HANQING ZHU

Graduate Research Assistant ⋄ ECE Department ⋄ University of Texas at Austin
hqzhu@utexas.edu ⋄ (512)200-6791 ⋄ Personal website ⋄ Google scholar

RESEARCH INTERESTS

My research centers on efficient AI computing, from emerging AI systems to hardware-/system-aware AI algorithms.

- Hardware-Software Co-Design for Emerging AI Hardware and Computing Systems
- Efficient Training and Inference Techniques for Vision Models and Large-Scale Foundation Models

EDUCATION

The University of Texas at Austin (UT-Austin), TX, USA *Aug. 2020 - Dec. 2025 (Expected)*
Ph.D. Candidate, Department of Electrical and Computer Engineering
Advisor: David Z. Pan; Co-advisor: Ray T. Chen
First year (2020-2021) conducted part-time in China due to COVID-19
Honors: ECE Graduate Achievement Award; Graduate School Continuing Fellowship Nominee
GPA: 3.93/4.00

Shanghai Jiao Tong University (SJTU), Shanghai, China *Sept. 2016 - Jun. 2020*
B.E., Microelectronics Science and Engineering
Graduated with Highest Honors
Rank: 2^{nd}/57; GPA: 3.81/4.00

HONORS AND AWARDS

Texas ECE Graduate Achievement Award	UT Austin	2024
UT Graduate School Continuing Fellowship Nomination (1 of 2 nominees in the entire ECE department)	UT Austin	2024
1st Place in IEEE/ACM MLCAD FPGA Macro-Placement Contest	MLCAD	2023
MLSys Student Travel Award	MLSys	2023
Winner of Robert S. Hilbert Memorial Optical Design Competition	Synopsys	2022
DAC Young Fellow	DAC	2021
Shanghai Outstanding Graduate	Shanghai City	2020
Departmental Excellent Undergraduate Thesis	SJTU	2020
Hongyi Scholarship	SJTU	2019
Outstanding Undergraduate Scholarship	SJTU	2019
Samsung Scholarship	SJTU	2018
Zhiyuan College Honors Scholarship	SJTU	2018
1st Prize, National Mathematical Contest in Modeling	Shanghai Division	2018
Academic Excellence Scholarship	SJTU	2017-2019

PROFESSIONAL EXPERIENCE

Meta AI, CA, USA *May 2024 – Oct 2024*
Research Scientist Intern, Efficient Large-scale Training
Mentor: Dr. Jinwon Lee

- Memory-efficient training techniques for large language models; submitted to MLSys. [P2]
- Communication-efficient methods for large-scale ads model training.

Lightelligence Inc., MA, USA *May 2023 – Sept 2023*
Software Research Intern, Low-bit Noise-aware Training for Photonic AI Chips
Mentor: Dr. Weifeng Zhang

- Low-precision noise-aware training for state-of-the-art photonic AI accelerators.

Google Brain, CA, USA *Jul 2022 – Nov 2022*
Student Researcher, Google Brain, RL-based Chip Placement
Mentor: Dr. Joe Jiang

- Chip placement with reinforcement learning. Integrate and tune DREAMPlace for the RL chip placer.

SELECTED RESEARCH PROJECTS

I published papers across top conferences in design automation, computer architecture, and machine learning, e.g., DAC, ICCAD, TCAD, HPCA, Neurips, ICCV.

Efficient ML Training and Inference, including:

- Memory-efficient LLM pre-training optimization method, APOLLO, which **matches or even outperforms Adam in perplexity** while delivering **superior memory savings compared to Galore**[P2].
  - APOLLO **first** enables LLM pre-training with low-rank auxiliary optimizer states **without SVD**.
  - APOLLO-mini achieves negligible optimizer state overhead, requiring only **rank-1 space**.
  - Achieve $> 3\times$ pre-training throughput of LLaMA-7B compared to Adam on a 4×A100-80GB setup.
- Efficient sparse training on-chip/on-device tailored for self-learnable AI hardware[C4].

Emerging Hardware and Accelerators for AI, including:

- *First-of-its-kind* photonic Transformer accelerator with circuit-architecture co-design. [C17]

HW-SW Co-design and Optimization for Efficient and Reliable AI Systems, including:

- Circuit/system-aware quantization and compression strategies for CNNs and Transformers. [C1, C4, C5, C9, C15, C17, J2]
- AI-driven simulation for photonic device design. [C18]

INVITED TALKS

- Invited talk at Lightelligence, 2023
  - "Towards Reliable and Self-Learnable Photonic Neural Network from the Lens of Software-Hardware Co-design"

PROFESSIONAL SERVICE

- **Conference Reviewer:** ICML, NeurIPS, ICLR, AAAI, DAC, ICCAD, FPGA, AICAS
- **Journal Reviewer:** TNNLS, TCAD, Photonic Network Communications

MENTORING & TEACHING & VOLUNTEER EXPERIENCES

- Mentor for senior undergraduates' capstone project 2023
- TA at EE316: Digital Logic Design Fall 2022
- Conference Volunteer, the IEEE International Symposium on Circuits and Systems (ISCAS) 2022
- Volunteer teacher at Eryuan No.2 high school, Yunnan, China Aug. 2017- Sept. 2017

PUBLICATIONS (* DENOTED CO-FIRST AUTHOR)

Preprint Papers

[P2] **Hanqing Zhu***, Zhenyu Zhang*, Wenyan Cong, Xi Liu, Sem Park, Vikas Chandra, Bo Long, David Z. Pan, Zhangyang Wang, Jinwon Lee. "APOLLO: Greater Memory Savings than GaLore, Superior Performance to AdamW." under submission to MLsys 2025, will arxiv and open-source soon.

[P1] Chen, Guojin, Keren Zhu, Seunggeun Kim, Hanqing Zhu, Yao Lai, Bei Yu, and David Z. Pan. "LLM-Enhanced Bayesian Optimization for Efficient Analog Layout Constraint Generation." arXiv preprint arXiv:2406.05250 (2024).

Conference Papers

[C18] **Hanqing Zhu**, Wenyan Cong, Guojin Chen, Shupeng Ning, Ray Chen, Jiaqi Gu, and David Z. Pan, "PACE: Pacing Operator Learning to Accurate Optical Field Simulation for Complicated Photonic Devices," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2024 [paper; code;]

[C17] **Hanqing Zhu**, Jiaqi Gu, Hanrui Wang, Zixuan Jiang, Zhekai Zhang, Rongxin Tang, Chenghao Feng, Song Han, Ray T. Chen, David Z. Pan, " Lightening-Transformer: A Dynamically-operated Optically-interconnected Photonic Transformer Accelerator," in *IEEE International Symposium on High Performance Computer Architecture(HPCA)*, Mar. 2024 (Acceptance Rate: 18.3% (75 of 410)) [paper; code;]

[C16] Zixuan Jiang, Jiaqi Gu, **Hanqing Zhu**, and David Z. Pan, "Pre-RMSNorm and Pre-CRMSNorm Transformers: Equivalent and Efficient Pre-LN Transformers," in *Conference on Neural Information Processing Systems (NeurIPS)*, Dec 10 - Dec 16, 2023 **(Spotlight)**. (Acceptance Rate: 26.1%)

[C15] **Hanqing Zhu**, Jiaqi Gu, Hanrui Wang, Rongxin Tang, Zhekai Zhang, Chenghao Feng, Song Han, Ray T. Chen, David Z. Pan, " DOTA: A Dynamically-Operated Photonic Tensor Core for Energy-Efficient Transformer Accelerator," in *Conference on Machine Learning and Systems (MLSys), Workshop on Systems for Next-Gen AI Paradigms (SNAP)*, Jun 4 - Jun 8, 2023

[C14] Jiaqi Gu, Chenghao Feng, **Hanqing Zhu**, David Z. Pan, and Ray T. Chen, "Light-AI Interaction: The Convergence of Photonic AI and Cross-layer Circuit-Architecture-Algorithm Co-design," in *Conference on Machine Learning and Systems (MLSys), Workshop on Systems for Next-Gen AI Paradigms (SNAP)*, Jun 4 - Jun 8, 2023

[C13] Jiaqi Gu, Chenghao Feng, **Hanqing Zhu**, David Z. Pan, and Ray T. Chen, "Light-AI Interaction: The Convergence of Photonic AI and Cross-layer Circuit-Architecture-Algorithm Co-design," in *SPIE Photonics West*, Jan., 2023

[C12] Chenghao Feng, Rongxing Tang, Jiaqi Gu, **Hanqing Zhu**, David Z. Pan, and Ray T. Chen, "Optically Interconnected, Hardware-Efficient, Electronic-Photonic Neural Network using Compact Multi-Operand Photonic Devices," in *SPIE Photonics West*, Jan., 2023

[C11] Jiaqi Gu, Zhengqi Gao, Chenghao Feng, **Hanqing Zhu**, Ray Chen, Duane S Boning, and David Z. Pan, " NeurOLight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation," in *Conference on Neural Information Processing Systems (NeurIPS)*, Nov 26 - Dec 4, 2022. **(Spotlight)**

[C10] Harrison Jin, **Hanqing Zhu**, Keren Zhu, Thomas Leonard, Mahshid Alamdar, David Z. Pan, and Jean Anne C. Incorvia, "Design of Domain Wall-Magnetic Tunnel Junction Analog Content Addressable Memory using Current and Projected Prototype Data," in *Annual Conference on Magnetism and Magnetic Materials (MMM)*, Minneapolis, MN, October 31 - November 4, 2022.

[C9] **Hanqing Zhu**, Keren Zhu, Jiaqi Gu, Harrison Jin, Ray Chen, Jean Anne Incorvia and David Z. Pan, "Fuse and Mix: MACAM-Enabled Analog Activation for Energy-Efficient Neural Acceleration" in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Oct., 2022

[C8] Chenghao Feng, Jiaqi Gu, **Hanqing Zhu**, Zhoufeng Ying, Zheng Zhao, David Z. Pan, and Ray T. Chen, "Optoelectronically Interconnected Hardware-Efficient Deep Learning using Silicon Photonic Chips," in *Smart Photonic and Optoelectronic Integrated Circuits (SPIE)*, Mar., 2022

[C7] Chenghao Feng, Jiaqi Gu, **Hanqing Zhu**, David Z. Pan, and Ray T. Chen, "Design and Experimental Demonstration of A Hardware-Efficient Integrated Optical Neural Network," in *Smart Photonic and Optoelectronic Integrated Circuits (SPIE)*, Mar., 2022

[C6] Jiaqi Gu, **Hanqing Zhu**, Chenghao Feng, Zixuan Jiang, Mingjie Liu, Shuhan Zhang, Ray T. Chen, and David Z. Pan, "ADEPT: Automatic Differentiable DEsign of Photonic Tensor Cores," in *ACM/IEEE Design Automation Conference (DAC)*, Jul., 2022

[C5] **Hanqing Zhu**, Jiaqi Gu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "ELight: Enabling Efficient Photonic In-Memory Neurocomputing with Life Enhancement," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan., 2022.

[C4] Jiaqi Gu, **Hanqing Zhu**, Chenghao Feng, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "L2ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-situ Subspace Optimization," in *Conference on Neural Information Processing Systems (NeurIPS)*, Dec., 2021.

[C3] Jiaqi Gu, **Hanqing Zhu**, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "Towards Memory-Efficient Neural Networks via Multi-Level in situ Generation," in *International Conference on Computer Vision (ICCV)*, Oct., 2021.

[C2] Chenghao Feng, Jiaqi Gu, **Hanqing Zhu**, David Z. Pan, and Ray T. Chen, "Experimental Demonstration of a WDM-based Integrated Optical Decoder for Compact Optical Computing," in *Conference on Lasers and Electro-Optics*, May, 2021.

[C1] Jiaqi Gu, Zheng Zhao, Chenghao Feng, **Hanqing Zhu**, Ray T. Chen, and David Z. Pan, "ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-bit Controls," in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Mar., 2020.

Journal Papers

[J8] Shupeng Ning*, **Hanqing Zhu***, Chenghao Feng, Jiaqi Gu, Zhixing Jiang, Zhoufeng Ying, Jason Midkiff, Sourabh Jain, May H. Hlaing, David Z. Pan, and Ray T. Chen, "Photonic-Electronic Integrated Circuits for High-Performance Computing and AI Accelerator," in IEEE Journal of Lightwave Technology (JLT), July. 2024.

[J7] Feng, Chenghao, Jiaqi Gu, **Hanqing Zhu**, Shupeng Ning, Rongxing Tang, May Hlaing, Jason Midkiff, Sourabh Jain, David Z. Pan, and Ray T. Chen. "Integrated multi-operand optical neurons for scalable and hardware-efficient deep learning." Nanophotonics 13, no. 12 (2024): 2193-2206.

[J6] Jiaqi Gu, **Hanqing Zhu**, Chenghao Feng, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "M3ICRO: Machine Learning-Enabled Compact Photonic Tensor Core based on PRogrammable Multi-Operand Multimode Interference," in *APL Machine Learning*, Jan. 2024.

[J5] Harrison Jin, **Hanqing Zhu**, Keren Zhu, Thomas Leonard, Jaesuk Kwon, Mahshid Alamdar, Kwangseok Kim, Jungsik Park, Naoki Hase, David Z. Pan, Jean Anne C. Incorvia, "Domain Wall-Magnetic Tunnel Junction Analog Content Addressable Memory Using Current and Projected Data" in *IEEE Transactions on Nanotechnology*, 2024.

[J4] Chenghao Feng*, Jiaqi Gu*, **Hanqing Zhu**, Zhoufeng Ying, Zheng Zhao, David Z. Pan, and Ray T. Chen, "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," in *ACS Photonics*, 2022.

[J3] Jiaqi Gu, Chenghao Feng, **Hanqing Zhu**, Zheng Zhao, Zhoufeng Ying, Mingjie Liu, Ray T. Chen and David Z. Pan, "SqueezeLight: A Multi-Operand Ring-Based Optical Neural Network with Cross-Layer Scalability," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jul., 2022.

[J2] **Hanqing Zhu**, Jiaqi Gu, Chenghao Feng, Mingjie Liu, Zixuan Jiang, Ray T. Chen, and David Z. Pan, "ELight: Towards Efficient and Aging-Resilient Photonic In-Memory Neurocomputing," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, Jun., 2022.

[J1] Jiaqi Gu, Chenghao Feng, **Hanqing Zhu**, Ray T. Chen and David Z. Pan, "Light in AI: Toward Efficient Neurocomputing with Optical Neural Networks - A Tutorial," in *IEEE Transactions on Circuits and Systems–II: Express Briefs (TCAS-II)*, Apr., 2022.