

Haoran Ni

haoran.ni@smail.nju.edu.cn | haoran_ni@seas.harvard.edu
+86 139 1207 2569 | +1 510 345 9405

Education

Nanjing University

B.S. in Computer Science and Technology

- GPA: 4.60/5.00; weighted average: 92.0/100; major rank: 8/197.

- Selected coursework: Algorithms 97, Data Structures 98, Operating Systems 98, Computer Systems Laboratory 97, Rust Programming 99, Cryptography 99, and Graphics 100.

University of California, Berkeley

Summer Exchange Student, University-Level Exchange Program

- GPA: 4.00/4.00.

Nanjing, China

Expected Jun 2027

Berkeley, CA, USA

Jun 2025 – Aug 2025

Research Experience

Harvard John A. Paulson School of Engineering and Applied Sciences

Research Intern, Harvard MadSys Lab — ML Systems

Advisor: Prof. Juncheng Yang

- Design routing algorithms across multiple upstream providers to optimize cost and latency.
- Analyze ML artifact-storage workloads on Hugging Face and design a delivery network to improve model-download serving.
- Help develop and maintain a model-inference serving system with 100+ DAU and 10B daily token generation.

Institute of Computer Software, Nanjing University

Research Intern, SPAR Group — Operating Systems

Advisor: Prof. Yanyan Jiang

- Designed and built a system for competitive-programming contest environments.
- Deployed the system in NOIP, CSP, and LMCC competitions, serving 5,000+ participants; its technical approach was adopted as a reference by the NOI Technical Committee and secured purchase interest from key high schools in Jiangsu and Hubei.

Cambridge, MA, USA

Mar 2026 – Present

Nanjing, China

Sep 2024 – Feb 2026

Publications

RouteWise: Latency–Cost Optimization for Multi-Provider LLM Routing

Muxin Tian*, **Haoran Ni***, Yiyang Zhai, Yangsun Park, and Juncheng Yang

Accepted, EuroSys'27 (Co-first author)*

LatentStore: An Efficient Latent-First Storage System for AI-Generated Images

Zirui Wang, Yunjia Zheng, Tingfeng Lan, Zhaoyuan Su, **Haoran Ni**, Juncheng Yang, and Yue Cheng

Accepted, NSDI'27

To Store or To Regenerate? A Cost Model for AI-Generated Content at Scale

Yunjia Zheng, Zirui Wang, **Haoran Ni**, Tingfeng Lan, Zhaoyuan Su, Yue Cheng, and Juncheng Yang

Under review, SIGMETRICS'27

AutoX: Discovering Deployable Lossless Codecs for Evolving AI Artifacts

Tingfeng Lan, **Haoran Ni**, Muhammad Safiullah, Zhaoyuan Su, Zirui Wang, Yunjia Zheng, Juncheng Yang, and Yue Cheng

Under review, FAST'27

Honors & Awards

National Scholarship

Dec 2025

National Student Computer System Capability Challenge (East China), Second Prize (3/44)

Dec 2024

Teaching & Mentoring Experience

School of Computer Science, Nanjing University

Peer Mentor

2025–2026

Languages

Chinese (native); English (fluent; TOEFL 111, speaking 27)