

Liang Qiao(乔良)

☎ + (86)17335160870 ✉ ql1an9@mail.ustc.edu.cn 🌐 qiaolian9 📦 qiaolian9.github.io

Education

- **USTC(中国科学技术大学)** Sep. 2020 - Jun. 2026
Ph.D. candidate in Computer Science; Advisor: Prof. Hong An (安虹教授)
Interests: **MLSys: Deep Learning Compiler & Sparse Attention**; GPA: 4.01/4.3
Hefei
- **Nankai University(南开大学)** Sep. 2016 - Jun. 2020
Bachelor of Management, Business School
Tianjin

Selected Publications

- Pruner: A Draft-then-Verify Exploration Mechanism to Accelerate Tensor Program Tuning** 📄 🌐 ASPLOS 2025
CCF-A
[Liang Qiao](#), Jun Shi, Xiaoyu Hao, Xi Fang, et al. (including Hong An, Xulong Tang, NIO chip teams, etc.)
- FcTC-UNet: Fine-grained Combination of Transformer and CNN for Thoracic Organs Segmentation** 📄 EMBC 2022
CAAI-B
[Liang Qiao](#), Qiang Liu, Jun Shi, Minfan Zhao, Hongyu Kan, Zhaohui Wang, Hong An, et al.
- CIEplorer: Microarchitecture-Aware Exploration for Tightly Integrated Custom Instruction** 📄 🌐 ICS 2025
CCF-B
Xiaoyu Hao, Sen Zhang, [Liang Qiao](#), Qingcai Jiang, Jun Shi, Junshi Chen, Hong An, Xulong Tang, et al.
- Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study** 📄 🌐 The Lancet Oncology
July. 2024
Anindo Saha, Joeran S Bosma, Jasper J Twilt, and [others \(included in PI-CAI consortium\)](#) SCI Q1, IF=35.9

Others Publications 📄

Projects

- Sparse Attention for MMDits** Jan. 2025 - Now
Leader
• A training-free row(Q) and column(KV) sparse mechanism guided by the prompt part of the attention score.
• Sparse kernel implementation (based on [flashinfer](#)) for Q-sparse Attention and corresponding Pre-/Post-Linears.
• Prepare for submitting to ICLR'26.
- Deep Learning Compiler: Auto-tuning (🌐 Pruner)** Jun. 2023 - Oct. 2024
Leader
• In collaboration with [Xulong Tang](#) from University of Pittsburgh, and [NIO chip\(神玢\) team](#).
• Pruner is a "**Draft-then-Verify**" exploration mechanism that accelerates the schedule search process.
• MoA-Pruner uses a momentum online adaptation strategy to address the cross-platform online unawareness.
- Torch2Tensor** 🌐 Spring 2023
• A toy compiler for generating Tensor Program from Torch (based on Torch FX, TVM Relax & [broccoli](#)).
- Others Projects:** Huawei NPU system simulator, medical image segmentation, etc. Aug. 2020 - May. 2022

Experiences

- Alibaba Cloud** May. 2022 - Aug. 2022
Summer Internship
• Working on fault prediction and anomaly detection based on AI algorithms.
Hangzhou

Others

- Talk:** Pruner: A Draft-then-Verify Exploration Mechanism to Accelerate Tensor Program Tuning 3rd Apr. 2025
• **Session 9A: ML Compilers @ ASPLOS'25**, Rotterdam, The Netherlands
- Patent:** “算子配置方法、深度学习编译器、智能设备及存储介质” (CN118862770A) Oct. 2024
- Competition:** The 5th place (**Team: Swangeese**) @ **The PI-CAI Challenge** May. 2022 - Apr. 2023
- Scholarship:** First-class academic scholarship (Master's & Doctoral) 2021 - 2024