Amir Taherin

PhD Candidate in Computer Engineering – Northeastern University
✉ taherin.a@northeastern.edu • amirtaherin.github.io
Last Updated: December 5, 2025

Professional Summary

PhD candidate specializing in **edge AI systems**, **efficient LLM inference**, **GPU/SoC performance analysis**, and **robotics-focused VLA pipelines**. Experienced in leading multi-student research teams, deploying LLMs on NVIDIA Jetson platforms, and conducting system-level profiling across quantization formats and hardware generations. Strong record of interdisciplinary collaboration with robotics, systems, and ML groups. Proven ability to design, optimize, and evaluate large-scale embedded AI workloads with attention to latency, power, memory behavior, and real-world deployment constraints.

Education

PhD in Computer Engineering — **Northeastern University**
Advisors: Profs. David Kaeli and Yanzhi Wang — *2020–Present*

MS in Computer Science — **University of Rochester**
2018–2020

MS in Computer Systems Architecture — **Sharif University of Technology**
Advisor: Prof. Alireza Ejlali — *2014–2016*

BS in Computer Engineering — **K. N. Toosi University of Technology**
2006–2011

Research Experience

NUCAR Laboratory, Northeastern University — **Boston, MA**
Graduate Research Assistant — *2021–Present*

Generalist Robotic Learning

- Led a multi-student research team to build an **end-to-end VLA pipeline** for generalist robotic policy learning, maintaining direct collaboration with industry.
- Deployed and benchmarked VLA models (OpenVLA, OpenVLA-oft, SpatialVLA, and our VOTE) on **NVIDIA Jetson AGX Orin** under various power budgets and on **high-performance GPUs** (H100, A100, V100, A6000).
- Upgraded OpenVLA-oft by replacing the LLM backbone with Qwen and Moxin models and redesigning the action head to achieve faster inference on embedded platforms.
- Integrated the full VLA stack with Kinova robotic arms for real-world manipulation experiments.

LLM Inference on Edge Devices

- Led a multi-student research team to perform a **system-level characterization of LLM inference** on embedded edge platforms.
- Deployed and benchmarked 13 LLMs (1B-8B parameters) from diverse model families–including LLaMA, Qwen, Gemma, Granite, Mistral, Phi, and Moxin–on **NVIDIA Jetson AGX Orin and Xavier**.
- Evaluated prompt and instruction-following workloads using the **HuggingFace Transformers** stack and the **IFEval** benchmark suite.
- Executed all models using **Llama.cpp** under multiple **quantization formats** (Q8, Q6, Q4) to study quantization-induced effects on latency, memory behavior, and throughput.
- Analyzed how quantization interacts with **SoC architecture, memory hierarchy, DVFS behavior**, and **GPU scheduling policies** on Xavier (Volta) and Orin (Ampere).
- Designed a unified profiling pipeline capturing **TTFT**, **token latency**, **effective memory bandwidth**, **KV-cache behavior**, **thermal characteristics**, and **GPU/CPU/Memory power** metrics.

Visual Inference on Edge Devices

- Developed a lightweight, workload-aware framework for adaptive visual inference that improves energy efficiency and increases object-detection accuracy **without violating real-time constraints** on embedded systems.
- Designed a parallel **Bayesian Optimization** method to balance the trade-off between high-resolution visual inference and runtime efficiency by adaptively adjusting input resolution.
- Implemented **reinforcement learning** baselines for dynamic resolution selection and evaluated their performance relative to the proposed BO-based approach.

Goodwill Laboratory, Northeastern University — **Boston, MA**
Graduate Research Assistant — *2020–2021*

HPC Reliability and Failure Analysis

- Conducted reliability analysis of large-scale **GPU-accelerated supercomputers.**
- Analyzed multi-year failure and repair logs from the Tsubame-2 and Tsubame-3 supercomputers to characterize fault behavior across generations of **multi-GPU compute nodes**.
- Identified systemic trends in GPU, node, and interconnect failures, and examined recovery behavior of large-scale HPC systems.

Systems Group, University of Rochester — **Rochester, NY**
Graduate Research Assistant — *2018–2020*

Energy-Efficient 360° Video Rendering on FPGAs

- Developed an **algorithm-architecture co-designed** system for real-time 360° **AR/VR video rendering** targeting FPGA acceleration.
- Addressed the prohibitive on-chip memory footprint of naive AR/VR rendering pipelines by restructuring the underlying video processing algorithm.
- Implemented the system on Zynq UltraScale+ MPSoC Evaluation Kit, enabling energy-efficient 360° video processing without loss of performance compared to commercial AR/VR rendering systems.

ESRLab, Sharif University of Technology — **Tehran, Iran**
Graduate Research Assistant — *2014–2017*

Reliability-Aware Energy Management in Mixed-Criticality Systems

- Proposed a **reliability-aware energy management** framework for mixed-criticality systems, targeting safe energy reduction in non-safety-critical workloads.
- Designed and evaluated three optimization techniques–**Monotonous-DVFS**, **Stretch**, and a combined DVFS/Stretch method–to exploit slack and degrade low-criticality service levels in a controlled manner.
- Achieved high energy savings with bounded service degradation while preserving system reliability, validated through extensive experiments.

Publications

Preprints

2025: **Cross-Platform Scaling of Vision-Language-Action Models from Edge to Cloud GPUs**. **Amir Taherin**, Juyi Lin, Arash Akbari, Arman Akbari, Pu Zhao, Weiwei Chen, David Kaeli, Yanzhi Wang, arXiv:2509.11480, 2025.

2025: **VOTE: Vision-Language-Action Optimization with Trajectory Ensemble Voting**. Juyi Lin, **Amir Taherin**, Arash Akbari, Arman Akbari, Lei Lu, Guangyu Chen, Taskin Padir, Xiaomeng Yang, Weiwei Chen, Yiqian Li, Xue Lin, David Kaeli, Pu Zhao, Yanzhi Wang, arXiv:2507.05116, 2025.

Journal Papers

2018: **Reliability-Aware Energy Management in Mixed-Criticality Systems**. **Amir Taherin**, Mohammad Salehi, Alireza Ejlali. *IEEE Transactions on Sustainable Computing*, doi:10.1109/TSUSC.2018.2801123, 2018.

Conference Papers

2021: **Examining Failures and Repairs on Supercomputers with Multi-GPU Compute Nodes**. **Amir Taherin**, Tirthak Patel, Giorgis Georgakoudis, Ignacio Laguna, and Devesh Tiwari. *In The 51st Annual IEEE/IFIP International*

Conference on Dependable Systems and Networks, doi:10.1109/DSN48987.2021.00043, 2021.

2020: **Energy-Efficient 360-Degree Video Rendering on FPGA via Algorithm-Architecture Co-Design**. Qiuyue Sun, **Amir Taherin**, Yawo Siatitse, and Yuhao Zhu. *In The 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '20). Association for Computing Machinery*, doi:10.1145/3373087.3375317, 2020.

2015: **Stretch: Exploiting Service Level Degradation for Energy Management in Mixed-Criticality Systems**. **Amir Taherin**, Mohammad Salehi, Alireza Ejlali. *The CSI Symposium on Real-Time and Embedded Systems and Technologies (RTEST)*, doi:10.1109/RTEST.2015.7369846, 2015.

Technical Skills

Programming: C/C++ (OpenMP, MPI, pthreads), Python, CUDA, Verilog, Bash, MATLAB, x86/ARM Assembly

ML/AI: PyTorch, TensorRT-LLM, ONNX, Llama.cpp

Dev. Boards: NVIDIA Jetson AGX Orin, Jetson AGX Xavier, Zynq UltraScale+ MPSoC ZCU104 Evaluation Kit

CAD Tools: Synopsys (*Design Compiler, HSPICE, PrimePower, Platform Architect*), Cadence (*Virtuoso, SoC Encounter*), Mentor Graphics (*ModelSim*), Xilinx (*ISE Design Suite, Vivado HLS, SDSoC*), Simulink

Typesetting: LaTeX, TeX, Markdown

Research Interests

- Computer Architecture, SoC, and GPU Design
- AI Acceleration, Edge Computing, and Efficient Inference
- Neural Network Optimization and Deployment
- Parallel, Heterogeneous, and Real-Time Systems

Teaching Experience

Teaching Assistant, High Performance Computing — **Northeastern University**
Course Instructor: Prof. David Kaeli — *Fall 2022*

Teaching Assistant, Parallel and Distributed Computing — **University of Rochester**
Course Instructor: Prof. Sandhya Dwarkadas — *Spring 2020*

Teaching Assistant, Programming Languages Design and Implementation — **University of Rochester**
Course Instructor: Prof. Michael L. Scott — *Fall 2019*

Teaching Assistant, Computer Organization — **University of Rochester**
Course Instructor: Prof. Yuhao Zhu — *Spring 2019*

Teaching Assistant, Embedded Systems Design — **Sharif University of Technology**
Course Instructor: Prof. Alireza Ejlali — *Spring 2016*

Teaching Assistant, Logic Design — **Sharif University of Technology**
Course Instructor: Prof. Shaahin Hessabi — *Spring 2016*

Teaching Assistant, Advanced Logic Design — **Sharif University of Technology**
Course Instructor: Prof. Alireza Ejlali — *Spring 2015*

Honors & Awards

2022: **Quantum Excellence** in Quantum Simulation from IBM Qiskit Global Summer School

2021: **Quantum Excellence** in Quantum Machine Learning from IBM Qiskit Global Summer School

2016: **Ranked 3rd** in cumulative GPA among all students of computer architecture (41 students), Sharif University of Technology, Tehran, Iran.

2015: **National Talent Award** for exceptional GPA from Sharif University of Technology, Tehran, Iran.

Academic Services

TETCSI-2018: *Reviewer* IEEE Transactions on Emerging Topics in Computing

RTEST-2017: *Reviewer* The CSI Symposium on Real-Time and Embedded Systems and Technologies (RTEST)

RTEST-2015: *Reviewer* The CSI Symposium on Real-Time and Embedded Systems and Technologies (RTEST)

Languages

- **English**: Full professional proficiency (TOEFL 115/120)
- **Persian**: Native