Ashkan Moradifirouzabadi

UC San Diego, 9500 Gilman Dr, La Jolla, CA, 92093 • ashkan@ucsd.edu

https://scholar.google.com/citations?user=i7vjNG4AAAAJ&hl=en
https://www.linkedin.com/in/ashkan-moradi-fir98/

RESEARCH INTERESTS

Hardware-Algorithm Co-Design, LLM Optimization, In-memory Computing, VLSI, Computer Architecture

EDUCATION

Jan 2022 - Present **Expected Graduation: 3/2027**	**University of California, San Diego** Ph.D. in Computer Engineering **(GPA: 4.00/4.00)** Advisor: Prof. Mingu Kang	**La Jolla, CA**
Jan 2022 - March 2024	**University of California, San Diego** M.S. in Computer Engineering **(GPA: 4.00/4.00)** Advisor: Prof. Mingu Kang	**La Jolla, CA**
Sep 2016 - Feb 2021	**University of Tehran** B.Sc. in Electrical Engineering **(Overall GPA: 3.73/4.00, Major GPA: 4.00/4.00)**	**Tehran, Iran**

EXPERIENCE

- **Technology Research Intern, Intel Corporation, Hillsboro, OR**
  Hardware Accelerator Design for Large Language Models — Jun 2024 - Sep 2024
  Manager: Ian Young
- **Research Assistant, Vertically-integrated VLSI Information Processing (VVIP) Lab, UC San Diego, La Jolla, CA**
  Developing Large-scale Hardware Accelerators for Deep Learning Models from Architecture and Algorithm Design to Chip Fabrication, PI: Mingu Kang — Jan 2022 - Present

PUBLICATIONS

* Equal Contribution

1. Priyansh Bhatnagar*, **A. Moradifirouzabadi***, S. Yang, S. Lee, J. Choi, and M. Kang, "STAR-KV: Low-Rank KV Cache Compression via Soft Thresholding for Adaptive Rank Control," in *Forty-third International Conference on Machine Learning (ICML)* (**Spotlight**), Seoul, South Korea, Jul. 2026.
2. **A. Moradifirouzabadi**, D. S. Dodla, and M. Kang, "An Analog and Digital Hybrid Attention Accelerator for Transformers with Charge-based In-memory Computing," in *IEEE 50th European Solid State Electronics Research Conference (ESSERC),* Bruges, Belgium, Sep. 2024.
3. **A. Moradifirouzabadi** and M. Kang, "HyAtt: A Hybrid Attention Accelerator for Transformers with Analog In-memory Computing," in *IEEE Journal of Solid-State Circuits (JSSC),* Sep. 2025.
4. A. Yazdanbakhsh*, **A. Moradifirouzabadi***, Z. Li*, and M. Kang, "Sparse attention acceleration with synergistic in-memory pruning and on-chip recomputation," in *Proc. of the 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Chicago, USA, Oct. 2022.
5. **A. Moradifirouzabadi***, Priyansh Bhatnagar*, et. al., "TAILOR: Token-Adaptive Low-Rank KV Cache Compression for PIM-Accelerated LLM Inference," Under Review at MICRO.
6. **A. Moradifirouzabadi**, M. Kang, "End-to-end Acceleration of Generative Models with Runtime Regularized KV Cache Management," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, Jun. 2025.
7. C. E. Song*, **A. Moradifirouzabadi***, T. Rosing, and M. Kang, "Efficient Transformer Accelerator via Reconfiguration for Encoder and Decoder Dual Modes with Sparsity-Aware Data Mapping," in *Proc. of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, Newport Beach, USA, Aug. 2024.
8. K. Fan, **A. Moradifirouzabadi**, et. al., "SpecPCM: A Low-power PCM-based In-Memory Computing Accelerator for Full-stack Mass Spectrometry Analysis," in *IEEE Journal on Exploratory Solid-State Computational Devices and Circuit (JXCDC)*, Nov. 2024.
9. S. Pinge, **A. Moradifirouzabadi**, et. al., "FeNOMS: Enhancing Open Modification Spectral Library Search with In-Storage Processing of Ferroelectric NAND (FeNAND) Flash," in *International Conference on Computer-Aided Design (ICCAD)*, Oct. 2025.

TECHNICAL SKILLS

Programming Languages: Python, C/C++, TCL, MATLAB

Machine Learning Frameworks: PyTorch

Hardware Development: Verilog, SystemVerilog, Cadence Innovus, Cadence Virtuoso, Synopsis Design Compiler, Cadence Xcelium, Xilinx Vivado, Altera Quartus, Modelsim, Altium Designer, PCB Artist

SELECTED COURSES

ECE 260B: VLSI Integrated Circuits & Systems Design (A)
ECE 260C: VLSI Advanced Topics (A)
ECE 226: Deep Learning Acceleration on Hardware (A^+)
CSE 251A: Principles of Machine Learning: Learning Algorithms (A)
ECE 284: VLSI Implementation for Machine Learning (A^+)
ECE 285: Intro to Visual Learning (A)
CSE 240A: Principles of Computer Architecture (A)
CSE 257: Search and Optimization (A)

HONORS AND AWARDS

- UCSD Electrical and Computer Engineering Department Ph.D. Fellowship, CA, USA, 2022.
- Finalist for The Best Undergraduate Project Award, Tehran, Iran, 2021.
- Ranked 263rd among 160,000+ Participants in The National University Entrance Exam for B.Sc., Tehran, Iran, 2016.
- IEEE/ACM International Symposium on Microarchitecture Student Travel Grant, IL, USA, 2022.

SELECTED PROJECTS

- **Adaptive Optimal Rank Selection for Low-rank KV cache Compression:** Compressing the KV cache through finding the optimal rank of key/value projections via soft thresholding, achieving 4x compression with minimal loss, and reaching 20c compression when combined with a low-rank-aware mixed-precision quantization. With custom Triton-based GPU kernels, we achieve 6.9x attention speedup and 3.1x end-to-end generation throughput (token/s) improvement compared to Pytorch SDPA. **(*ICML'25 Spotlight*)**
- **An Analog and Digital Hybrid Attention Accelerator for LLMs with Charge-based In-memory Computing (Fabricated Chip):** Design and implement a 9T-1C SRAM bitcell with analog in-memory computing capability in a 16Kb CIM cache. The chip also includes a digital core (~150K gates) for Attention related computations in Large Language Models. The successful 2.00 by 1.60 taped-out 65nm die results in 1.65 TOPS/W SoC energy efficiency with the capability to work on up to 1.1 GHz. **(*ESSERC'24, JSSC'25*)**
- **Sparse Attention Acceleration in LLMs with Synergistic In-memory Pruning and On-chip Recomputation:** Deploying a lightweight low-precision thresholding technique for the Attention mechanism using in-memory computing scheme in ReRAM to eliminate redundant off-chip memory access and on-chip computation for BERT, ALBERT, GPT-2, and ViT models. The proposed work results in 94% data movement reduction with only 0.34% accuracy degradation, thanks to on-chip recomputation capability. **(*MICRO'22*)**
- **Token-Adaptive Low-rank KV Cache Compression:** Compressing the KV cache through applying SVD to the projection layers with token-adaptive rank selection, achieving 60% compression with minimal loss. The PIM-enabled system schedules operations on HBM-based PIM units or xPU, achieving 1.58x and 1,96x speedup compared to SoTA GPU-only and GPU-PIM designs. **(Under review at MICRO)**
- **Accelerator Design for Ultra-Compressed Transformers:** Design and fabrication of a neural engine to accelerate the computation of ultra-compressed transformer blocks achieved by combining tensor decomposition, mixed-precision quantization, and token pruning. **(On going)**
- **End-to-end Acceleration of Generative Models with Runtime Regularized KV Cache Management:** Designing an algorithm-hardware framework to compress the KV cache context in generative model by keeping the same number of tokens across all inputs, heads, and layers, accompanied by a memory management unit for efficient memory accesses and a task scheduling mechanism for batch processing. **(*JETCAS'25*)**
- **Reconfigurable Accelerator for Encoder and Decoder Transformer Layers in LLMs:** Designing a 2-D array-based architecture for processing Transformers with reconfiguration for the processing of encoder and decoder blocks and token pruning capability with task-based bit-precision reconfigurability in the Attention computations. **(*ISLPED'24*)**
- **Hyperdimensional In-memory Computing of Mass Spectrometry Analysis using NVM:** A cross-disciplinary study to design a system for hyperdimensional mass spectrometry analysis with in-memory computing using various non-volatile memory devices such as PCM and 3D FeNAND, from device and circuit to architecture, compiler, and algorithm. **(*JXCDC'24, ICCAD'25*)**
- **GAN for Medical Image Synthesis:** Exploiting a Deep Convolutional Generative Adversarial Network (DCGAN) to augment a Chest X-Ray dataset to increase a VGG classifier's accuracy for pneumonia detection. *(SP23, ECE285)*
- **Reinforcement Learning for Safe Driving:** Designing a Q-Learning algorithm for a Markov Decision Process in a sample autonomous driving scenario. (WI24, CSE257)
- **Branch Predictor Design:** Designing a customized TAGE branch predictor in C. The custom BP beat both Tournament and GShare predictors on a set of different integer, floating point and memory instructions. (SP22, CSE240A)