Rongxiang Wang (434) 257-9768 | rxpwang.github.io | waq9hw@virginia.edu

EDUCATION

University of Virginia, **PhD in CS** (GPA 4.0/4.0) Aug. 2022 – Current

Research theme: on-device MLsys; current focus: efficient serving systems for speech models. Advisor: Felix Lin

Tsinghua University, Bachelor of Engineering in Automation Aug. 2017 – Jun. 2022

EXPERIENCE

ZOOM Communications, Inc, GenAI Research Intern May. 2025 – Nov. 2025

Research theme: On-device LLM, enabling dynamic sparsity for quantized LLM with new GPU kernel design.

PUBLICATIONS

[1] "Enabling Dynamic Sparsity in Quantized LLM Inference," Rongxiang Wang, Kangyuan Shu, Felix Xiaozhu Lin, in Arxiv, 2025 (in submission)

[2] "WhisperFlow: speech foundation models in real time," Rongxiang Wang, Zhiming Xu and Felix Xiaozhu Lin, in MobiSys, 2025 (Acceptance rate: 18.02%, 42 accepted in total 233 submission)

[3] "Turbocharge Deep Speech Understanding with Pilot inference," Rongxiang Wang and Felix Xiaozhu Lin, in MobiCom, 2024 (Acceptance rate: 19.09%, 55 accepted in total 288 submission in Winter run)

[4] "Proto: A Guided Journey through Modern OS Construction," Wonkyo Choe*, Rongxiang Wang*, Afsara Benazir*, Felix Xiaozhu Lin, in SOSP, 2025 (Acceptance rate: ~20%)

[5] "AnA: An Attentive Autonomous Driving System," Wonkyo Choe, Rongxiang Wang, and Felix Xiaozhu Lin, in ASPLOS, 2025 (Acceptance rate: 12.63%, 74 accepted in total 586 submission in Spring + Summer run)

[6] "Profiling Apple Silicon Performance for ML Training," Dahua Feng*, Zhiming Xu*, Rongxiang Wang, and Felix Xiaozhu Lin, in Arxiv, 2025

[7] "Accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data with scCASE," Songming Tang, Xuejian Cui, Rongxiang Wang, Sijie Li, Siyu Li, Xin Huang, and Shengquan Chen, in Nature Communication, 2024

[8] "ASTER: accurately estimating the number of cell types in single-cell chromatin accessibility data," Shengquan Chen, Rongxiang Wang, Wenxin Long, and Rui Jiang, in Bioinformatics, 2023

RESEARCH PROJECTS

Dynamic-sparsity-aware low-bit LLM inference system University of Virginia, May 2025 – Nov 2025

- Targets on-device LLM decoding on consumer GPUs, resolving the mismatch between dynamic activation sparsity and low-bit quantization.
- Proposes Zigzag weight layout and sparsity-aware GPU kernels, enabling structured skipping, load-balanced parallelism, and efficient index collection.
- Achieves up to 1.55× decoding speedup over dense 4-bit baselines with negligible accuracy loss (<0.4 perplexity) across Llama-2/3 (7B, 8B, 13B, 70B) models on Apple A19 Pro, M2, M2 Pro, and M2 Max.

A real time serving system for speech foundation model University of Virginia, Jan 2024 – Dec 2024

- Targets edge devices with heterogeneous processors, aiming to speed up speech foundation models for streaming speech processing and reduce the per word latency of the system.

- Proposes hush word, a novel adversarial attack to reduce the encoding redundancy. Proposes beam pruning, a reference guide method tailored for streaming tasks to reduce the decoding redundancy. Proposes CPU/GPU pipelinging to better utilize the hardware resources and reduce the overall latency.
- Achieves a 2x per word latency speedup, with as low as 0.5s per word latency and ~7 W power consumption.

On-device deep speech understanding University of Virginia, Mar 2023 – Dec 2023

- Targets mobile devices with limited hardware capabilities, aiming to speed up attention-based model local processing in streaming scenarios, as well as reduce offloading in collaboration with the cloud.
- Proposes pilot inference, which periodically processes partial data to attain tentative information that helps with local processing speed up and selective offloading. Proposes beam reduction, beam search termination prediction and CTC prefix scoring speedup for local execution, selective offloading for cloud collaboration.
- Achieves a 2x local processing speed up and reduces offloading by 50% in collaboration with the cloud.

Profiling Apple Silicon performance for ML training University of Virginia, Aug 2024 – Dec 2024

- Targets machine learning training on apple silicon scenarios, aiming to better understand the training performance and how it compares to other hardware platforms.
- Proposes profiling settings that covers state-of-the-art generative models as well as micro-benchmarks on basic machine learning kernels. Design and conduct experiments on different hardware platforms.
- Provide a comprehensive report and discussion on the profiling results. Deliver suggestions to practitioners.

Redesign of autonomous driving stack University of Virginia, Aug 2022 – Feb 2023

- Targets in-vehicle autonomous driving scenarios, aiming to reduce the computational cost of the autonomous driving system and make it more responsive.
- Proposes an attentive method that helps the perception module better focus on important regions based on the information from the downstream planning module.
- Reduces the computation cost of the autonomous driving system by 44% and avoids collisions by 2x.

Single cell genomic data enhancement (undergraduate thesis) Tsinghua University, Aug 2021 – May 2022

- Targets single-cell data with massive sampling loss, aiming to recover the data and enhance the data quality.
- Proposes an optimization-based low-rank matrix factorization method to assist with data recovery.
- Improves the data quality, as reflected in clustering metrics by 30%.

TEACHING

Teaching Assistant for the undergraduate Operating System (CS4414) in University of Virginia, Spring 2023, Spring 2024, Spring 2025. Responsibility: help ~150 students to understand the course project, which builds an OS prototype on Raspberry Pi3. Paper on the teaching OS published in SOSP 2025.

AWARDS

Outstanding Graduate Research Award, UVA Department of Computer Science, 2024 – 2025

Early-Stage Research Award, UVA LINK LAB, 2024 – 2025

School Scholarships For Academic Excellence, Tsinghua University, 2017 – 2018

Gold Medal, China National Biology Olympiad, 2016

SKILLS

Programming Languages Python, C, C++, Metal **Software** Pytorch, Numpy, Pandas, Linux

Specializations Machine learning, GPU programming, Transformer-based foundation models