50 Vassar Street
Cambridge, MA 02139
☎ *+1 857-209-1101*
✉ *qinghao@mit.edu*
🌐 *tonyhao.xyz*

Qinghao Hu

Research Interests

My research focuses on building **efficient and scalable machine learning systems**. Specifically, I develop full-stack infrastructure that pushes the efficiency frontier across the foundation-model lifecycle, spanning datacenter scheduling, large-scale pre-training, post-training with reinforcement learning, and model serving. My work emphasizes **algorithm–system co-design** for emerging workloads (long-context, multimodal, reasoning, agentic), and extends to broader system scenarios (storage, networking, robotics).

Experience

2024.08–Present **Massachusetts Institute of Technology**
- Postdoc in Department of Electrical Engineering and Computer Science
- Advisor: Prof. Song Han

2024.02–2024.04 **ETH Zürich**
- Academic Visitor in ETH System Group
- Advisor: Prof. Ana Klimovic

2024.01–2024.08 **Nanyang Technological University**
- Research Assistant Professor

Education

2020–2023 **Nanyang Technological University**
- Ph.D. in Computer Science
- Advisor: Prof. Tianwei Zhang and Prof. Yonggang Wen

2018–2020 **National University of Singapore**
- S.M. in Electrical and Computer Engineering
- Advisor: Prof. Kelvin Fong

2014–2018 **Zhejiang University**
- B.Eng. in Electronic Information Science and Technology
- Advisor: Prof. Wenyan Yin

Selected Awards & Honors

2023	**Google Ph.D. Fellowship** (*Only-ever* recipient from Singapore in Systems & Networking)
2023	**Distinguished Paper Award** of ASPLOS ’23
2023	**Best Paper Award** of WAIC ’23
2024	**Rising Star** in ML and Systems
2024	**Best Ph.D. Thesis Award** of Nanyang Technological University (College of Computing)
2024	**National Scholarship** for Outstanding International Graduates
2018	**Best Undergraduate Thesis Award** of Zhejiang University
2018	**Outstanding Graduate** of Zhejiang University

Publications

* : Equal Contribution, † : System Lead, ‡ : My Mentee

Conference & Journal Papers

ASPLOS '26 **Taming the Long-Tail: Efficient Reasoning RL Training with Adaptive Drafter**
Qinghao Hu*, Shang Yang*, Junxian Guo, Xiaozhe Yao, Yujun Lin, Yuxian Gu, Han Cai, Chuang Gan, Ana Klimovic, Song Han
Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2026

EuroSys '26 **Zeppelin: Balancing Variable-length Workloads in Data Parallel Large Model Training**
Chang Chen, Tiancheng Chen, Jiangfei Duan, Qianchao Zhu, Zerui Wang, **Qinghao Hu**, Peng Sun, Xiuhong Li, Chao Yang, Torsten Hoefler
European Conference on Computer Systems (EuroSys), 2026

NeurIPS '25 **Jet-Nemotron: Efficient Language Model with Post Neural Architecture Search**
Yuxian Gu, **Qinghao Hu**†, Shang Yang, Haocheng Xi, Junyu Chen, Song Han, Han Cai
Neural Information Processing Systems (NeurIPS), 2025

NeurIPS '25 **Scaling up Reasoning to Long Videos in VLMs**
Yukang Chen, Wei Huang, Baifeng Shi, **Qinghao Hu**†, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, Song Han
Neural Information Processing Systems (NeurIPS), 2025

SOSP '25 **Sailor: Automating Distributed Training over Dynamic, Heterogeneous, and Geo-distributed Clusters**
Foteini Strati, Zhendong Zhang, George Manos, Ixeia Sánchez Périz, **Qinghao Hu**, Tiancheng Chen, Berk Buzcu, Song Han, Pamela Delgado, Ana Klimovic
ACM Symposium on Operating Systems Principles (SOSP), 2025

MLSys '25 **LServe: Efficient Long-sequence LLM Serving with Unified Sparse Attention**
Shang Yang*, Junxian Guo*, Haotian Tang, **Qinghao Hu**, Guangxuan Xiao, Jiaming Tang, Yujun Lin, Zhijian Liu, Yao Lu, Song Han
Conference on Machine Learning and Systems (MLSys), 2025

ICLR '25 **LongVILA: Scaling Long-Context Visual Language Models for Long Videos**
Yukang Chen*, Fuzhao Xue*, Dacheng Li*, **Qinghao Hu***†, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, Song Han
International Conference on Learning Representations (ICLR), 2025

EuroSys '25 **DeltaServe: Multi-Tenant Language Model Serving via Delta Compression**
Xiaozhe Yao, **Qinghao Hu**, Ana Klimovic
European Conference on Computer Systems (EuroSys), 2025

NSDI '24 **Characterization of Large Language Model Development in the Datacenter**
Qinghao Hu*, Zhisheng Ye*, Zerui Wang*, Guoteng Wang, Meng Zhang, Qiaoling Chen, Peng Sun, Dahua Lin, Xiaolin Wang, Yingwei Luo, Yonggang Wen, Tianwei Zhang
USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2024
🏅 Invited Submission to USENIX ;login:

CSUR '24 **Deep Learning Workload Scheduling in GPU Datacenters: A Survey**
Zhisheng Ye*, Wei Gao*, **Qinghao Hu***, Peng Sun, Xiaolin Wang, Yingwei Luo, Tianwei Zhang, Yonggang Wen
ACM Computing Surveys (CSUR), 2024

SC '24 **TorchGT: A Holistic System for Large-scale Graph Transformer Training**
Meng Zhang*‡, Jie Sun*, **Qinghao Hu**, Peng Sun, Zeke Wang, Yonggang Wen, Tianwei Zhang
Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2024

ICDE '24 **Sylvie: 3D-adaptive and Universal System for Large-scale Graph Neural Network Training**
Meng Zhang‡, **Qinghao Hu**, Cheng Wan, Haozhao Wang, Peng Sun, Yonggang Wen, Tianwei Zhang
IEEE International Conference on Data Engineering (ICDE), 2024

WWW '24 **FedDSE: Distribution-aware Sub-model Extraction for Federated Learning over Resource-constrained Devices**
Haozhao Wang, Yabo Jia, Meng Zhang, **Qinghao Hu**, Hao Ren, Peng Sun, Yonggang Wen, Tianwei Zhang
The Web Conference (WWW), 2024

IWQoS '24 **Lins: Reducing Communication Overhead of ZeRO for Efficient LLM Training**
Qiaoling Chen‡, **Qinghao Hu**, Guoteng Wang, Yingtong Xiong, Ting Huang, Xun Chen, Yang Gao, Hang Yan, Yonggang Wen, Tianwei Zhang, Peng Sun
International Symposium on Quality of Service (IWQoS), 2024

OSDI '23 **Hydro: Surrogate-Based Hyperparameter Tuning Service in Datacenters**
Qinghao Hu, Zhisheng Ye, Meng Zhang, Qiaoling Chen, Peng Sun, Yonggang Wen, Tianwei Zhang
USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2023

ASPLOS '23 **Lucid: A Non-Intrusive, Scalable and Interpretable Scheduler for Deep Learning Training Jobs**
Qinghao Hu*, Meng Zhang*, Peng Sun, Yonggang Wen, Tianwei Zhang
Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2023
🎖 Distinguished Paper Award

ATC '22 **Primo: Practical Learning-Augmented Systems with Interpretable Models**
Qinghao Hu, Harsha Nori, Peng Sun, Yonggang Wen, Tianwei Zhang
USENIX Annual Technical Conference (ATC), 2022

SC '21 **Characterization and Prediction of Deep Learning Workloads in Large-Scale GPU Datacenters**
Qinghao Hu, Peng Sun, Shengen Yan, Yonggang Wen, Tianwei Zhang
Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2021
🎖 Top-3 Cited Paper in SC '21 (as of December 2025)

In Submission

Preprint **Universal and Efficient Load Balancing for RL Training of Large Multimodal Models**
Zerui Wang‡, **Qinghao Hu**, Jiecheng Zhou‡, Chang Chen, Haojie Duanmu, Xingcheng Zhang, Peng Sun, Dahua Lin

Preprint **Plan, Imagine, then Act: Steering Your VLA with Efficient Visually Grounded Planning**
Zhuoyang Zhang, Shang Yang, **Qinghao Hu**, Luke J. Huang, James Hou, Yufei Sun, Yao Lu, Song Han

Preprint **LLMFabric: Heterogeneous and Decentralized Large-Scale Machine Learning Serving System**
Xiaozhe Yao, Youhe Jiang, Ilia Badanin, **Qinghao Hu**, Binhang Yuan, Imanol Schlag, Eiko Yoneki, Ana Klimovic

Preprint **RL in the Wild: Characterizing RLVR Training in LLM Deployment**
Jiecheng Zhou‡, **Qinghao Hu**, Yuyang Jin, Zerui Wang, Peng Sun, Yuzhe Gu, Wenwei Zhang, Mingshu Zhai, Xingcheng Zhang, Weiming Zhang

Preprint **Semantic-Aware Scheduling for GPU Clusters with Large Language Models**
Zerui Wang*‡, **Qinghao Hu***, Ana Klimovic, Tianwei Zhang, Yonggang Wen, Peng Sun, Dahua Lin

Preprint **LoongTrain: Efficient Training of Long-Sequence LLMs with Head-Context Parallelism**
Diandian Gu, Peng Sun, **Qinghao Hu**, Ting Huang, Xun Chen, Yingtong Xiong, Guoteng Wang, Qiaoling Chen, Shangchun Zhao, Jiarui Fang, Yonggang Wen, Tianwei Zhang, Xin Jin, Xuanzhe Liu

Preprint **An Empirical Study of LLM Serving in Confidential GPUs**
Eunseong Park, Timo Thans, Vishnu Kumar Kalidasan, **Qinghao Hu**, Wenjie Xiong

Preprint **Efficient Training of Large Language Models on Distributed Infrastructures: A Survey**
Jiangfei Duan*, Shuo Zhang*, Zerui Wang*‡, Lijuan Jiang, Wenwen Qu, **Qinghao Hu**, Guoteng Wang, Qizhen Weng, Hang Yan, Xingcheng Zhang, Xipeng Qiu, Dahua Lin, Yonggang Wen, Xin Jin, Tianwei Zhang, Peng Sun

Preprint **InternEvo: Efficient Long-Sequence Large Language Model Training via Hybrid Parallelism and Redundant Sharding**
Qiaoling Chen‡, Diandian Gu, Guoteng Wang, Xun Chen, Yingtong Xiong, Ting Huang, **Qinghao Hu**, Xin Jin, Yonggang Wen, Tianwei Zhang, Peng Sun

Professional Service

Workshop Organizer

CVPR '25	Workshop: Efficient Large Vision Models Workshop (ELVM)
MICRO '24	Workshop: Hardware and Architectural Support for Security and Privacy (HASP)

Conference Reviewer

ICLR '26	International Conference on Learning Representations
CVPR '26	Conference on Computer Vision and Pattern Recognition
ICLR '25	International Conference on Learning Representations
EuroSys '25	EuroSys Conference – Shadow Committee Member
EuroSys '24	EuroSys Conference – Shadow Committee Member
EuroSys '23	EuroSys Conference – Shadow Committee Member
OSDI '22	USENIX Operating Systems Design and Implementation – AE Committee Member
ATC '22	USENIX Annual Technical Conference – AE Committee Member
EuroSys '22	EuroSys Conference – AE Committee Member

SOSP '21	Symposium on Operating Systems Principles – AE Committee Member

Journal Reviewer

TPDS	IEEE Transactions on Parallel and Distributed Systems
TACO	ACM Transactions on Architecture and Code Optimization
TOCS	ACM Transactions on Computer Systems
CSUR	ACM Computing Surveys