Jiesong Liu

984-325-3859 | jliu93@ncsu.edu | linkedin.com/in/jiesong-liu | website

Education

North Carolina State University Raleigh, NC
Ph.D. in Computer Science *Aug. 2023 – 2028 (Expected)*

- Advisor: Dr. Xipeng Shen
- Research Interests: (1) **ML Systems & Scaling**: Distributed training/inference, TPU/GPU scaling, and system-algorithm co-design. (2) **Agentic Workflows**: Scalable infrastructure for multi-modal agents and compiler-driven optimization. (3) **Efficient AI**: Model compression (quantization, merging), post-training, and reasoning.

Renmin University of China Beijing
Bachelor of Computer Science *Sep. 2019 – June 2023*

- GPA: 3.8/4.0

Research Experience

North Carolina State University Aug. 2025 – Present
Research Assistant *Raleigh, NC*

Distributed Optimization Framework for Scalable Agentic Workflows

- **Developed a distributed optimization framework** for multi-agents tool usage, targeting latency reduction in distributed environments. Integrated LLM-driven compiler optimizations to automate the generation of efficient execution kernels for diverse agent tasks.
- Engineered **Multi-Level API Fusion** to merge heterogeneous operators (e.g., GroundingDINO + SAM) into composite kernels, eliminating serialization overhead via **in-memory tensor passing**.
- Achieved **1.83x speedup** (9.48s $\rightarrow$ 3.56s) on complex image editing pipelines by reducing network RTTs and optimizing GPU memory management (warm-start caching).

Google DeepMind May 2025 – Dec. 2025
Student Researcher *Mountain View, CA*

Personalizing LLM with Interpretable Embeddings for Better User Understanding

- Advanced a novel **User-LLM** framework that leverages interpretable user embeddings to enable Large Language Models (Gemma) to reason about user timelines and preferences.
- Designed a **hybrid encoder-decoder architecture** that fuses user signals into the LLM backbone; achieved a **6% accuracy gain** on MovieLens benchmarks.
- Developed a **Latent Reasoning** mechanism for post-training LLM, enabling the model to generate user profiles and interpretable justifications for its recommendations.
- Scaled training on **TPU clusters** using Xmanager; optimized data pipelines and reduced memory overhead to facilitate efficient end-to-end fine-tuning of the User LLM.

North Carolina State University Nov. 2021 – May 2025
Research Assistant *Raleigh, NC*

Adaptive Speculative Decoding for Large Language Models

- Developed the first **on-the-fly adaptive speculative decoding** framework, a drop-in solution that dynamically adjusts speculation window size (γ) and draft model selection without ahead-of-time training. (*ACL 2025*)
- Designed multiple **agile online adaptation methods**, including state machine-based mechanisms, cache-enabled FSM, reinforcement learning-based approaches, and token accuracy-based online optimization to maximize decoding efficiency.
- Achieved **3.55–16.48% speed improvement** over standard speculative decoding and **1.2–3.4×** speedups over default LLM inference across various **LLM architectures, GPUs, and inference tasks**.

Fourier Token Merging for Efficient Image Generation

- Proposed a lightweight framework that transforms intermediate tokens via **Discrete Fourier Transform (DFT)**, utilizing low-frequency structural priors to perform robust token clustering and merging. (*NeurIPS 2025*)

- Achieved up to **25% reduction in inference latency** on Stable Diffusion v1.5 while maintaining visual fidelity (comparable FID scores).
- Derived theoretical error bounds for token merging, mathematically proving that frequency-domain clustering minimizes attention approximation error compared to spatial-domain methods.

Efficient Deep Learning for Microcontrollers

- Proposed **TREC (Transient Redundancy Elimination-based Convolution)** as a novel compilation optimization to exploit transient data redundancies in DNNs. Introduced a **learned LSH clustering mechanism** to group similar input vectors and reuse computation results, reducing **computational overhead** by over 96% and achieving **3.4-5× speedups** with minimal accuracy loss. *(ASPLOS 2023)*
- Proposed **Generalized Reuse**, a framework that expands computation reuse strategies in neural networks, **yielding 1.03-2.2× inference speedups** or **1-8% accuracy improvements** across diverse architectures. *(ASPLOS 2025)*
- Developed analytical models to efficiently navigate the **exponential reuse pattern search space**, enabling systematic selection of optimal reuse strategies by predicting accuracy-latency trade-offs.
- Designed **space-efficient reuse and clustering techniques** by embedding the LSH matrix into the weight matrix to enable large-scale DNN deployment on microcontrollers with less than 1 MB of memory, overcoming extreme hardware constraints.

Publication

[**NeurIPS'25**]"Fourier Token Merging: Understanding and Capitalizing Frequency Domain for Efficient Image Generation "; **Jiesong Liu**, Xipeng Shen.

[**ACL'25**]"A Drop-In Solution for On-the-Fly Adaptation of Speculative Decoding in Large Language Models"; **Jiesong Liu**, Brian Park, Xipeng Shen.

[**ASPLOS'25**]"Generalizing Reuse Patterns for Efficient DNN on Microcontrollers"; **Jiesong Liu**, Bin Ren, Xipeng Shen.

[**NeurIPS'24**]"UQ-guided Hyperparameter Optimization for Iterative Learners"; **Jiesong Liu**, Feng Zhang, Jiawei Guan, Xipeng Shen.

[**VLDB'24**]"A Systematic Study on Early Stop Metrics in HPO and the Implications of Uncertainty"; Jiawei Guan, Feng Zhang, **Jiesong Liu**, Xipeng Shen.

[**TC'24**]"Enabling Efficient Deep Learning on MCU with Transient Redundancy Elimination"; **Jiesong Liu**, Feng Zhang, Jiawei Guan, Hsing-Hsuan Sung, Xiaoyong Du, Xipeng Shen.

[**TPDS'24**] "G-Learned Index: Enabling Efficient Learned index on GPU"; **Jiesong Liu**, Feng Zhang, Lv Lu, Xiaoyong Du, Guoliang Li, Dong Deng.

[**ASPLOS'23**] "Space Efficient TREC for Enabling Deep Learning on Microcontrollers"; **Jiesong Liu**, Feng Zhang, Jiawei Guan, Hsing-Hsuan Sung, Xiaoyong Du, Xipeng Shen.

[**VLDB'23**] "Approximating Probabilistic Group Steiner Trees in Graphs"; Shuang Yang, Yahui Sun, **Jiesong Liu**, Xiaokui Xiao, Ronghua Li, Zhewei Wei.

[**TPDS'22**] "Exploring Query Processing on CPU-GPU Integrated Edge Device"; **Jiesong Liu**, Feng Zhang, Hourun Li, Dalin Wang, Weitao Wan, Xiaokun Fang, Jidong Zhai, Xiaoyong Du.

[**NeurIPS'22**] "TREC: Transient Redundancy Elimination-based Convolution"; Jiawei Guan, Feng Zhang, **Jiesong Liu**, Hsing-Hsuan Sung, Ruofan Wu, Xiaoyong Du, Xipeng Shen.

Honors

2024 NCSU Travel Award

2023 NCSU University Graduate Fellowship

2023 NCSU Graduate Merit Awards

2020 & 2021 National Scholarships of China (highest scholarship for Chinese undergraduate)

2022 SenseTime Scholarship (30 students selected from across China)

2018 National Olympiad in Informatics (NOI)