Jiaqi Mo

Madison, WI | mo29@wisc.edu | jmo25.github.io | github.com/jmo25

Research Interests

Efficient Machine Learning, Model Architectures, ML Systems

Education

University of Wisconsin-Madison, B.S. in Computer Sciences — Jan. 2025 – Expected Dec. 2026

- Overall GPA: 3.96/4.00, Dean's List (Every Eligible Semester)

Hangzhou City University, Data Science — Sep. 2020 – Dec. 2024

- Overall GPA: 3.83/4.00, Ranking: 2/41
- Zhejiang Provincial Government Scholarship, 2024

Publications & Manuscripts

Efficient ML: Architectures and Systems

Synthetic Languages as a Testbed for Lightweight Language Model Design

Jiaqi Mo*, John Cooper*, Frederic Sala
Under review at **ACL Rolling Review** (NAACL 2027)

The Effect of Layer Ordering in Hybrid Architectures

John Cooper*, Mingchen Ma*, ***Jiaqi Mo****, Shengqi Qiu*, Nathaniel Bastian, Ilias Diakonikolas, Frederic Sala
Under review at **ICLR 2027**

FORGE: Procedural Weight Generation for Sparse MoE Inference

Chanyoung Moon, Travis Pence, Daisuke Yamada, ***Jiaqi Mo***, Yanxin Zhang, Vikas Singh, Karthikeyan Sankaralingam
ASPLOS 2027

Fast Sandwich Products in Clifford Algebra

Travis Pence, Daisuke Yamada, ***Jiaqi Mo***, Chanyoung Moon, Karthikeyan Sankaralingam, Vikas Singh
NeurIPS 2026

KernelCascade: Stateful Model Cascading for Cost-Effective Kernel Optimization

Jiaqi Mo*, Yichi Zhang*, Mingyuan Wu, Minjia Zhang
Under review at **MLSys 2027**

FleetServe: A Framework for Fleet-Scale Robot Intelligence in the Cloud

Hongyu Yao, Ziwen Zeng, ***Jiaqi Mo***, Minjia Zhang
Under review at **MLSys 2027**

Earlier and Other Work

MamFusion: Multi-Mamba with Temporal Fusion for Partially Relevant Video Retrieval

Xinru Ying*, ***Jiaqi Mo****, Jingyang Lin, Canghong Jin, Fangfang Wang, Lina Wei†
ICME 2025 (Oral)

Mamba-Based Multimodal Continual Learning for Audio-Visual Classification with Prototype-Enhanced Anti-Forgetting Mechanism

Jingyang Lin*, Xinru Ying*, ***Jiaqi Mo***, Lina Wei†, Fangfang Wang, Canghong Jin, Guanlin Chen
ACM MM 2025 Workshop (MMGR)

OS-Omni: A Cross-Platform Benchmark for Generalist Computer-Using Agents

Hui Shen, Yunta Hsieh, Jianing Ma, Ziyuan Liu, Qi Han, . . . , ***Jiaqi Mo***, . . . , Qiushi Sun, Mi Zhang, Ping Luo, Wenhu Chen, Ngai Wong
NeurIPS 2026 Evaluations and Datasets Track

MathGen: Revealing the Illusion of Mathematical Competence through Text-to-Image Generation

Ruiyao Liu, Hui Shen, Ping Zhang, Yunta Hsieh, Yifan Zhang, . . . , ***Jiaqi Mo***, . . . , Jing Xiong, Xin Wang, Ziyuan Liu, Hangrui Cao, Ngai Wong

ECCV 2026 Workshop (MUCG); under review at **TMLR**

Research Experience

Research Assistant, Advisor: Prof. Frederic Sala — May 2026 – Present
Sprocket Lab, **University of Wisconsin-Madison**

- Conducted systematic evaluations of synthetic languages as low-cost proxies for model architecture selection, demonstrating a Kendall's τ of 0.67 with English-trained models using a 400$\times$ smaller vocabulary.
- Ran architecture sweeps of Transformer–Mamba hybrids across multiple tasks, showing that small changes in layer ordering produce large performance gaps and that the best hybrid architecture depends on the task.

Research Assistant, Advisor: Prof. Minjia Zhang — May 2026 – Present
Supercomputing System and AI Lab, **University of Illinois Urbana-Champaign**

- Developed KernelCascade, a master–apprentice cascade for LLM-based kernel optimization that reuses and refines kernels from a stronger model, retaining most of its performance at a fraction of the optimization cost.
- Designing the robot–cloud protocol, per-robot network emulation, and a deadline-aware serving gateway for FleetServe, a framework for fleet-scale cloud robot control.

Research Assistant, Advisor: Prof. Vikas Singh — Oct. 2025 – May 2026
Vertical Research Group, **University of Wisconsin-Madison**

- Evaluated rotor-based layers from Clifford algebra as parameter-efficient replacements for the linear layers of dense and Mixture-of-Experts LLMs.
- Conducted compression and recovery fine-tuning experiments on models up to 80B parameters, evaluating generation-quality degradation under aggressive compression and recovery through fine-tuning.

Research Assistant, Advisor: Prof. Lina Wei — Apr. 2024 – Dec. 2024
Vision Multimodal Lab, **Hangzhou City University**

- Co-designed MamFusion, a video retrieval framework that integrates Mamba state-space modules to efficiently locate the moments relevant to a text query in long, untrimmed videos.
- Developed bidirectional text-to-video and video-to-text temporal fusion to strengthen cross-modal alignment, achieving state-of-the-art results on standard benchmarks.

Teaching Experience

University of Wisconsin-Madison, WI, USA
Peer Mentor — Sep. 2026 – Present
CS 540: Introduction to Artificial Intelligence

Skills

Programming Languages: Python, C, SQL, Java, MATLAB
Tools and Frameworks: PyTorch, Git, Docker, LaTeX, SGLang