Jeff (Junze) **Ma**

2025 Huron Pkwy, Ann Arbor, MI 48104, United States

(917) 753-0823 | jeffjma@umich.edu | jeff.junzema.com | majunze2001 | junze-ma

Summary

I'm a second-year Ph.D. student in CSE at the University of Michigan, advised by **Prof. Mosharaf Chowdhury.** I build **efficient software systems** for **Generative AI**, with a recent focus on Any-to-Any models and LLM agents.

Education

University of Michigan — *Ann Arbor, U.S.A*
Ph.D. Student in Computer Science and Engineering — *Aug. 2024 - Present*

New York University — *New York, U.S.A*
B.A. in Computer Science (Honors), Minor in Mathematics and Web Development and Applications — *Sept. 2021 - May. 2024*

- **GPA**: 4.0/4.0; **Honors**:Summa Cum Laude, Dean's List (2021-2024), Presidential Honors Scholars

Publications & Preprints

- **Jeff J. Ma***, Jae-Won Chung*, Akshay Jajoo, Myungjin Lee, Mosharaf Chowdhury. **Cornserve: Efficiently Serving Any-to-Any Multimodal Models.** *Under submission*
- Jae-Won Chung, **Jeff J. Ma**, Ruofan Wu, Jiachen Liu, Oh Jun Kweon, Yuxuan Xia, Zhiyu Wu, Mosharaf Chowdhury. **The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization.** *NeurIPS 2025 Datasets & Benchmarks (Spotlight)*
- Runyu Lu, Shiqi He, Wenxuan Tan, Shenggui Li, Ruofan Wu, **Jeff J. Ma**, Ang Chen, Mosharaf Chowdhury. **TetriServe: Efficient DiT Serving for Heterogeneous Image Generation** *Preprint 2025*
- Zachary DeStefano, **Jeff J. Ma**, Joseph Bonneau, Michael Walfish. **NOPE: Strengthening domain authentication with succinct proofs.** *SOSP 2024.*

Research Experience

Any-to-Any Model Serving — *Ann Arbor, U.S.A*
Graduate Student Research Assistant, SymbioticLab, University of Michigan — *Aug. 2024 - Present*

- **Cornserve**: an efficient online serving system for generic *Any-to-Any models*. It allows developers to describe a model as a computation graph and applies an automated planner with an Mixed-Integer Linear Programming solver to generate an optimized disaggregation and colocation deployment plan for the model.

GenAI Energy Efficiency — *Ann Arbor, U.S.A*
Graduate Student Research Assistant, SymbioticLab, University of Michigan — *Jun. 2025 - Present*

- **ML.ENERGY Benchmark**: a benchmark suite and tool that measures inference energy consumption under realistic service environments and performs automated energy optimization recommendations.

Zero-Knowledge Domain Proving — *New York, U.S.A*
Research Assistant, Courant Institute of Mathematical Sciences, New York University — *Aug. 2023 - May 2024*

- **NOPE**: a new mechanism for server authentication that uses Zero-Knowledge Proofs to verify domain ownership efficiently and securely. It improves security and reliability, and reduces reliance on Certificate Authorities while enabling compatibility with existing TLS infrastructure.

GPU Memory Disaggregation — *New Haven, U.S.A*
Research Assistant, Department of Computer Science, Yale University — *May 2023 - Apr. 2024*

- Proposed a system architecture that attaches a remote memory pool to GPUs to mitigate the GPU memory capacity for datacenter workload.
- Designed and implemented new GPU page fault mechanism in NVIDIA UVM driver; constructed kernel modules paired with RDMA daemons.

Teaching

Spring 2024	**CSCI-UA.0480 Computer Networks**	*NYU*
Fall 2023	**CSCI-UA.0202 Operating Systems**	*NYU*
Fall 2022	**CSCI-UA.0102 Data Structures**	*NYU*

Skills

Programming Languages: Python, C, JavaScript, Java, Bash

Tools and Frameworks: PyTorch, vLLM, Transformers, Kubernetes & K3s, Docker, OpenTelemetry, UCX, RDMA