MINGYU GUAN

(+1)470-334-6144 · my.guan@outlook.com

EDUCATION

Georgia Institute of Technology Aug. 2019 - Dec. 2025
Doctor of Philosophy in Computer Science Atlanta, GA
Advisors: Taesoo Kim and Anand Iyer

The Chinese University of Hong Kong (CUHK)* Aug. 2015 - May. 2019
Bachelor of Science in Computer Science with Honours, First Class Hong Kong, China

*Joint program offered by Sun Yat-Sen University and CUHK.

EXPERIENCE

Software Engineer Intern May. 2025 - Aug. 2025
Meta Platforms, Menlo Park

- Accelerating inference of *Latent Diffusion Models (LDM)* for image and movie generation;
- Developed an inference utility for fast benchmarking *DiT transformers*, supporting different configurations of *tensor parallelism (TP) and context parallelism (CP)*;
- Exploring and applying *async-TP and symmetric memory* on the combination of TP and CP to accelerate LDM inference.

Research Intern May. 2022 - Aug. 2022
Microsoft Research, Redmond

- Designed and implemented a novel heterogeneous Graph Neural Network (GNN) for compromised email detection on real-world email graphs[2];
- *Cooperated with a research team and a product team* to construct heterogeneous graphs from a large-scale noisy enterprise email data set and built an automatic system for detecting compromised email accounts.

SELECTED PROJECTS

Fast and Adaptive Graph-based RAG Aug. 2024 - Present

- Designing adaptive and query-aware graph-based RAG techniques with fast indexing and retrieval.
- Enabling caching intermediate states based on the unique pattern in the graph-based RAG.
- Devising a cache-aware scheduler to optimize resource usage, while serving user queries within SLOs.

Model Training Provenance with Confidential Computing March. 2024 - Present

- Identified the performance and privacy challenges in model training provenance;
- Designed proof generation and verification protocol with low overhead and security guarantees;
- Implemented on AMD SEV-SNP, supporting the latest LLMs including Open-R1 and Llama models.

E^3: High-throughput Inference for Early-exit LLMs May. 2023 - Sep. 2024

- Enabled efficient batching techniques during inference for existing early-exit (EE) LLMs, e.g. CALM (based on Google T5 model);
- Extended non-EE LLMs and compressed models, e.g., Llama family models and distilled BERT, to their EE counterparts;
- Accelerated inference goodput for autoregressive LLMs (2.8-3.8x) and compressed models (1.67x).

Distributed System for Dynamic Graph Neural Networks May. 2021 - Dec. 2023

- Supported efficient dynamic GNN (DGNN) training in large-scale distributed settings;
- Leveraged computational structure in the GNN-RNN approach to propose cross-layer optimizations;
- Enabled efficient distributed training that reserves both structure and time dependencies in dynamic graphs;
- Outperformed state-of-the-art GNN frameworks by up to 10.7x on various DGNNs and workloads.

Processing Billion-scale Dynamic Graphs on a Single Machine Jan. 2020 - Jul. 2021

- Introduced the design of cell abstraction, allowing a significant reduction in overall storage space as well as enabling a simple, yet effective load-balancing strategy;
- Proposed an API and execution model tailored for streaming graphs by incorporating a hybrid edge- and vertex-centric API coupled with the *edgeChanged* API to allow a timely reaction to graph changes;
- Designed a technique for concurrent analytics on streaming graphs, which fully exploits the similarities in data access among concurrent graph processing jobs.

PUBLICATIONS

[1] Sujin Park, **Mingyu Guan**, Xiang Cheng, and Taesoo Kim. Principles and Methodologies for System Performance Optimization. *The 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Boston, MA, Jul, 2025.

[2] **Mingyu Guan**, Jack W. Stokes, Qinlong Luo, Fuchen Liu, Purvanshi Mehta, Elnaz Nouri, and Taesoo Kim. Heterogeneous Graph Neural Network on Semantic Tree. *In Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, Feb 2025.

[3] **Mingyu Guan**, Saumia Singhal, Taesoo Kim, and Anand Padmanabha Iyer. ReInc: Scaling Training of Dynamic Graph Neural Networks. *arXiv preprint* arXiv:2501.15348, 2025.

[4] Anand Iyer, **Mingyu Guan**, Yinwei Dai, Rui Pan, Swapnil Gandhi, and Ravi Netravali. Improving DNN Inference Throughput Using Practical, Per-Input Compute Adaptation. *In Proceedings of the 30th Symposium on Operating Systems Principles (SOSP)*, Austin, TX, USA, Nov 2024.

[5] **Mingyu Guan**, Anand Padmanabha Iyer, and Taesoo Kim. DynaGraph: Dynamic Graph Neural Networks at Scale. *In Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems and Network Data Analytics (GRADES-NDA)*, Philadelphia, PA, Jun 2022.

HONORS AND GRANTS

2023 **Student Grant Award**, 17$^{\text{th}}$ USENIX OSDI — Boston, MA
2019 **Deans List**, CUHK — Hong Kong, China
2019 **Rev Mak Shuet Kwong Memorial Scholarship**, CUHK — Hong Kong, China
2017 **First Prize Academic Scholarship**, SYSU — Guangzhou, China
2016 **Jetta Scholarship for Outstanding Students**, SYSU — Guangzhou, China

TALKS

[1] TAITEE: Bridging the Trust Gap From Claims to Proof in AI Model Training, *Confidential Computing Summit*, San Francisco, CA, USA, Jun 2025.

SERVICES

- **Artifact Evaluation Committee**, SOSP '24.
- **External Review Committee**, ATC '24.

TEACHING EXPERIENCE

Graduate Teaching Assistant Georgia Institute of Technology, Atlanta

- CS8803 Systems for AI: Large Language Models, Spring 2024
- CS3251 Computer Networking, Spring 2020

SKILLS

Language Python, C/C++, SQL, Bash Script
Frameworks PyTorch, DGL, PyG, gRPC