Vima Gupta

vimagupta.github.io

Email: vgupta345@gatech.edu
+14703349450 linkedin.com/in/vima-gupta

Education

Georgia Institute of Technology — Atlanta, GA
PhD Computer Science, specializing in Systems for AI; advised by Dr. Anand Iyer — *Aug'23-Dec'27*
M.S. Computer Science, specializing in Computing Systems (thesis track) — *Jan'21-May'23*

Relevant Coursework: Systems for Machine Learning, Advanced Operating Systems, Statistical Machine Learning

Birla Institute of Technology and Science (BITS), Pilani — Pilani, India
Bachelors of Engineering in Electrical & Electronics Engineering — *Aug'14-May'18*

Publications

- M. Hu, A. Gupta, J. Yuan, **V. Gupta**, T. Kim, X. Xu, J. Kulkarni, O. Dekel, V. Adve, C. Mendis, "VTC: DNN Compilation with Virtual Tensors for Data Movement Elimination," To appear in OSDI 2026.
- J. M. Cherian, A. M. Bharadwaj, **V. Gupta**, A. P. Iyer, "CHAI: CacHe Attention Inference for Text2Video."
- Z. Yang, K. Inani, K. Kabra, **V. Gupta**, A. P. Iyer, "SAFuzz: Semantic-Guided Adaptive Fuzzing for LLM-Generated Code."
- **V. Gupta**, O. Duran, G. Ananthanarayanan, A. Iyer "Alleviating GPU Memory Pressure in LLM Serving through Fine-grained Model Merging" (Under submission)
- **V. Gupta**, Jae Hyung Ju, Kartik Sinha, Ada Gavrilovska, Anand Iyer, "LYNX - Efficient MoE inference for workload-agnostic LLM serving" (Under submission)
- **V. Gupta** and S. Varma, "Understanding Infinity: Neural Network Models of Becoming a Cardinal Principle Knower" (CogSci '24) [Paper]
- **V. Gupta** and S. Varma, "Learning to count: a neural network model of the successor function" Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44, 2022. [Poster]
- **V. Gupta** and R. Singhal, "Performance analysis of a visible light vehicle-to-vehicle wireless communication system" 2019, International Conference on Microwave Integrated Circuits, Photonics and Wireless Networks (IMICPW). IEEE, 2019 **(Best Paper Award)** [Paper]

Research Internships

Adaptive Scheduling for Multi-model LLM Serving — *May'25 – Aug'25*
Research Advisor: Dr. Roshan Dathathri; Microsoft Research

- Developed adaptive pool-level KV-aware scheduler for multi-instance, multi-region LLM inference, optimizing SLO attainment and GPU utilization across agentic workloads, multi-turn conversation graphs.

Kairos: Adaptive Kernel Dispatch for Latency-Sensitive LLM Workloads — *May'24 – July'24*
Research Advisors: Dr. Ondřej Čertík, Dr. Janardhana Kulkarni, Dr. Abhinav Jangda; AI Frameworks & Microsoft Research

- Designed adaptive CUDA kernel system that dynamically tunes launch parameters, reducing latency by up to 30% versus static heuristics on SOTA flash-attention kernels validated on large-scale GPU clusters like H100s and A100s.

Research Experience

Designing Efficient LLM Inference Systems — *Aug'23 – Now*
Advised by Prof. Anand Iyer, Georgia Tech — Atlanta, GA

- Designed and implemented [LYNX], an efficient large-scale MoE serving system that dynamically reduces active experts per batch based on runtime insights, achieving 25% latency reductions with under 1% accuracy loss.
- Developed Sandhi, a system that exploits opportunities for fine-grained model merging for a given set of models to alleviate memory capacity pressure by upto 49% for agentic workflows while adhering to accuracy bound of 2%.

Work Experience

Cerebras Systems — *May'22 – July'22*
ML Frameworks Intern, Backend — Atlanta, GA

- Converted the block sparse attention graph in BigBird, an NLP transformer, to match with existing highly optimized full attention kernel, from Tensorflow to MLIR lowering, at compile time for improved performance.
- The transformation was implemented through an MLIR graph match and rewrite pattern, automated in C++.

PACE: Physical Activity and Care for Everyone — *May'21 – Dec'21*
Co-founder, CREATE-X — Atlanta, GA

- Developed a real-time exercise-feedback system for Android using Google's Mediapipe, including pose-detection models, error-tolerance heuristics, and extensible exercise-library interface for remote physical-training applications.
- Performed structured user-needs and market analysis to guide system design; built the product website and co-developed the cross-platform mobile prototype (iOS/Android) showcased during the CREATE-X incubator demo.

Arm Embedded Technologies *May'18 – Dec'20*
Design Engineer Bengaluru, India

- Led a sub-team of three interns to design an IoT subsystem for the open-source ecosystem. Synthesis, floorplanning and PnR for high performance cores, ultra low power machine learning accelerators and octa-core clusters.
- Youngest engineer selected consecutively for technical poster presentation at Arm's Global Engineering Conference.

Achievements and Extra-curricular activities

- Received the OSDI'24 travel grant to attend USENIX OSDI 2024, ATC 2024 conference in Santa Clara, California.
- Won 3rd position at Klaus Poster Symposium, held across College of Computing, Georgia Institute of Technology.
- Awarded the **EDIC fellowship** at EPFL, Lausanne, one among fifty candidates selected across the world.
- Awarded the **Adobe Research Women in Technology scholarship** 2022 from candidates across North America
- Student Organizations: Secretary at Quantum computing Association (2021), India Club Finance Leader (2021), English Drama Club Co-ordinator (2016-2017), Logistics head at Department of Controls (2015-2017).
- Awarded bronze medal for basketball in BITS Open Sports Meet, 2015
- Awarded 'Most Outgoing Student of the Year' in high school, 2012
- Secured All India 3rd rank a national quizzing competition, 'Kaho What's My Idea' hosted by Derek'O'Brien, 2011

Mentorship

Jae Hyung Ju *Aug'25 – present*
PhD Student

- Co-designed scalable expert reduction techniques to accelerate low-latency MoE serving.

Oytun Kuday Duran *May'25 – present*
PhD Student

- Developed noise-sensitive heuristics to enable merging of co-located LLMs and alleviate capacity bottlenecks.

Kartik Sinha *Jan'24 – Dec'24*
MS Student → Citadel Securities

- Implemented expert restriction techniques and evaluated policies that further accelerate low-latency MoE inference.

Skills and Teaching Experience

Programming skills – C++ (DSA and OOP), Python, C, OpenMP, OpenMPI, Assembly, MATLAB, Agile practices
Python Libraries and software suites – PyTorch, Numpy, Matplotlib, Tensorflow, Streamlit, Qemu, Libvirt, Vtune
Graduate Teaching Assistant — Head GTA for advanced graduate research-based class on Systems for AI managing supervising eight research projects. Held office hours and delivered tutorials on LLM fundamentals and data visualization; CS 6476 Computer Vision (OMSCS): Designed and graded assignments for 500+ students.