

Ahan Gupta

Linkedin

Github: <https://github.com/spikerhead01234>

Email : ahangupta.96@gmail.com

Mobile : +1-415-966-5501

EDUCATION

- **University of Illinois Urbana-Champaign** Champaign, IL
PhD in Computer Science - Advisor: Minjia Zhang Aug 2022 - Present
- **National University of Singapore** Singapore
Bachelor of Computing in Computer Science Aug 2017 - Dec 2021

EXPERIENCE

- **Google DeepMind** Mountain View, CA
Student Researcher May 2024 - November 2024
 - Investigated novel low-rank compression techniques to reduce KV-cache sizes for 10B+ parameter LLMs.
- **Citadel** Hong Kong
Software Engineering Intern May 2021 - Aug 2021
 - Designed an authentication library to enable developers to integrate authentication logic with different services
 - Contributed to a tool that monitors AWS usage of different desks
 - Designed and built a monitoring tool that enables traders to track internal services' uptime and accuracy
- **Google** Singapore
Software Engineering Intern May 2020 - Aug 2020
 - Designed Asynchronous Web APIs via OpenAPI for authorisation microservice in MojaLoop network
 - Designed database Schemas & built infrastructural groundwork to enable integration with said databases
 - Implemented APIs that enable secure FIDO signature validation in HapiJS and TypeScript
 - Merged all code into production

PUBLICATIONS

- Davis Wertheimer*, Haochen Shen*, Ahan Gupta*, Derrick Liu, Yu Chin Fabian Lim, Mudhakar Srivatsa, Raghu K. Ganti, Minjia Zhang, Naigang Wang. A Self-Pruning Transformer: Extreme KV-Cache Compression with Universal Attention. COLM 2026.
- Muyan Hu, Ahan Gupta, Jiachen Yuan, Vima Gupta, Taeksang Kim, Xin Xu, Janardhan Kulkarn, Ofer Dekel, Vikram Adve, Charith Mendis. VTC: DNN Compilation with Virtual Tensors for Data Movement Elimination. OSDI 2026.
- Hoa La*, Ahan Gupta*, Alex Morehead, Jianlin Cheng, Minjia Zhang. MegaFold: System-Level Optimizations for Accelerating Protein Structure Prediction Models. ISC 2026. **Kernels adopted by AMD for OpenFold3.**
- Ahan Gupta*, Zhihao Wang*, Neel Dani, Masahiro Tanaka, Olatunji Ruwase, Minjia Zhang. AutoSP: Unlocking Long-Context LLM Training Via Compiler-Based Sequence Parallelism. ICLR 2026. **Adopted by Microsoft DeepSpeed. Featured by PyTorch Foundation blog.**
- Yueming Yuan, Ahan Gupta, Jianping Li, Sajal Dash, Feiyi Wang, Minjia Zhang. X-MoE: Enabling Scalable Training for Emerging Mixture-of-Experts Architectures on HPC Platforms. SC 2025. **Best Student Paper.**
- Ahan Gupta, Yueming Yuan, Devansh Jain, Yuhao Ge, David Aponte, Yanqi Zhou, Charith Mendis. SPLAT: Optimized GPU code generation framework for SParse reguLAR ATtention. OOPSLA 2025.
- Ahan Gupta, Hao Guo, Yueming Yuan, Yanqi Zhou, Charith Mendis. FLuRKA: Fast fused Low-Rank & Kernel Attention. In Submission 2025. Preprint link: <https://arxiv.org/abs/2306.15799>.

* Denotes Equal Contribution

SERVICE

- **ACM TACO Reviewer:** 2025, 2026
- **ISCA AEC:** 2024

OPEN SOURCE CONTRIBUTIONS

- **DeepSpeed:** (42k stars)
- **Enzyme:** (1.6k stars)