YIFAN (EVELYN) GONG

(she / her / hers) ◇ Research Scientist/Engineer ◇ Adobe Research, Seattle, WA 98103

gong.yifa@northeastern.edu Personal Website Yifan Gong

EDUCATION

Northeastern University Boston, MA
Ph.D. in Computer Engineering, advised by Prof. Yanzhi Wang Sep 2019 – Aug 2024

- With a focus on **energy-efficient deep learning and artificial intelligence systems, accelerations of deep neural networks including large-scale models for GenAI, trustworthy machine learning**

University of Toronto Toronto, ON, Canada
Master of Applied Science (Thesis-based with Fellowship), advised by Prof. Baochun Li Sep 2017 – Sep 2019

- With a focus on deep reinforcement learning and its applications

Xidian University Xi'an, Shaanxi, China
Bachelor of Engineering (**Valedictorian**, **with highest honor**), GPA: 3.83/4.0 (rank $\mathbf{1}^{st}$) Sep 2013 – Jun 2017
Education Experimental Class **(Undergraduate honor program)**

RESEARCH INTERESTS

- Accelerations of Emerging Large-Scale Models for GenAI such as Diffusion Models and LLMs
- Hardware and Software Co-Design for Artificial Intelligence Acceleration
- Energy-Efficient Deep Learning and Artificial Intelligence Systems
- Model Compression
- Scalable and Robust Deep Learning

SELECTED SCHOLARSHIP, HONORS, AND AWARDS

MIT EECS Rising Star	08/2024
First Place in **DAC Ph.D. Forum**	06/2024
ML and Systems Rising Star	05/2024
CVPR Travel Grant	04/2024
DAC Young Fellow	04/2024
ICCAD Student Scholar Award	09/2023
College of Engineering **Outstanding Graduate Student in Teaching** of Northeastern University	04/2023
College of Engineering **Dean's Fellowship** of Northeastern University	2019-2020
ECE Student Fellowship of University of Toronto	2017-2019
Valedictorian of Xidian University	06/2017
Excellent Graduate of Xidian University (**10 of 5180**)	06/2017
Goodix Scholarship for Science and Technology	12/2016
National Scholarship (1%)	10/2015, 10/2016
Special Prize in National English Competition for College Students **(0.1%)**	05/2016
Role Model Outstanding Student	11/2014, 11/2015
Provincial $\mathbf{1}^{st}$ Prize in CUMCM	11/2015

PUBLICATIONS

Submitted, † means equal contribution.

[I3] Xuan Shen, Peiyan Dong, Zhenglun Kong, **Yifan Gong**, et al, "Qualiquant: Qualified Quantization-Aware Training for Mobile Language Models", *under review*.

[I2] Changdi Yang, Zheng Zhan, **Yifan Gong**, et al, "FairSMOE: Mitigating Multi-Attribute Fairness Problem with Sparse Mixture-of-Experts", *under review*.

[I1] Zhenglun Kong, Zheng Zhan, **Yifan Gong**, et al, "Fusion-X: Advancing LLM Ability with Adaptive Heterogeneous Model Integration", *under review*.

Conference Proceedings, † means equal contribution.

[C26] Xuan Shen, Hangyu Zheng, **Yifan Gong**, Zhenglun Kong, Changdi Yang, Zheng Zhan, Yushu Wu, Xue Lin, Yanzhi Wang, Pu Zhao, Wei Niu, "Sparse Learning for State Space Models on Mobile", in **ICLR** 2025.

[C25] Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Yanyu Li, **Yifan Gong**, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, Zhihao Shu, Wei Niu, Pu Zhao, Yanzhi Wang, Jiuxiang Gu, "Lazydit: Lazy Learning for the Acceleration of Diffusion Transformers", in **AAAI** 2025.

[C24] Yuguang Yao†, Jiancheng Liu†, **Yifan Gong**†, Xiaoming Liu, Yanzhi Wang, Xue Lin, Sijia Liu, "Can Adversarial Examples Be Parsed to Reveal Victim Model Information?", in **WACV** 2025.

[C23] **Yifan Gong**, Zheng Zhan, Yanyu Li, Yerlan Idelbayev, Andrey Zharkov, Kfir Aberman, Sergey Tulyakov, Yanzhi Wang, Jian Ren, "Efficient Training with Denoised Model Weights", in **ECCV** 2024.

[C22] **Yifan Gong**, Zheng Zhan, Qing Jin, Yanyu Li, Yerlan Idelbayev, Xian Liu, Andrey Zharkov, Kfir Aberman, Sergey Tulyakov, Yanzhi Wang, Jian Ren, "E^2GAN: Efficient Training of Efficient GANs for Image-to-Image Translation", in **ICML** 2024.

[C21] **Yifan Gong**, Yushu Wu, Zheng Zhan, Pu Zhao, Liangkai Liu, Chao Wu, Xulong Tang, Yanzhi Wang, "Lotus: learning-based online thermal and latency variation management for two-stage detectors on edge devices", in **DAC** 2024.

[C20] **Yifan Gong**†, Yushu Wu†, Zheng Zhan, Geng Yuan, Yanyu Li, Qi Wang, Chao Wu, Yanzhi Wang, "MOC: Multi-Objective Mobile CPU-GPU Co-optimization for Power-efficient DNN Inference", in **ICCAD** 2023.

[C19] **Yifan Gong**, Pu Zhao, Zheng Zhan, Yushu Wu, Chao Wu, Zhenglun Kong, Minghai Qin, Caiwen Ding, Yanzhi Wang, "Condense: A Framework for Device and Frequency Adaptive Neural Network Models on the Edge", in **DAC** 2023.

[C18] **Yifan Gong**, Zheng Zhan, Pu Zhao, Yushu Wu, Chao Wu, Caiwen Ding, Weiwen Jiang, Minghai Qin, Yanzhi Wang, "All-in-One: A Highly Representative DNN Pruning Framework for Edge Devices with Dynamic Power Management", in **ICCAD** 2022.

[C17] **Yifan Gong**, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, Sijia Liu, "Reverse Engineering of Imperceptible Adversarial Image Perturbations", in **ICLR** 2022.

[C16] **Yifan Gong**, Zheng Zhan, Zhengang Li, Wei Niu, Xiaolong Ma, Wenhao Wang, Bin Ren, Caiwen Ding, Xue Lin, Xiaolin Xu, Yanzhi Wang, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", in **GLSVLSI** (invited) 2020.

[C15] **Yifan Gong**, Baochun Li, Ben Liang, Zheng Zhan, "Chic: Experience-driven Scheduling in Machine Learning Clusters", in **IWQoS** 2019.

[C14] Chao Wu†, **Yifan Gong**†, Liangkai Liu†, Mengquan Li, Yushu Wu, Xuan Shen, Zhimin Li, Geng Yuan, Weisong Shi and Yanzhi Wang, "AyE-Edge: Automated Deployment Space Search Empowering Accuracy yet Efficient Real-Time Object Detection on the Edge", in **ICCAD** 2024.

[C13] Zheng Zhan†, Zhenglun Kong†, **Yifan Gong**†, Yushu Wu, Zichong Meng, Hangyu Zheng, Xuan Shen, Stratis Ioannidis, Wei Niu, Pu Zhao, Yanzhi Wang, "Exploring Token Pruning in Vision State Space Models", in **NeurIPS** 2024.

[C12] Yushu Wu†, **Yifan Gong**†, Pu Zhao, Yanyu Li, Zheng Zhan, Wei Niu, Hao Tang, Minghai Qin, Bin Ren, Yanzhi Wang, "Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution", in **ECCV** 2022.

[C11] Zheng Zhan†, **Yifan Gong**†, Pu Zhao, Geng Yuan, Wei Niu, Yushu Wu, Tianyun Zhang, Malith Jayaweera, David Kaeli, Bin Ren, Xue Lin, Yanzhi Wang, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search", in **ICCV** 2021.

[C10] Xuan Shen, Pu Zhao, **Yifan Gong**, Zhenglun Kong, Zheng Zhan, Yushu Wu, Ming Lin, Chao Wu, Xue Lin, Yanzhi Wang, "Search for Efficient Large Language Models", in **NeurIPS** 2024.

[C9] Zheng Zhan, Yushu Wu, **Yifan Gong**, Zichong Meng, Zhenglun Kong, Changdi Yang, Geng Yuan, Pu Zhao, Wei Niu, Yanzhi Wang, "Fast and Memory-Efficient Video Diffusion Using Streamlined Inference", in **NeurIPS** 2024.

[C8] Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, **Yifan Gong**, Xuan Shen, Xue Lin, Pu Zhao, Yanzhi Wang, "Rethinking Token Reduction for State Space Models", in **EMNLP Main** 2024.

[C7] Peiyan Dong, Zhenglun Kong, Xin Meng, Pinrui Yu, **Yifan Gong**, Geng Yuan, Hao Tang, Yanzhi Wang, "HotBEV: Hardware-oriented Transformer-based Multi-View 3D Detector for BEV Perception", in **NeurIPS** 2023.

[C6] Zifeng Wang, Zheng Zhan, **Yifan Gong**, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy, "DualHSIC: HSIC-Bottleneck and Alignment for Continual Learning", in **ICML** 2023.

[C5] Sizhe Chen, Geng Yuan, Xinwen Cheng, **Yifan Gong**, Minghai Qin, Yanzhi Wang, Xiaolin Huang, "Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors", in **ICLR** 2023.

[C4] Xiaolong Ma, Geng Yuan, Zhengang Li, **Yifan Gong**, Tianyun Zhang, Wei Niu, Zheng Zhan, Pu Zhao, Ning Liu, Jian Tang, Xue Lin, Bin Ren, Yanzhi Wang, "Blcr: Towards Real-Time DNN Execution with Block-based Reweighted Pruning", in **ISQED** 2022.

[C3] Zifeng Wang, Zheng Zhan, **Yifan Gong**, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy, "Sparcl: Sparse continual learning on the edge", in **NeurIPS** 2022.

[C2] Geng Yuan, Xiaolong Ma, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, **Yifan Gong**, Zheng Zhan, Chaoyang He, Qing Jin, Siyue Wang, Minghai Qin, Bin Ren, Yanzhi Wang, Sijia Liu, Xue Lin, "Mest: Accurate and fast memory-economic sparse training framework on the edge", in **NeurIPS** 2021.

[C1] Peiyan Dong, Siyue Wang, Wei Niu, Chengming Zhang, Sheng Lin, Zhengang Li, **Yifan Gong**, Bin Ren, Xue Lin, Dingwen Tao, "RTMobile: Beyond Real-Time Mobile Acceleration of RNNs for Speech Recognition", in **DAC** 2020.

Book Chapters

[B1] "Machine Learning Systems for High Performance and Dependability: The Role of Hardware Design", Springer.

Journal Papers

[J3] Yuguang Yao, Xiao Guo, Vishal Asnani, **Yifan Gong**, Jiancheng Liu, Xue Lin, Xiaoming Liu, Sijia Liu, "Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks", Foundations and Trends® (**FnT**) in Privacy and Security, 2024.

[J2] **Yifan Gong**, Geng Yuan, Zheng Zhan, Wei Niu, Zhengang Li, Pu Zhao, Yuxuan Cai, Sijia Liu, Bin Ren, Xue Lin, Xulong Tang, Yanzhi Wang, "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time

Mobile Acceleration", ACM Transactions on Design Automation of Electronic Systems (**TODAES**), 2021.

[J1] Tong Jian, **Yifan Gong**, Zheng Zhan, Runbin Shi, Nasim Soltani, Zifeng Wang, Jennifer Dy, Kaushik Chowdhury, Yanzhi Wang, Stratis Ioannidis, "Radio Frequency Fingerprinting on the Edge", IEEE Transactions on Mobile Computing (**TMC**), 2021.

Patents

[P1] Yanzhi Wang, **Yifan Gong**, Zheng Zhan, "Computer-implemented methods and systems for privacy-preserving deep neural network model compression", US Patent App. 17/176,340, 2021.

PROFESSIONAL EXPERIENCES

Adobe Research Seattle, WA
Research Scientist/Engineer Aug 2024 – present

- *Work on efficient generative models and foundation model optimizations*
  Manager: Dr. Kevin Wampler

Snap Inc. Santa Monica, CA
Ph.D. Research Intern @ Creative Vision Group May 2023 – Aug 2023

- *Work on efficient distillation of GANs from diffusion models and building of a weight generator with diffusion process to predict GAN weights*
  Mentor: Dr. Jian Ren, Dr. Sergey Tulyakov

IBM Research Cambridge, MA
Ph.D. Research Intern @ MIT-IBM Watson AI Lab May 2021 – Aug 2021

- *Work on improving vision transformers by attention graph*
  Mentor: Dr. Quanfu Fan

TEACHING EXPERIENCES

Teaching Assistant of Advances in Deep Learning @ Northeastern University

- Taught a 2-hour course about how to use SOTA deep learning frameworks such as PyTorch and TensorFlow
- Prepared course materials and final project topics
- Held office hours to address the questions of students in the class
- Assigned and graded homework, quiz, and final projects

Teaching Assistant of Operating System and Computer Fundamentals @ University of Toronto

- Gave lab demonstrations to students
- Marked and graded tests and exams

PARTICIPATING GRANTS

NSF #1937500 Sep 2019 - Sep 2023
Primary student representative for Prof. Yanzhi Wang's group
"RTML: Large: Efficient and Adaptive Real-Time Learning for Next Generation Wireless Systems" - $1,199,000
PI: Stratis Ioannidis, Jennifer Dy, Kaushik Chowdhury, Tommaso Melodia, Yanzhi Wang

NSF #1909172 Sep 2019 - Sep 2024
Primary student representative for Prof. Yanzhi Wang's group
"CNS Core: Small: Collaborative: Content-Based Viewport Prediction Framework for Live Virtual Reality Streaming" - $171,229
PI: Yanzhi Wang

DARPA Reverse Engineering of Deceptions (RED) Oct 2020 - Oct 2024
Primary student representative from NEU
"Intelligent Diagnosis for Machine and Human-Centric Adversaries" - $1,000,000

PI: Sijia Liu, Xiaoming Liu, Xue Lin

NSF # 2312158 Oct 2023 - Sep 2026
Primary student representative for Prof. Yanzhi Wang's group
"Collaborative Research: CSR: Small: Expediting Continual Online Learning on Edge Platforms through Software-Hardware Co-designs" - $250,000
PI: Yanzhi Wang

MENTORING EXPERIENCES

Anh-Dung Dinh *CS PhD @ University of Sydney, summer intern @ Adobe* Summer 2024
Project: Anycost Diffusion Model with Dynamic Capacity

Sizhe Chen *Master @ Shanghai Jiao Tong University, now a CS PhD @ UC Berkeley* Fall 2022
Publication: Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors (ICLR-23)

Xuan Shen *ECE PhD @ Northeastern University* 2023-2024
Publication: Search for Efficient Large Language Models (NeurIPS-24)

Yushu Wu *ECE PhD @ Northeastern University* 2022-2024
Publication: Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution (ECCV-22), MOC: Multi-Objective Mobile CPU-GPU Co-Optimization for Power-Efficient DNN Inference (ICCAD-23), Fast and Memory-Efficient Video Diffusion Using Streamlined Inference (NeurIPS-24)

INVITED TALKS

Tutorials

[T4] "Reverse Engineering of Deceptions: Foundations and Applications", @ **CVPR'23**.

Invited Seminars

[T3] "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration", in ROAD4NN @ **DAC'21**.

[T2] "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", @ **GLSVLSI'20**.

[T1] "Towards Best Possible Deep Learning Acceleration on the Edge - A Compression-Compilation Co-Design Framework", in MGHPCC @ **SC'20**.

PROFESSIONAL SERVICES

Conference Reviewer: ICLR, NeurIPS, ICML, ICCV, ECCV, CVPR, AAAI, ISCAS, AICAS, AdvML

Journal Reviewer: TCAD, TODAES

Mentorship: OurCS Workshop for Undergraduates in Computer Science at CMU

REFERENCES

Dr. Yanzhi Wang
Associate Professor, Faculty Fellow
Northeastern University
yanz.wang@northeastern.edu

Dr. David R. Kaeli
COE Distinguished Professor
Northeastern University
kaeli@ece.neu.edu

Dr. Stratis Ioannidis
Professor
Northeastern University
ioannidis@ece.neu.edu

Dr. Xiaoming Liu
MSU Foundation Professor
Michigan State University
liuxm@cse.msu.edu

Dr. Jennifer Dy
Professor, EAI director
Northeastern University
jdy@ece.northeastern.edu

Dr. Baochun Li
Professor, Associate Chair
University of Toronto
bli@ece.toronto.edu