Shi, Qian Chen, Hongyu Zhu
- **3rd grade**: Dongyun Zou, Jidong Chen, Yichen You, Ruiqi Xie, Qinghao Han, Yi Ge

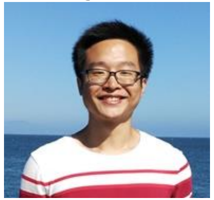
Alumni in 2024

- **Graduate Students**: Zixuan Zhou (graduate with honors, Bytedance), Junbo Zhao (Huawei)
- **Undergraduate Students**: Luning Wang (UMich)
- **Interns**: Peiran Yao, Lidong Guo, Haofeng Huang, Yuming Lou, Xianying Chen, Rui Wan, Luyue Zhang

Full-Time Researcher

Dr. Zhihang Yuan
(Infinigence-AI)

Co-Advisor on Many Projects



Dr. Zinan Lin
(Microsoft Research)

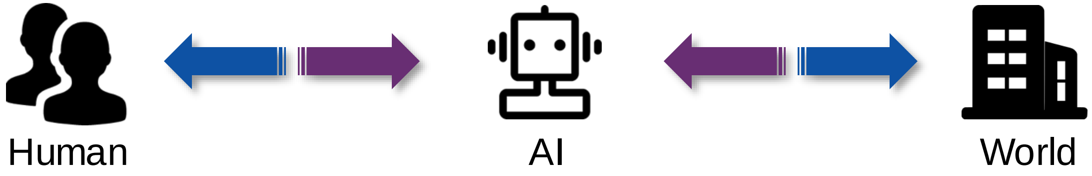
***Academic collaboration with folks from**: MSR, SJTU, HKU, Georgia Tech, KUL, UAlberta, …*

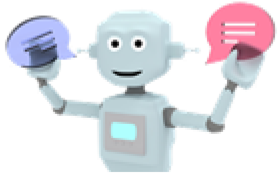
Application Goal

Goal: Let AI better interact with us and the world to serve us

Interaction with Human

Interaction with World



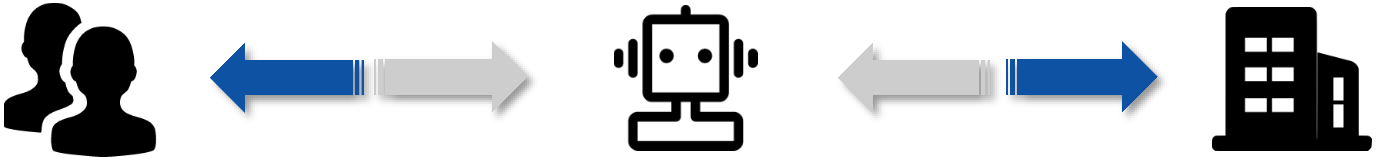


- Understand **human's instructions**
- Improve their **sensory experiences**

- Understand the **physical world**
- Make decisions to **influence the world**

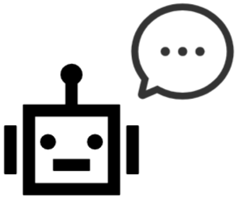
Generative AI

Interaction with Human | **Interaction with World**



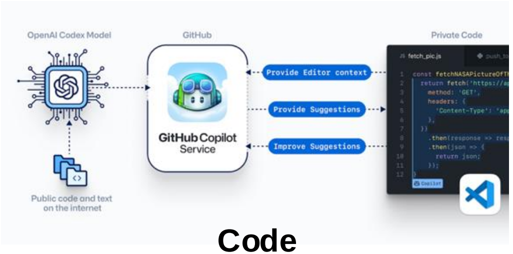
Human | AI | World

Generative tasks



Dialogue Generation

Image Generation



Code Generation

3D Assets Generation

Generative AI

AIGC, which uses generative models to generate content that satisfies human instructions, aims to make the content creation process more efficient and accessible[1].

Language Generation

Large Language Models: LLaMA-2-7B[2]

Visual Generation

Video Diffusion Models: Sora[3]

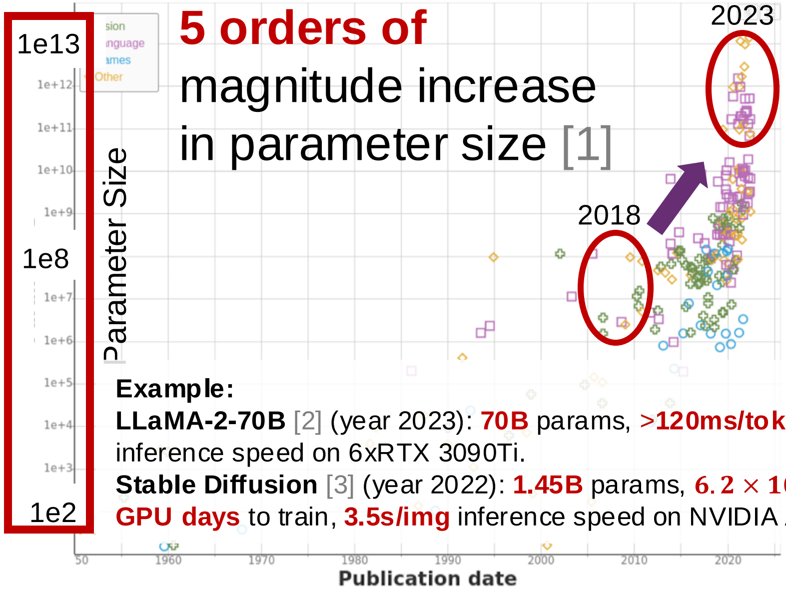
[1] Cao, Yihan, et al. "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt." *arXiv 2023*.
[2] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv 2023*.
[3] Brooks, Peebles, et al., "Video generation models as world simulators." *2024*.

Trend of Generative Models

The model size of generative models has being rapidly increasing

2018 - 2023

5 orders of magnitude increase in parameter size [1]



Example:

LLaMA-2-70B [2] (year 2023): **70B** params, **>120ms/token** inference speed on 6xRTX 3090Ti.

Stable Diffusion [3] (year 2022): **1.45B** params, 6.2×10^3 **GPU days** to train, **3.5s/img** inference speed on NVIDIA A100.

Stable Diffusion 1.5[3]
~1B Params

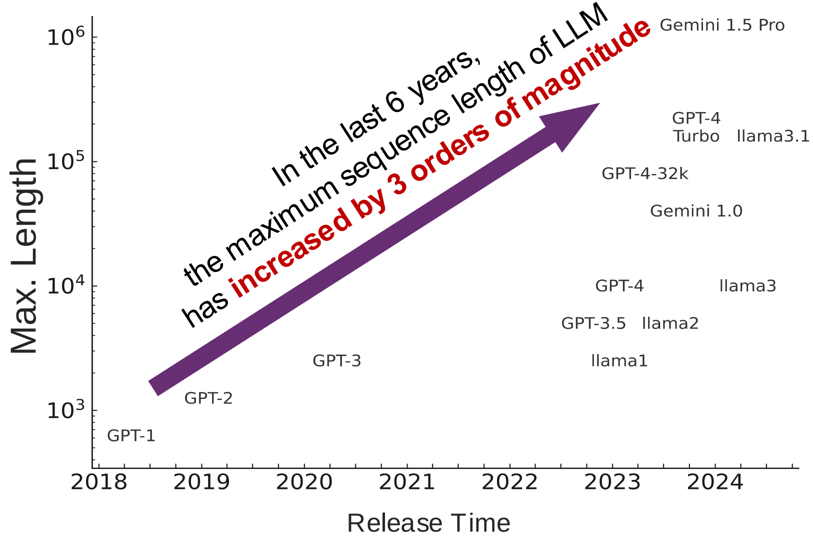
Flux[4]
~12B Params

[1] Villalobos et al. “Machine Learning Model Sizes and the Parameter Gap.” arXiv 2022.
[2] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv 2023*.
[3] Rombatch et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.
[4] black-forest-labs/flux: Official inference repo for FLUX.1 models

Trend of Generative Models

The input & output length has being rapidly increasing

Longer Sequence Length for Language



Higher Resolution / Longer Video Length for Vision

OpenAI
Meta AI
Google

OpenSORA[4]
generate Videos

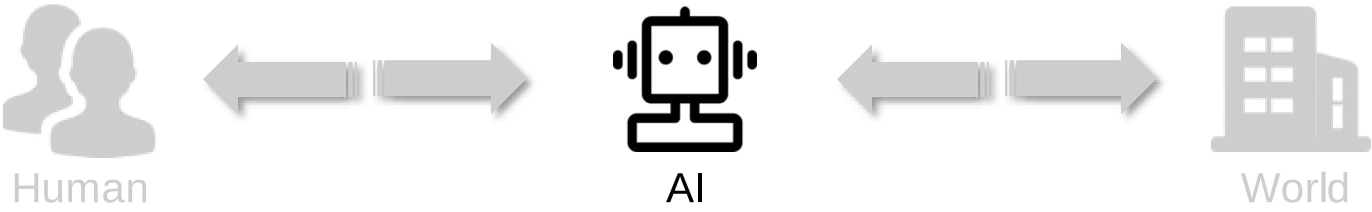
Pixart-sigma[5]
generates 4K image

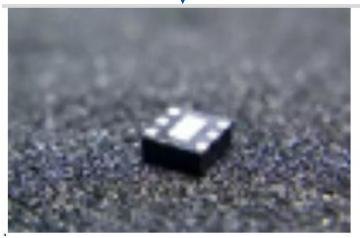
[1] Achiam, Josh, et al. "Gpt-4 technical report." arXiv 2023.
[2] Reid, Machel, et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." arXiv 2024.
[3] Dubey, Abhimanyu, et al. "The llama 3 herd of models." arXiv 2024.
[4] hpcaitech, "Open-SoRA: Democratizing Efficient Video Production for All." https://github.com/hpcaitech/Open-Sora
[5] Chen, Junsong et al. "PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation." arXiv 2024.

Application Scenario

Interaction with Human

Interaction with World





Sensor
Wearable Device

Mobile Phone
IoT Device

Smart City
Auto-driving Car

Smart City
Auto-driving Car

Challenge and Research Goal

- As the model size is scaling up, the demands for computing power are increasing
- Due to real-time, usable, privacy and other application demands, physical limitations of the scenario, as well as cost control considerations, models need to be deployed on computing devices with limited computing power and low storage, and are required to run under low budgets.
- How to deploy “large” generative models and satisfy the application’s efficiency requirements while maintaining algorithmic performance?

Our goal is to **improve the efficiency (e.g., latency, throughput, storage)** of generative models to satisfy the application requirement.

Research Overview

Research Goal：Efficient model inference for AIGC application

Application

Language Generation

Large Language Models
(e.g., LLaMA-2-7B)

Visual Generation

Diffusion Models
(e.g., Stable Diffusion 3)

Methodology: System-aware algorithm-level and model-level optimization

Technique

Algorithm-level

- **Diffusion Timestep Compression** — Tackling Many Timesteps of Diffusion
- **Non-Autoregressive Generation** — Tackling Full Autoregressive Generation of LLMs

Model-level

- **Structure Design**
- **Model Compression**

Research Summary

Efficient LLM/VLM

Overview

Survey [Under Review]

Survey on efficient LLM inference techniques

Algorithm-level

SoT [ICLR'24]

Parallel generation via prompting.
1.91~2.39x speed-up

Model-level

Sparse Attention

MoA [Under Review]

Decide the heterogeneous elastic rule of the attention span for each head.
5.5~6.7x throughput improvement

Pruning

EEP [Under Review]

Search the pruning pattern for MoE and use expert merging for finetuning.
48%~71% memory reduction,
1.11~1.40x speed-up,
better performance

Quantization

LLM-MQ [NeurIPS'23 Workshop]

Mixed-precision quantization.
2.8-bit quantization

MBQ [Under Review]

Modality-balanced quantization for VLM.
acc. improvement on MMMU: W3 up to **5.4%**, W4A8 up to **3.8%**

QLLM-Eval [ICML'24]

Evaluating the effect of quantization.
Providing knowledge and suggestions

Efficient Vision Generation

Algorithm-level

Time Step Compression

LCSC [Under Review]

Linear combination of checkpoints.
15~23x training acceleration,
1.25~2x timestep compression

USF [ICLR'24]

OMS-DPM [ICML'23]

Search for optimal diffusion schedulers.
1.5~2x speed-up

DD [Under Review]

Distill AR into Flow Matching, can achieve **>100x** speedup for Image AR model

Fast Compression

FlashEval [CVPR'24]

10x evaluation acceleration

Model-level

Quantization

MixDQ [ECCV'24]

Mixed-precision quantization.
3x memory decrease,
1.5x speed-up

ViDiT-Q [Under Review]

Quantization for DiT.
2.5x memory improvement,
1.5x speed-up

Local Attn. & Act. Sharing

DiTFastAttn [NeurIPS'24]

Window & reused attention for DiT.
1.6x speed-up

Research Summary

Overview

Survey
[Under Review]

Survey on efficient LLM inference techniques

Algorithm-level

SoT
[ICLR'24]

Parallel generation via prompting.
1.91~2.39x speed-up

Model-level

Sparse Attention

MoA
[Under Review]

Decide the heterogeneous elastic rule of the attention span for each head.
5.5~6.7x throughput improvement

Pruning

EEP
[Under Review]

Search the pruning pattern for MoE and use expert merging for finetuning.
48%~71% memory reduction,
1.11~1.40x speed-up,
better performance

Quantization

LLM-MQ
[NeurIPS'23 Workshop]

Mixed-precision quantization.
2.8-bit quantization

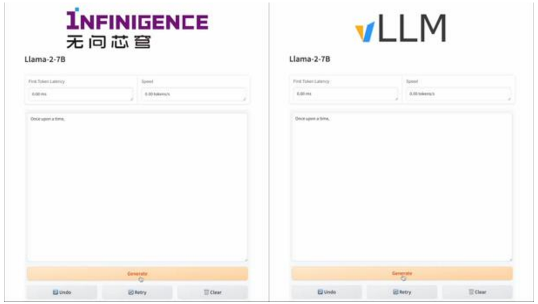
MBQ
[Under Review]

Modality-balanced quantization for VLM.
acc. improvement on MMMU: W3 up to **5.4%**, W4A8 up to **3.8%**

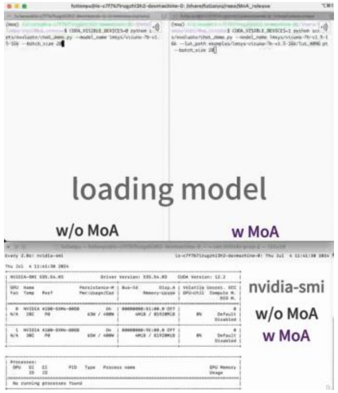
QLLM-Eval
[ICML'24]

Evaluating the effect of quantization.
Providing knowledge and suggestions

Efficient LLM/VLM



LLaMA-2-7B on AMD MI210
2× throughput improvement



Vicuna-7B on Nvidia-A100
batch size 20
end-to-end latency 2.3x

Research Summary

Algorithm-level
Time Step Compression

LCSC
[Under Review]

Linear combination of checkpoints.
15~23x training acceleration,
1.25~2x timestep compression

USF
[ICLR'24]

OMS-DPM
[ICML'23]

Search for optimal diffusion schedulers.
1.5~2x speed-up

DD
[Under Review]

Distill AR into Flow Matching, can achieve **>100x** speedup for Image AR model

Fast Compression

FlashEval
[CVPR'24]

10x evaluation acceleration

Model-level
Quantization

MixDQ
[ECCV'24]

Mixed-precision quantization.
3x memory decrease,
1.5x speed-up

ViDiT-Q
[Under Review]

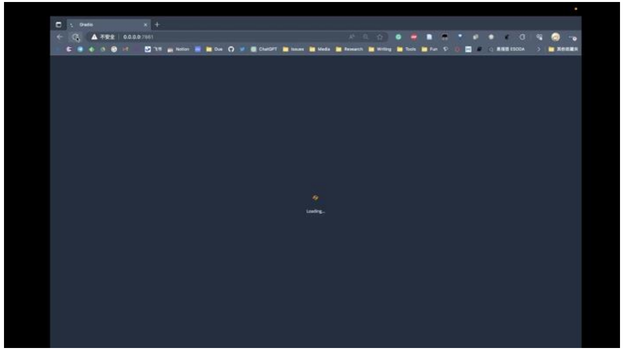
Quantization for DiT.
2.5x memory improvement,
1.5x speed-up

Local Attn. & Act. Sharing

DiTFastAttn
[NeurIPS'24]

Window & reused attention for DiT.
1.6x speed-up

Efficient Vision Generation



Stable Diffusion on a single NVIDIA A100 GPU, Achieving **6.9×** speed-up and reducing **1.5×** memory

Pixart-Sigma, 2K generation on NVIDIA A100 GPU
1.8x latency speedup

OpenSORA, 512x512x16 Frames, on NVIDIA A100 GPU,
2x Memory Savings, **1.7x** latency speedup

References

- Efficient LLM/VLM
  1. **SoT**: "Skeleton-of-Thought: Large Language Models Can Do Parallel Decoding." ICLR 2024. https://arxiv.org/abs/2307.15337
  2. **LLM-MQ**: "LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment." NeurIPS Workshop' 23.
  3. **QLLM-Eval**: "Evaluating Quantized Large Language Models." ICML 2024. https://arxiv.org/pdf/2402.18158
  4. **Survey**: "A Survey on Efficient Inference for Large Language Models." arXiv 2024. https://arxiv.org/abs/2404.14294
  5. **MoA**: "MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression." Under review. https://arxiv.org/abs/2406.14909
  6. **EEP**: "Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models." Under review. https://arxiv.org/abs/2407.00945
  7. **MBQ**: "MBQ: Modality-Balanced Quantization for Large Vision-Language Models." Under review.
- Efficient Vision Generation
  1. **OMS-DPM**: "OMS-DPM: Optimizing the Model Schedule for Diffusion Probabilistic Models." ICML 2023. https://arxiv.org/abs/2306.08860
  2. **USF**: "A Unified Sampling Framework for Solver Searching of Diffusion Probabilistic Models." ICLR 2024. https://arxiv.org/abs/2312.07243
  3. **FlashEval**: "FlashEval: Towards Fast and Accurate Evaluation of Text-to-image Diffusion Generative Models." CVPR 2024. https://arxiv.org/abs/2403.16379
  4. **LCSC**: "Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models Better." Under review. https://arxiv.org/abs/2404.02241
  5. **MixDQ**: "MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization. " ECCV 2024. https://arxiv.org/abs/2405.17873
  6. **ViDiT-Q**: "ViDiT-Q: Efficient and Accurate Quantization of DiTs for Image and Video Generation. " Under review. https://arxiv.org/abs/2406.02540
  7. **DiTFastAttn**: "DiTFastAttn: Attention Compression for DiT Models." NeurIPS 2024. https://arxiv.org/abs/2406.08552
  8. **DD**: "Distilling Autoregressive Models into Few Steps 1: Image Generation." Under review.

We're Now Working On ...

- **[Application-driven]** Applying and analyzing efficiency techniques on *multi-modality understanding models & video generative models*, to use them well
- **[Application-driven]** Developing methods for efficient *long-context inference*
- **[Application-driven]** *Pushing to the edge*: We want high compression ratio or a small model from scratch
  - Training-free techniques -> Training-based techniques
  - *Integrating efficiency techniques together*, to understand their interplay and use them well
  - How can we still *inherit the knowledge* well, or there is not difference from training from scratch?
- **[Algorithm-driven]** *Developing efficient generative algorithm*: Combining the benefits of data-space autoregressive models and flow matching

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

Thank You!

We're looking for self-motivated students, interns, and other form of collaborations! If interested, please drop me an email with yours thoughts and information.

Team Leader: **Xuefei Ning** foxdoraame@gmail.com
Lab Leader: **Prof. Yu Wang** yu-wang@tsinghua.edu.cn

Team Website
https://nics-effalg.com/

Book
"Efficient Deep Learning: Model Compression and Design"

Sponsors

HUAWEI, OPPO, Baidu 百度, TOYOTA, CHINA TOWER 中国铁塔, zongmu, Mercedes-Benz, 美团

(Super) Close Collaboration

INFINIGENCE 无问芯穹

https://cloud.infini-ai.com/promotion

Interns Wanted!

Welcome to follow the TechReview Wechat official account

Welcome to Our Talk Session

Time	Speaker	Position	Topic	Title
16:45 – 17:00	Zhihang Yuan	Researcher	Efficient Visual Generation	**[NeurIPS24]** DiTFastAttn: Attention Compression for Diffusion Transformer Models
17:00 – 17:15	Tianchen Zhao	Ph.D. Student	Efficient Visual Generation	ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformer for Image and Video Generation
17:15 – 17:30	Enshu Liu	Master Student	Efficient Visual Generation	Distilling Autoregressive Models into Few Steps for Image Generation
17:30 – 17:45	Enshu Liu	Master Student	Efficient Visual Generation	Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models Better
17:45 – 18:00	Tianyu Fu	Ph.D. Student	Efficient LLM	MoA: Mixture of Sparse Attention for Automatic Large Language Model Compression
18:00 – 18:15	Enshu Liu	Master Student	Efficient LLM	Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models
18:15 – 18:30	Shiyao Li	Ph.D. Student	Reasoning of LLM	**[NeurIPS’24]** Can LLMs Learn by Teaching for Better Reasoning? A Preliminary Study
18:30 – 18:45	Lidong Guo	Ph.D. Student	3D Modelling	**[NeurIPS’24]** Rad-NeRF: Ray-decoupled Training of Neural Radiance Field