Yuwei An

ayw.sirius19@gmail.com | Google Scholar | Website

Education

Carnegie Mellon University, MS in Electrical and Computer Engineering — Sept 2023 – May 2025

- GPA: 4.0/4.0 (Transcript)

Tsinghua University, BS in Computer Science and Technology — Sept 2019 – June 2023

- GPA: 3.77/4.0 (Transcript)

Publications

Multiverse: Your Language Models Secretly Decide How to Parallelize and Merge Generation

Xinyu Yang*, Yuwei An*, Hongyi Liu, Tianqi Chen, Beidi Chen

NeurIPS 2025 Spotlight

LMCACHE: An Efficient KV Cache Layer For Enterprise-scale LLM Inference

Yihua Cheng*, Yuhan Liu*, Jiayi Yao*, Yuwei An, Xiaokun Chen, Shaoting Feng, Yuyang Huang, Samuel Shen, Kuntai Du, Junchen Jiang

HyperRAG: Enhancing Quality-Efficiency Tradeoffs in Retrieval-Augmented Generation with Reranker KV-Cache Reuse

Yuwei An, Yihua Cheng, SeoJin Park, Junchen Jiang

Strata: Hierarchical Context Caching for Long Context Language Model Serving

Zhiqiang Xie, Ziyi Xu, Mark Zhao, Yuwei An, Vikram Sharma Mailthody, Scott Mahlke, Michael Garland, Christos Kozyrakis

IFMoE: An Inference Framework Design for Fine-grained MoE

Yuwei An, Zhuoming Chen, Beidi Chen

NeurIPS 2024 ML For Systems workshop

Controllable Mesh Generation Through Sparse Latent Point Diffusion Models

Zhaoyang Lyu*, Jinyi Wang*, Yuwei An, Ya Zhang, Dahua Lin, Bo Dai

CVPR 2023

Open-Source Project

SGLang — github:SGLang

- SGLang is a fast serving framework for large language models and vision language models.
- Lead Developer for the torch compile backend and piecewise cuda graph support.

LMCache — github:LMCache

- LMCache is an LLM serving engine extension to reduce TTFT and increase throughput.
- Lead Developer for the integration of LMCache and SGLang for better global KV Cache management.

Experience

Research Intern@CMU *Advisor: Beidi Chen* — Sept 2023 - May 2025

- *Multiverse Project*: Build Multiverse, a new generative modeling framework that enables native parallel generation and achieves comparable reasoning ability as autoregressive model.
- *IFMoE Project*: Build IFMoE, a fast MoE inference framework with speculative decoding for expert Layer.
- *Herd Project*: Build Herd, an algorithm to achieve multi-batch inference with dynamic pruning.
- *HyFlow Project (In progress)*: Build HyFlow, a system design for DAG-based program service such as MCTS and Agent workflow, achieving high throughput and maintaining SLO guarantees.
- Also Collaborated with: Tianqi Chen, Chenyan Xiong

Research Intern@Uchi/TensorMesh *Advisor: Junchen Jiang* Sept 2024 - Sept 2025

- *HyperRAG Project*: Build HyperRAG, a high-throughput serving system for retrieval-augmented generation (RAG) which focuses on optimizing the accuracy-latency tradeoff curve with KV-Cache management.
- *LMCache Project*: Contribute to LMCache on: Integration with SGLang, one of the most popular inference engine. Mainly work on offloading-computation overlap and layerwise KV Cache offloading.
- Also Collaborated with: Seojin Park

Research Intern@UCLA/SGLang *Advisor: Ying Sheng & Harry Xu* Aug 2025 - Now

- *Agent-Prism Project (In progress)*: Built Agent-Prism, an inference system designed for SLO-aware, high-throughput service in multi-agent inference. Equipped with dynamic KV-cache allocation, it implements an adaptive scheduler to efficiently serve multi-agent, multi-model workloads.
- *SGLang Project*: Led the development of torch.compile integration and piecewise CUDA Graph support for SGLang, improving prefill performance and enabling KV-cache offloading during decode.
- Also Collaborated with: Lianmin Zheng

Undergrad Research Experience Sept 2019 - July 2023

- Work with *Dai Bo* @ Shanghai AI Lab on 3D mesh generation
- Work with *Fanjin Zhang* @ Tsinghua University on Graph Neural Network benchmark
- Work with *Youyou Lu* @ Tsinghua University on Multi Tenant Embedding serving

Teaching

TA @ CMU 18789 Deep Generative Model 2025

Tutorial @ SIGCOMM 2025: Networking for Stateful LLM Inference 2025

Awards

Outstanding Graduate Student of Computer Science and Technology Department, Tsinghua University 2023

Tsinghua University Excellent Academic Scholarship 2022

Tsinghua University Excellent Academic Scholarship 2021

Tsinghua University Freshman Scholarship 2019

Technologies

Languages: Cpp, C, Java, Objective-C, SQL, Matlab, Python, Cuda, Typescript

Software: PyTorch, React, Spring Boot, .NET