Siyuan Feng

Assistant Professor

syfeng@sii.edu.cn | syfeng.net | GitHub | Google Scholar

Experience

Assistant Professor - Shanghai Innovation Institute — Jul 2025 - Present

- Research area: distributed machine learning systems, including inference, training, and RL systems for LLMs and other models.
- Exploring new programming paradigms for machine learning models using compilation techniques.
- Major focus on hardware-native machine learning systems.

Education

Ph.D., Computer Science - Shanghai Jiao Tong University — Sep 2020 - Jun 2025

- Research area: machine learning systems and machine learning compilers.
- Advisor: Prof. Yong Yu and Prof. Weinan Zhang at APEX Data & Knowledge Management Lab.
- Worked closely with Prof. Tianqi Chen on Apache TVM and the MLC community.

Bachelor of Engineering, Computer Science - Shanghai Jiao Tong University — Sep 2016 - Jul 2020

- B.Eng. in Computer Science and Technology, Zhiyuan Honors Program.
- Member of ACM Honors Class, an elite CS program for top 5% talented students.

Teaching

System for Artificial Intelligence - Instructor — Fall

- Graduate course; 3 credits; 48 hours over 16 weeks
- Covers modern AI computing hardware, programming paradigms, deep learning frameworks, machine learning compilers, distributed training, and inference systems.
- Designed around theoretical study and hands-on projects for building production-grade AI systems.

Open Source Projects

Nex Agentic Models - Hugging Face Model Scope 47.1k+ — 2025 - Present

- Worked as one of the project leaders for the Nex agentic model line, from Nex-N1 environment-construction research to Nex-N2 open model releases including Pro, mini, and fp8 variants.

Apache TVM - GitHub ★ 13,556 ⑂ 3,915 — 2019 - Present

- Led TensorIR and co-led TVM Unity/Relax, shaping compiler IRs for tensor programs and dynamic end-to-end ML workloads.
- Contributed to TVMScript, Meta-Schedule, runtime, frontend infrastructure, and project governance.

MLC-LLM - GitHub ★ 22,920 ⑂ 2,081 — 2023 - Present

- Co-developed a universal LLM deployment stack across CUDA, ROCm, Metal, Vulkan, OpenCL, WebGPU, mobile, desktop, and browser runtimes.
- Built compiler-backed model optimization, distributed inference, and OpenAI-compatible serving paths.

Web-LLM - GitHub ★ 18,329 ⑂ 1,321 — 2023 - Present

- Built browser-native LLM inference on WebGPU and WebAssembly as the web runtime in the MLC ecosystem.

PyPTO - GitHub ★ 95 ⑂ 75 2025 - Present

- Developed a Python-first tile/tensor programming paradigm for hardware-native AI systems with white-box compilation.

Selected Publications

Expert-as-a-Service: Towards Efficient, Scalable, and Robust Large-scale MoE Serving - SC 2026

Ziming Liu, Boyu Tian, Guoteng Wang, Zhen Jiang, Peng Sun, Zhenhua Han, Tian Tang, Xiaohe Hu, Yanmin Jia, Yan Zhang, He Liu, Mingjun Zhang, Yiqi Zhang, Qiaoling Chen, Shenggan Cheng, Mingyu Gao, Yang You†, **Siyuan Feng**†

DistFlow: A Fully Distributed RL Framework for Scalable and Efficient LLM Post-Training - ICML 2026

Zhixin Wang, Jiaming Xu, Tianyi Zhou, Mingjun Zhang, Liming Liu, Jiarui Hu, Dian Yang, Tongyu Wang, Ping Zhang, Jinlong Hou, **Siyuan Feng**†, Yuan Qi†, Yuan Cheng†

ReSpec: Towards Optimizing Speculative Decoding in Reinforcement Learning Systems - MLSys 2026

Qiaoling Chen, Zijun Liu, Peng Sun, Shenggui Li, Guoteng Wang, Ziming Liu, Yonggang Wen, **Siyuan Feng**, Tianwei Zhang

Productively Deploying Emerging Models on Emerging Platforms: A Top-Down Approach for Testing and Debugging - ISSTA 2025

Siyuan Feng*, Jiawei Liu*, Ruihang Lai, Charlie F. Ruan, Yong Yu, Lingming Zhang, Tianqi Chen

Relax: Composable Abstractions for End-to-End Dynamic Machine Learning - ASPLOS 2025

Ruihang Lai*, Junru Shao*, **Siyuan Feng***, Steven S. Lyubomirsky*, Bohan Hou, Wuwei Lin, Zihao Ye, Hongyi Jin, Yuchen Jin, Jiawei Liu, Lesheng Jin, Yaxing Cai, Ziheng Jiang, Yong Wu, Sunghyun Park, Prakalp Srivastava, Jared Roesch, Todd C. Mowry, Tianqi Chen

TensorIR: An Abstraction for Automatic Tensorized Program Optimization - ASPLOS 2023

Siyuan Feng*, Bohan Hou*, Hongyi Jin, Wuwei Lin, Junru Shao, Ruihang Lai, Zihao Ye, Lianmin Zheng, Cody Hao Yu, Yong Yu, Tianqi Chen

CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario - WWW 2019

Huichu Zhang, **Siyuan Feng**, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, Zhenhui Li

*equal contribution; †corresponding author

Open Source Community

Apache Software Foundation community member - ASF Mar 2024 - Present

Apache TVM project management committee (PMC) member - Apache TVM Mar 2022 - Present