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; GPA:4.0 Atlanta, GA

The Chinese University of Hong Kong Aug. 2017 - May. 2019
Bachelor of Science in Computer Science with Honours, First Class Hong Kong, China

EXPERIENCE

Graduate Research Assistant May. 2020 - Present
Advisors: Taesoo Kim and Anand Iyer *Georgia Institute of Technology, Atlanta*

- Building a trustworthy training framework to enable AI model provenance, while preserving the flexibility of custom training algorithms as well as data privacy with confidential computing.
- Building an adaptive and query-aware graph-based RAG with fast indexing and retrieval techniques.
- Built a high-throughput serving system with efficient batching techniques for early-exit large language models (LLMs) [3].
- Built a scalable distributed graph deep learning system for dynamic graphs, enabling efficient training on real-world graphs with billions of edges[1,4].

Research Intern May. 2022 - Aug. 2022
Mentor: Jay Stokes *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.

Undergraduate Research Assistant May. 2018 - Apr. 2019
Advisor: James Cheng *The Chinese University of Hong Kong, Hong Kong*

- Supported Distributed Online Analytical Processing (OLAP) on a general distributed system (Husky);
- Implemented SQL engine and customized query optimization rules on Husky using Apache Calcite.

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 proof of concept on Intel TDX, supporting the latest LLMs including BERT and Llama family 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] **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.

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

[3] 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.

[4] **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, June 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

TEACHING EXPERIENCE

Graduate Teaching Assistant Georgia Institute of Technology, Atlanta

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

SERVICES

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

SKILLS

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