Natalia Frumkin (Natasha)

(408)-314-2208 • nfrumkin@utexas.edu • www.nfrumkin.com

Education

The University of Texas Austin — **Electrical and Computer Engineering**
PhD — *2020-2025*

Expected Graduation: December 2025
Research Focus - Post-Training Quantization & Model Compression for Language & Vision Models
Advisor: Diana Marculescu

The University of Texas Austin — **Electrical and Computer Engineering**
MS — *2024*

Boston University — **Computer Engineering**
BS, Magna Cum Laude — *2020*

Research Interests

Model Quantization: INT4 & FP4 Computation, KV-Cache Compression, Vector Quantization

Efficient AI Inference: Compressing State-Space Models, Diffusion Models, and Diffusion Language Models

Industry

Research Scientist Intern, Student Researcher
Meta — *May-Nov '23*

Project: Quantized Representations for Large-Scale Information Retrieval
· Contributed to an industrial-scale retrieval system for Instagram & Facebook Reels
· Tailored multiple training and post-training quantization techniques to dense similarity-search-based retrieval

Research Intern
Arm Ltd. — *Summer '22*

Project: Quantizing Vision Transformers
· Developed a quantization framework for DeiT & ViT models
· Conducted performance analysis of ViT models on an internal hardware simulator

Research Intern
Advanced Micro Devices (AMD) — *Summer '21*

Project: Hardware-Aware Neural Architecture Search
· Created a neural architecture search framework for AMD Research
· Hand-designed multiple neural networks for a scientific application
· Incorporated hardware metrics into the ProxylessNAS search space

Supercomputing Intern
Los Alamos National Labs — *Summer '20*

Project: Bottleneck Analysis on Virtual HPC clusters
· Designed bash scripts for monitoring, data transfer, and visualization
· Participated in HPC cluster management bootcamp where we set up a 10 node cluster by hand, and automatically through Ansible & Warewulf

AI/Data Science Intern
Red Hat — *Summer '18*

Project: Data Science on Prometheus Metrics
· Designed an anomaly detection system for time series data
· Implemented various forecasting models (ARIMA, Fourier Analysis, Prophet)
· My personal internship project has 145 stars and 53 forks on GitHub

Research

PhD Candidate — **U.T. Austin**
Energy-Aware Computing Lab — *2020–Current*
Research focus: Advancing post-training quantization: optimization methods for efficient AI inference

Distinguished Summer Research Fellow — **Boston University**
Information and Data Sciences Lab — *2019 – 2020*
Research focus: an ad-hoc online learning algorithms for divergence learning

Clare Boothe Luce Scholar — **Boston University**
Visual Information Processing Lab — *2017 – 2019*
Research focus: a privacy-preserving indoor localization and tracking system

Awards

Qualcomm Technologies, Inc.
Qualcomm Innovation Fellowship Finalist — *2024*

The University of Texas at Austin
Graduate School Mentoring Fellowship — *2022*

The University of Texas at Austin
Cockrell School of Engineering Fellowship — *2020*

Selected Publications

See full publication list on Google Scholar.

[1] **N. Frumkin** and Diana Marculescu. "Q-Sched: Pushing the Boundaries of Few-Step Diffusion Models with Quantization-Aware Scheduling". In: *under review* (2025).

[2] Hung-Yueh Chiang, Chi-Chih Chang, **N. Frumkin**, Mohammad Abdelfattah, Kai-Chiang Wu, and Diana Marculescu. "Quamba2: A Robust and Scalable Post-training Quantization Framework for Selective State Space Models". In: *ICML* (2025).

[3] Hung-Yueh Chiang, Chi-Chih Chang, **N. Frumkin**, Kai-Chiang Wu, and Diana Marculescu. "Quamba: A Post-Training Quantization Recipe for Selective State Space Models". In: *ICLR* (2024).

[4] **N. Frumkin**, Dibakar Gope, and Diana Marculescu. "Jumping through Local Minima: Quantization in the Loss Landscape of Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (*ICCV*)*. 2023, pp. 16978–16988.

[5] Tanvir Mahmud, **N. Frumkin**, and Diana Marculescu. "RL-Tune: A Deep Reinforcement Learning Assisted Layer-wise Fine-Tuning Approach for Transfer Learning". In: *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at* ICML. 2022.

[6] Hung-Yueh Chiang, **N. Frumkin**, Feng Liang, and Diana Marculescu. "MobileTL: On-device Transfer Learning with Inverted Residual Blocks". In: *AAAI* (2023).