Last updated: February 1, 2026

Hao Zhang

HDSI 440
UC San Diego 9500 Gilman Dr.
La Jolla, CA 92093

Phone: (412) 499-1581
Email: haozhang@ucsd.com
Homepage: https://cse.ucsd.edu/~haozhang

Research Interest

I study the intersection area of machine learning and systems. I am equally interested in designing strong, efficient, and secure machine learning models and algorithms, and in building scalable, practical distributed systems that can support real-world machine learning workloads.

Positions

University of California, San Diego

Assistant Professor, July 2023 - Present.

Snowflake, Menlo Park

Software Engineer (20% part-time), November 2023 - Present.

LMNet.ai, San Francisco

Co-founder, May 2023 - November 2023, acquired by Snowflake.

University of California, Berkeley

Postdoctoral Researcher, with Ion Stoica, Jan 2021 - July 2023.

Petuum Inc, Pittsburgh

Research Scientist, Jan 2020 – Jan 2023.

Director of Scalable ML, December 2017 – Jan 2020.

Tech Lead, May 2017 – December 2017.

Consultant, Jul 2016 – May 2017.

Microsoft Research Asia, Beijing

Research Intern, July 2013 - January 2014.

Microsoft, Shanghai

SDE Intern, September 2012 - April 2013.

Education

School of Computer Science, Carnegie Mellon University
Ph.D. in Computer Science, with Eric Xing, 2014 - 2020
Dissertation: Machine Learning Parallelism Could Be Adaptive, Composable, and Automated.

Department of Computer Science and Engineering, Shanghai Jiao Tong University
M.S. in Computer Science, 2011 - 2014

School of Computer Science and Engineering, South China University of Technology
B.E. in Computer Science, the Elite Class of Computer Science, 2008 - 2011

Awards and Honors

Google ML and System Junior Faculty Award, 2025.

Nvidia DGX B200 Compute Award (4 across the entire US), 2025.

AMD MI350x Compute Award, 2025.

MIT Technology Review Innovators Under 35 (TR35), China List, 2025

Google Research Award (three times), 2024, 2025.

#14 and #57 Most Cited AI Papers in 2023, by Zeta-alpha

Jay Lepreau Best Paper Award, OSDI 2021.

NVIDIA Pioneer Research Award, NeurIPS 2017.

Excellent Graduates (top 5%), Shanghai Jiao Tong University, 2014.

Scholarship for Graduates, Shanghai Jiao Tong University, 2011 - 2014.

Google Excellence Scholarship, Google Inc., 2013.

Early Graduate Honor (top 1%), South China University of Technology, 2011.

Excellent Undergraduates, South China University of Technology, 2008 - 2011.

1^{st} Class Scholarship (top 10%), South China University of Technology, 2008 - 2011.

Teaching

Instructor. DSC204A: Scalable Data Systems, UC San Diego, Fall 2025.

Instructor. CSE234 & DSC 291: Machine Learning Systems, UC San Diego, Winter 2025.

Instructor. DSC291: Machine Learning Systems, UC San Diego, Spring 2024.

Instructor. DSC204A: Scalable Data Systems, UC San Diego, Winter 2024.

TA. 10-708: Probabilistic Graphical Models, Carnegie Mellon University, Spring 2019.

TA. 16-791: Applied Data Science, Spring 2019.

TA. 10-701: Introduction to Machine Learning, Carnegie Mellon University, Fall 2015

Publications

See Google Scholar Profile for more details: 25821 citations as of February 1, 2026, with an h-index of 40.

Professional Service and Leadership

Organizer. The first FastVideo Meetup @ NeurIPS'25; the seventh vLLM Meetup @ Snowflake.

Department services. HDSI Masters Admissions Committee; HDSI Computational Resource Planning & Governance Committee; HDSI Undergraduate Program Committee; HDSI Industry Liaison Committee; HDSI Faculty Hiring Committees (twice); Hosted 10 faculty candidate interviews.

Founder and lead faculty. UCSD MLSYS Cohort.

Cofounder and advisor. LMSYS Org (Non-profit).

Organizer. UC Berkeley RISE Camp 2021, 2022.

Area Chair or Program Committee. ICLR, AAAI, UAI, ICML, NeurIPS, MLSYS, ATC, ASPLOS, COLM.

Reviewer. ICLR, NeurIPS, ACL, ECCV, AISTATS, ICML, NACCL, CVPR, ICCV, TPAMI, SCIS, IET Computer Vision, MVAP, TCC, VLDB, etc.

Volunteer. ICML, KDD, ATC, NeurIPS.

Funding

Raised approximately $1.5M or worth (including personal research funding, compute awards from NVIDIA, AMD, and Google etc.) from 2023 - 2025.

Tutorial and Invited Talks

See My lab's website for a list of my most recent talks. In the recent 3 years (2022 - 2025), I have given nearly **30 invited talks or tutorials** in universities, companies, industry venues, and major academic conferences. A partial list is below:

Tutorial, *Generating Video from Noise.* Nvidia Research Radar Talk Series.

Invited Talk, *The Future of AI Inference.* Nvidia Dynamo Day.

Invited Talk, *Fast Video Generation with Sliding Tile Attention.* Microsoft Research ACE Talk Series 2025.

Invited Talk, *DistServe: Disaggregating Prefill and Decoding for Goodput-optimized LLM Inference.* PyTorch Webinar 2024.

Invited Talk, *Lessons Learned from Running Chatbot Arena for 1 Year.* NSF Open-source Generated AI (OSGAI) Workshop.

Tutorial, *Welcome to the "Big Model" Era: Techniques and Systems to Train and Serve Bigger Models*. ICML 2022, with Zhuohan Li, Lianmin Zheng, and Ion Stoica.

Tutorial, *Simple and Automatic Distributed Machine Learning on Ray*. KDD 2021, with Zhuohan Li, Lianmin Zheng, and Ion Stoica.

Tutorial, *Simplifying and Automating Parallel Machine Learning via a Programmable and Composable Parallel ML System*. AAAI 2021, with Aurick Qiao, Qirong Ho and Eric Xing.

PhD Students

Junda Chen, PhD CS, UCSD (co-advised with Tajana Rosing; 2023 - present)

Yichao Fu, PhD CS, UCSD (2024 - present)

Lanxiang Hu, PhD ECE, UCSD (co-advised with Tajana Rosing; 2023 - present)

Mingjia Huo, PhD ECE, UCSD (co-advised with Tajana Rosing; 2025 - present)

Will Lin, PhD CS (2023 - present)

David Su, PhD Data Science (2025 - present)

Peiyuan Zhang, PhD CS (2024 - present)

Junli Wang, PhD CS (co-advised with Prithviraj (Raj) Ammanabrolu; 2025 - present)