Zhen Dong

https://dong-zhen.com | Google Scholar

RESEARCH INTERESTS

Efficient AI: Efficient MLLM, VideoGen & Embodied AI
LLM post-training, model compression, AI systems
Function-calling agents and multi-agent systems
Hardware-software co-design and AI for science
Efficient evaluation and alignment of foundation models

EDUCATION

Postdoc in CS at UC Berkeley, Advisor: Prof. Kurt Keutzer — *Oct. 2022 – Jun. 2023*
Ph.D. in CS at UC Berkeley, Advisor: Prof. Kurt Keutzer — *Aug. 2018 – Oct. 2022*
B.S. in EECS at Peking University (Rank 1st/327) — *Sept. 2014 – July 2018*

AWARDS

Winner of 2018-2020 Berkeley Fellowship
Winner of PhD Forum (2nd Place among all candidates) at DAC 2024
Doctoral Consortium at CVPR 2024
Best Paper Nomination at Practical DL Workshop at AAAI 2023
1st Place on EMCC 2020 Competition on Classification Track and Object Detection Track
2nd Place on Visual Wake Word Competition at CVPR 2019
AWS Research Credits Award 2021 and Google Cloud Research Credits Award 2022
1st Place Research Funding Proposal at Berkeley Deep Drive (BDD) 2019
Winner of SenseTime Scholarship in 2018
1st Prize in the National Olympiad in Physics (China), 1st Prize in National Physics Competition for College Students
Winner of Tang Lixin Scholarship (Highest honor for undergraduate academic and research excellence in China)
Winner of Tang Lixin 1st Prize Scholarship, Winner of Fang Zheng Scholarship
Princeton University Math Competition (PUMac) Top three among all participants in geometry group
Top Ten Undergraduate Research Award at PKU EECS
Outstanding Graduates at Peking University and Outstanding Graduates in Beijing

RESEARCH EXPERIENCE

Tenure-Track Assistant Professor, **UCSB** — *Starting Fall 2026*
Senior/Staff Research Scientist, **NVIDIA** — *Jun. 2025 - present*
Founding Member, **Nexusflow.AI** — *Jun. 2023 – Jun. 2025*
Research on Large Language Models (LLMs) with Function Calls: NexusRaven, NaxusRaven-V2, Athene-V2

- Propose novel methods to construct structured data and conduct instructional finetuning for NexusRaven models.
- Publicize benchmarks and a leaderboard that can measure the function calling abilities of LLMs.
- NexusRaven-V2 is around 7% better than GPT-4 on function calling, with only 13B parameters.
- NexusRaven-V2-13B gained 461 likes and **200k+ total downloads** on Huggingface. It ranked Top-5 on Huggingface Trending when released.

Ph.D./Postdoc, **Berkeley AI Research** (BAIR), UC Berkeley — *Dec. 2018 – Jun. 2023*
Advisor: Prof. Kurt Keutzer
Research on Hessian-aware Quantization, Pruning & Distillation: HAWQ (ICCV'19), HAWQ-V2 (NeurIPS'20), ZeroQ (CVPR'20), HAP (WACV'22), Quantization Review (BLPCV'22), QD-BEV (ICCV'23), NoisyQuant (CVPR'23)

- Propose a Hessian-based method to decide mixed-precision configuration and block-wise fine-tuning order.
- Prove theorem to use the trace of Hessian as sensitivity metric and conduct fast Pareto frontier optimization.
- Generalize to segmentation, 2D/3D object detection tasks and achieve state-of-the-art results.
- Conduct fast end-to-end quantization without fine-tuning and without using any training/test data.
- This line of work has obtained **3686 citations** to date.

Research on HW-SW Co-design:

HAWQ-V3 (ICML'21), CoDeNet (FPGA'21), HAO (FCCM'21), CSQ (DAC'23), EPIM (DAC'24)

Achieve hardware-aware quantization and utilize 4-bit Tensor Cores for inference acceleration.

Implement 4-bit kernels and mixed-precision support on TVM, achieve 7.4x compression ratio and 5.4x speedup compared to fp32.

Propose efficient deformable operations on embedded FPGAs and design new FPGA-core with ultra-low precision arithmetic.

HW-SW joint architecture search and efficient implementation of mixed-precision NNs on CPU/GPU/FPGAs/PIM (Processing-in-Memory).

This line of work has obtained **564 citations** to date.

Research on Efficient LLMs and Diffusion Models:

Q-BERT (AAAI'20), Q-Diffusion (ICCV'23), SqueezeLLM (ICML'24), PB-LLM (ICLR'24)

Propose sensitivity-based non-uniform quantization and dense-and-sparse decomposition for efficient handling of outliers.

Pioneer the usage of Hessian information to guide LLM quantization in both post-training quantization (PTQ) and quantization-aware training.

Implement 3/4-bit CUDA kernels and achieve 4.6x compression ratio compared to fp16 and 2.4x speedup when deployed on an A6000 GPU.

Propose timestep-aware calibration and split shortcut quantization to achieve 4-bit diffusion models at the first time.

This line of work has obtained **1292 citations** to date.

Research on Multi-agent Systems: MAgIC (EMNLP'24)

Pioneer the integration of probabilistic graphical modeling (PGM) to enhance the cognitive abilities of LLMs and obtain better interpretability.

Present a framework to evaluate LLM-powered multi-agent systems by employing social deduction games alongside game theory scenarios.

Research on Image & Video Generative Models:

PromptCoT (CVPR'24), ViewControl (IJCAI'24), D-Edit (AAAI'25), Meissonic, Magic-Me, VEditBench, KSort Arena

Propose new methods to achieve better control ability of generative diffusion models.

Develop novel efficient Arena algorithms for human-in-the-loop evaluation and alignment, and collect benchmarks for less costly auto-eval.

Present Meissonic-1B that elevates masked image modeling (MIM) text-to-image models to a level comparable to SDXL.

Research on AI for Science:

FastML (Frontiers in Big Data'22), High-momentum Particle Trigger Decisions (TRETS'24)

Review AI inference acceleration techniques and how they help dark matter search, morphology characterization, synthesis dynamics, etc.

Implement efficient AI on ASICs and FPGAs to reduce time cost and enable particle trigger decisions at CERN Large Hadron Collider (LHC).

Research Intern, **Bytedance AI** *Jan. 2023 – Apr. 2023*

Research Intern, **Nvidia AI** *Jun. 2021 – Sept. 2021*

Research Intern, **Facebook AI** *Jun. 2020 – Aug. 2020*

Research Intern, **SenseTime AI** *Apr. 2018 – Aug. 2018*

Undergraduate Visiting Researcher (UGVR), Electrical Engineering, **Stanford University** *Jun. 2017 – Sept. 2017*

Advisor: Prof. H.-S. Philip Wong

Research Assistant, Electrical Engineering and Computer Sciences, **Peking University** *Dec. 2016 – Jun. 2018*

Advisor: Prof. Jinfeng Kang

INDUSTRIAL COLLABORATIONS

NVIDIA, Intel, Amazon, Alibaba, Panasonic, Bytedance, Google, Meta, Apple, AMD, Nexusflow, Samsung, Tesla

OPENSOURCE

2025: NVIDIA-Nemotron-Nano-V2[9B base][12B base][pre-training dataset], Llama-Nemotron-Super-V1.5[8-bit model][dataset], NexusBench, R-KV

2024: K-Sort Arena [huggngface], Meissonic [huggingface], AtheneV2-Chat-72B[huggingface], AtheneV2-Agent [huggingface], D-Edit [huggingface]

2023: NexusRaven-V2 [huggingface][demo][leaderboard], NexusRaven [huggingface], Magic-Me [website][demo], SqueezeLLM, Q-Diffusion

2022: HAP, LOVEU-TGVE, AwesomeQuantizationPapers

2021: BitPack, CoDeNet

2020: ZeroQ, HAWQ

CORRESPONDING-AUTHOR PUBLICATIONS

[1] Z. Li, X. Liu, D. Fu, J. Li, Q. Gu, K. Keutzer, **Zhen Dong**✉. "K-Sort Arena: Efficient and reliable benchmarking for generative models via

K-wise human preferences," **[CVPR 2025]**.

[2] Yinsheng Li, **Zhen Dong**✉, Yi Shao✉. "DrafterBench: Benchmarking Large Language Models for Tasks Automation in Civil Engineering," arXiv 2025.

[3] Y. Shang, Z. Yuan, Q. Wu, **Zhen Dong**✉. "PB-LLM: Partially Binarized Large Language Models," **[ICLR 2024]**.

[4] L. Xu, Z. Hu, D. Zhou✉, H. Ren, **Zhen Dong**✉**,** et al. "MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration," **[EMNLP 2024]**.

[5] Z. Ma, D. Zhou, C. Yeh, X. Li, H. Yang, **Zhen Dong**✉**,** et al. "Magic-Me: Identity-Specific Video Customized Diffusion," arXiv 2024.

[6] R. Ma, Q. Zhou, Y. Jin, D. Zhou, B. Xiao, X. Li, Y. Qu, A. Singh, K. Keutzer, J. Hu, X. Xie, **Zhen Dong**✉**,** et al. "A Dataset and Benchmark for Copyright Protection from Text-to-Image Diffusion Models," arXiv 2024.

FIRST-AUTHOR PUBLICATIONS

[1] C. Wang*, **Zhen Dong*** ✉**,** et al. "EPIM: Efficient Processing-In-Memory Accelerators based on Epitome," **[DAC 2024]**.

[2] Y. Zhang*, **Zhen Dong***, et al. "QD-BEV: Quantization-aware View-guided Distillation for Multi-view 3D Object Detection," **[ICCV 2023]**.

[3] A. Gholami*, S. Kim*, **Zhen Dong*,** et al. "A Survey of Quantization Methods for Efficient Neural Network Inference", **[BLPCV 2022]** (Book of Low-Power Computer Vision).

[4] S. Yu*, Z. Yao*, A. Gholami*, **Zhen Dong***, et al. "Hessian-ßAware Pruning and Optimal Neural Implant," Oral, **[WACV 2022]**.

[5] Z. Yao*, **Zhen Dong***, et al. "HAWQV3: Dyadic Neural Network Quantization," **[ICML 2021]**.

[6] **Zhen Dong***, et al. "HAO: Hardware-aware neural Architecture Optimization for Efficient Inference," Oral, **[FCCM 2021]**.

[7] **Zhen Dong***, et al. "CoDeNet: Algorithm-hardware Co-design for Deformable Convolution," **[FPGA 2021]** Oral Presentation.

[8] **Zhen Dong**, et al. "HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks," **[NeurIPS 2020]**.

[9] S. Shen*, **Zhen Dong***, et al. "Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT," Spotlight, **[AAAI 2020]**.

[10] Y. Cai*, Z. Yao*, **Zhen Dong*,** et al. "ZeroQ: A Novel Zero Shot Quantization Framework," **[CVPR 2020]**.

[11] **Zhen Dong***, et al. "HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision," **[ICCV 2019]**.

[12] **Zhen Dong,** et al. "Ultra-low Bit Quantization for Visual Wake Word Challenge", 2nd Place at VWW Competition, **[CVPR 2019]**.

[13] **Zhen Dong**, et al. "Convolutional Neural Networks for Image Recognition and Online Learning Based on RRAM Devices," IEEE Transactions on Electron Devices **[TED 2018]**.

[14] **Zhen Dong**, et al. "RRAM-based Convolutional Neural Networks for High Accuracy Pattern Recognition," **[VLSI-SNW 2018]**, Oral Presentation.

[15] J. Kang***, Zhen Dong****, et al. China's patent about 3D RRAM.

SELECTED PUBLICATIONS (chronological order)

[1] "NVIDIA Nemotron Nano2: Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model," **[Tech Report 2025]**.

[2] J. Bai, T. Ye, W. Chow, E. Song, Q. Chen, X. Li, **Zhen Dong,** et al. "Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis," **[ICLR 2025]**.

[3] Zefan Cai, Wen Xiao, … **Zhen Dong,** Anima Anandkumar, Abedelkadir Asi, Junjie Hu. "R-KV: Redundancy-aware KV Cache Compression for Reasoning Models," **[arXiv 2025]**.

[4] J. Bai, **Zhen Dong,** et al. "Integrating View Conditions for Image Synthesis," **[IJCAI 2024]**.

[5] R. Zhang, Y. Luo, H. Yang, **Zhen Dong**, et al. "Efficient Deweahter Mixture-of-Experts with Uncertainty-Aware Feature-wise Linear Modulation," **[AAAI 2024]**.

[6] J. Yao, Y. Liu, **Zhen Dong,** et al. "PromptCoT: Align prompt distribution via adapted chain of thought," **[CVPR 2024]**.

[7] S. Kim, C. Hooper, A. Gholami, **Zhen Dong,** et al. "SqueezeLLM: Dense-and-Sparse Quantization," **[ICML 2024]**.

[8] A. Chen, H. Yang, Y. Gan, **Zhen Dong,** et al. "Split-Ensemble: Efficient OOD-aware ensemble via task and model splitting," **[ICML 2024]**.

[9] J. Campos, **Zhen Dong,** et al. "End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs," ACM Transactions on Reconfigurable Technology and Systems **[TRETS 2024]**.

[10] V. Srinivasan, **Zhen Dong,** et al. "NexusRaven: A Commercially-Permissive Language Model for Function Calling," **[FMDM@NeurIPS 2024]**.

[11] Z. Yuan, Y. Shang, Y. Zhou, **Zhen Dong,** et al. "LLM Inference Unveiled: Survey and Roofline Model Insights," arXiv 2024.

[12] J. Campos, **Zhen Dong,** et al. "End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs," OSCAR Workshop at **[ISCA 2023]**.

[13] L. Xiao, H. Yang, **Zhen Dong,** et al. "CSQ: Growing Mixed-Precision Quantization with Bi-level Continuous Sparsification," **[DAC 2023]**.

[14] Y. Liu, H. Yang, **Zhen Dong,** et al. "NoisyQuant: Noisy Bias-Enhanced Post-Training Activation Quantization for Vision Transformers,"

[CVPR 2023].

[15] X. Li, Y. Liu, L. Lian, H. Yang, **Zhen Dong,** et al. "Q-Diffusion: Quantizing Diffusion Models," **[ICCV 2023]**.

[16] M. Guo, **Zhen Dong,** et al. "SANA: Sensitivity-Aware Neural Architecture Adaptation for Uniform Quantization," **[Applied Sciences, 2023]**.

[17] T. Li, X. Chen, **Zhen Dong,** et al. "Domain-Adaptive Text Classification with Structured Knowledge from Unlabeled Data", Long Oral, **[IJCAI 2022]**.

[18] A. Deiana, …**Zhen Dong,** et al. "Applications and Techniques for Fast Machine Learning in Science", **[Frontiers in Big Data 2022]**.

[19] T. Li, X. Chen, S. Zhang, **Zhen Dong**, et al. "Cross-Domain Sentiment Classification with Contrastive Learning and Mutual Information Maximization," **[ICASSP 2021]**.

[20] P. Huang, Z. Li, **Zhen Dong**, et al. "Binary Resistive Switching Device Based Electronic Synapse with Spike-Rate-Dependent-Plasticity for Online Learning," ACS **[Applied Electronic Materials 2019]**.

[21] R. Han, P. Huang, Y. Xiang, C. Liu, **Zhen Dong**, et al. "A Novel Convolutional Computing Paradigm Based on NOR Flash Array with High Computing Speed and Energy Efficiency," published by IEEE Transactions on Circuits and Systems **[TCAS 2019]**.

[22] X. Wang, P. Huang, **Zhen Dong**, et al. "A Novel RRAM-based Adaptive-Threshold LIF Neuron Circuit for High Recognition Accuracy," published by International Symposium on VLSI Technology, Systems and Applications **[VLSI-TSA 2018]**.

[23] Z. Zhou, C. Liu, W. Shen, **Zhen Dong**, et al. "The Characteristics of Binary Spike-Time-Dependent Plasticity in HfO2-Based RRAM," Nanoscale Research Letters **[NRL 2018]**.

[24] P. Huang, D. B. Zhu, C. Liu, Z. Zhou, **Zhen Dong**, et al. "RTN based Oxygen Vacancy Probing Method for Ox-RRAM Reliability Characterization and Its Application in Tail Bits," published by International Electron Devices Meeting **[IEDM 2017]**.

ACADEMIC SERVICE

Program Committee Member: NeurIPS, ICML, CVPR, ICCV, EMNLP, ICLR, AAAI (Senior PC), ECCV, IJCAI, WACV, KDD, MLSys, TinyML, ECV, BLPCV

Reviewer for TPAMI (Transactions on Pattern Analysis and Machine Intelligence), TMLR (Transactions of Machine Learning Research), JMLR (Journal of Machine Learning Research), TNNLS (IEEE Transactions on Neural Networks and Learning Systems), IEEE Micro, TED (IEEE Transactions on Electron Devices), PR (Pattern Recognition), TCSVT (IEEE Transactions on Circuits and Systems for Video Technology), OJCAS (IEEE Open Journal of Circuits and Systems), JCST (Journal of Computer Science and Technology) and Fundamental Research (Elsevier)

TALKS & ORGANIZED WORKSHOPS & MEDIA

[1] NVIDIA-Nemotron-Nano-V2 8B is pretrained from scratch and can outperform Qwen3 8B. NVIDIA AI Official Post, NVIDIA ADLR Post, AI era Link to Post, Q-bit AI Link to Post.

[2] Llama-Nemotron-Super-V1.5 is on NVIDIA AI Official Post, Q-bit AI Link to Post. To date, ranked 1st of open-sourced LLMs on AA Index.

[3] I served as a panelist at the Global Green Development Summit (GGDS) 2025.

[4] R-KV gets recommended by Q-bit AI (量子位), Link to Post.

[5] Meissonic gets recommended by AI era (新智元) and 36Kr, Link to Post.

[6] KSort Arena gets recommended by Qingke Lab, Link to Post.

[7] I presented "Efficient Deep Learning via Quantization and Co-Design" at CVPR 2024 Doctoral Consortium and DAC 2024 PhD Forum.

[8] I co-organized the LOVEU (LOng-form VidEo Understanding) workshop at CVPR 2024.

[9] Q-Diffusion is featured in the newest TensorRT post.

[10] I co-organized the 3rd Workshop on Practical Deep Learning: Towards Efficient and Reliable LLMs at IEEE Conference on Artificial Intelligence (IEEE CAI) 2024.

[11] NexusRaven-V2-13B is presented at NeurIPS 2023 EXPO.

[12] NexusRaven and NexusRaven-V2 are recommended by: Deci AI Top 10 Under-13B LLMs, Huggingface's Post, Together AI's Post, etc.

[13] Invited Talk "Efficient Inference and Training of Large Neural Network Models" at Intel oneAPI DevSummit for AI and HPC 2023.

[14] Invited Talk "Hardware-Aware Efficient Deep Learning" at Peking University Institute of Artificial Intelligence (PKU-IAI), on June 11, 2023.

[15] I co-organized the LOVEU (LOng-form VidEo Understanding) workshop at CVPR 2023, Link to Zhihu.

[16] Invited to host the Practical DL Workshop at AAAI 2023 in Washington DC.

[17] Invited Talk "Efficient Deep Learning via Quantization and HW-SW Co-Design" at Hardware and Algorithms for Learning On-a-chip Workshop (HALO) at ICCAD 2022.

[18] My dissertation on "Hardware-aware Efficient Deep Learning" was defended on June 29, 2022.

[19] "Efficient Neural Networks through Systematic Quantization and Co-Design", virtually at Matchlab (Imperial College London), [slides].

[20] CoDeNet and HAO are presented at ML@B Seminar (Machine Learning at Berkeley).

[21] "Hessian-Aware Pruning and Optimal Neural Implant", WACV 2022, Hawaii, US, [slides].

[22] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2021, Berkeley, US.

[23] The book that I contributed to, "Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence", is online for ordering.

[24] "HAO: Hardware-aware neural Architecture Optimization for Efficient Inference", FCCM 2021 (online).

[25] "HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks", NeurIPS 2020.

[26] HAWQ-V2 gets recommended by JiangMen (将门) AI media (in Chinese), Link to ZhiHu.

[27] "Systematic Neural Network Quantization", NVIDIA GTC 2021.

[28] "Efficient Neural Networks through Systematic Quantization", BAIR/CPAR/BDD Seminar 2020, [slides].

[29] "HAWQ-V3: Dyadic Neural Network Quantization" is presented at TVM Conference 2020.

[30] "ZeroQ: A novel Zero-Shot Quantization Framework", Real-Time Intelligent Secure Explainable Systems (RISELab) Retreat 2020, Lake Tahoe (online), US, [slides].

[31] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2020, Santa Rosa, US.

[32] "Q-BERT: Hessian Based Quantization of BERT", AAAI 2020, New York, US, [slides].

[33] Q-BERT gets recommended by Synced (机器之心) AI media (in Chinese), Link to WeChat.

[34] Q-BERT gets recommended by AI.Science (Aggregate Intellect), Link to YouTube.

[35] "Hessian-Aware trace-Weighted Quantization", Beyond First-Order Methods in ML Workshop at NeurIPS 2019, Vancouver, Canada.

[36] Real-Time Intelligent Secure Explainable Systems (RISELab) Retreat 2019, Monterey, US.

[37] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2019, Berkeley, US.

[38] Visual Wake Word Challenge, LPIRC Workshop at CVPR 2019, Long Beach, US, [slides], [link].

[39] "RRAM Based Convolutional Neural Networks for High Accuracy Pattern Recognition and Online Learning Tasks", VLSI-SNW 2017, Kyoto, Japan, [slides].

TEACHING EXPERIENCE

Online Course of CS267 Parallel Computing on Moodle XSEDE: Course Coordinator
Applications of Parallel Computers, Berkeley CS 267: Head Graduate Student Instructor
Optimization Analytics, Berkeley INDENG 240: Graduate Student Instructor
Mathematical Programming, Berkeley INDENG 262A: Graduate Student Instructor
BAIR Mentoring Program for Underrepresented Undergraduates

GRANT & FUNDING WRITING

I have significantly contributed to the writing of the following grants and fundings:

[1] Canadian Research Council Grant 2025: Advanced Automation with MLLMs on Civil Engineering

[2] Intel Research Grant 2024: Efficient Deep Learning on Intel Processors and Networks

[3] Berkeley Deep Drive (BDD) 2023 Funding Proposal: Quantization on Vision Models for Real-time and Accurate Inference in ADAS/AV

[4] Panasonic Research Grant 2023: Controllable AI with LLM/VLM

[5] Intel Research Grant 2023: Efficient Distributed Training of Large-Scale Neural Networks

[6] Berkeley Deep Drive (BDD) 2022 Funding Proposal: Efficient Transformer Inference and Training for Fast Unsupervised Learning Through Attention-Aware Pruning

[7] Alibaba Berkeley Commons Grant 2022: DQRM: Deep Quantized Recommendation Models

[8] Panasonic Research Grant 2022: Sensitivity-aware DETR Quantization

[9] Alibaba Berkeley Commons Grant 2021: TASC: Topology-Aware Structured Communications for Efficient Deep Neural Network Training

[10] Berkeley Deep Drive (BDD) 2021 Funding Proposal: Real-time and Accurate Object Detection through Quantization of Transformer- and MLP-based Computer Vision Models

[11] Facebook Research Grant 2020: A Study of Communication Avoiding Algorithms for Training Large Scale Recommendation Systems

[12] Google Cloud Research Grant 2020: Hardware Software Co-Design for NLP and Recommendation Systems

[13] AWS Research Grant 2020: Hardware-aware Quantization with End-to-end Inference Acceleration

[14] Berkeley Deep Drive (BDD) 2020 Funding Proposal: Efficient Neural Networks Through Systematic Quantization

[15] Wave Computing Research Grant 2019: Model Compression of RoBERTa on NLP Tasks

[16] Google Cloud Research Grant 2019: Hessian-aware Mixed-Precision Quantization with Distillation

NOTABLE MENTORING EXPERIENCE

I had the privilege of mentoring many talented students over the years, listed below in chronological order.

1. Zhikai Li (2023-2025), PhD at CASIA, now Assistant Professor at CASIA
2.
3. Yaohui Cai (2019-2020), Undergrad at PKU, now PhD at Cornell
4. Daiyaan Arfeen (2019-2020), Undergrad at Berkeley, now PhD at CMU
5. Sheng Shen (2019-2020), MEng at Berkeley, then PhD at Berkeley, Meta Llama Team, xAI
6. Zhangcheng (Zach) Zheng (2020-2021), MEng at Berkeley, now at AWS
7. Eric Tan (2020-2021), MEng at Berkeley, now at Google
8. Yizhao Gao (2020-2022), Exchange student at Berkeley, now PhD at HKU
9. Lutfi Eren Erdogan (2021-2022), Undergrad at Berkeley, now at Narada.ai
10. Aishani Singh (2022-2023), The Harker School, now at CMU
11. Yijiang Liu (2022-2023), PhD at NJU, now Assistant Professor at NJU
12. Chenyu Wang (2022-2023), Undergrad at THU, now PhD at Princeton
13. Jinbin Bai (2022-2024), MS at NUS, now PhD at NUS
14. Yang Zhou (2022-2024), Visiting student at Berkeley, now PhD at CMU
15. Mingfei Guo (2022-2024), MS at Stanford, now at Nvidia
16. Vijay Anand Raghava Kanakagiri (2023-2024), Visiting student at Berkeley, now at Amazon

SKILLS

Programming languages: Python, C/C++, SQL, Matlab, Verilog

Platforms: PyTorch, JAX, Tensorflow(&Keras), MXNet, TensorRT, OpenMP/MPI/UPC++/CUDA, torch-CCL

Tools: Cadence, Gurobi, Xilinx Vivado & ISE, HSpice, Gradio, OriginLab, Modelsim