

WENJUN YU

David C. Lam Building 625, Hong Kong Baptist University, Kowloon Tong District, Hong Kong, China

Homepage: <https://freshduer.github.io/> ◇ Email: Duterfresh@outlook.com ◇ Phone: +852 46727901

EDUCATION

Hong Kong Baptist University
Ph.D. student in Computer Science
Supervisor: Amelie Chi Zhou

Sep 2025 – Jul 2029

Dalian University of Technology (985 & 211)
B.S. in Artificial Intelligence

Sep 2019 – June 2023
GPA: 3.89/4.0

RESEARCH INTEREST

My research interests lie broadly in DLRMs training/serving optimization, generative Recommender system acceleration and GPU Virtualization techniques.

PUBLICATIONS

When KV Meets Embeddings: Dynamic GPU Memory Allocation for Accelerating Generative Recommender Serving

Wenjun Yu, Amelie Chi Zhou, Shuguang Han (Alibaba – Xianyu)

(SC 2026, CCF-A) International Conference for High Performance Computing, Networking, Storage, and Analysis

Near-Zero-Overhead Freshness for Recommendation Systems via Inference-Side Model Updates

Wenjun Yu, Sitian Chen, Amelie Chi Zhou, Cheng Chen (Bytedance – Tiktok)

(HPCA 2026, CCF-A) IEEE International Symposium on High-Performance Computer Architecture

WORK EXPERIENCE

Generative Recommender System Intern

Dec. 2025 – Apr. 2026

Alibaba Group

Beijing, China

- Designed an adaptive dynamic GPU-memory partitioner for the KV cache and embedding cache to accelerate inference in the generative recommender-serving system (HSTU). Work is published in SC 2026.

Research Assistant

Feb. 2025 – Aug. 2025

Hong Kong Baptist University, Bytedance tiktok

Hong Kong, China

- Designed a LoRA-based online update mechanism for recommendation models, reducing communication overhead during model updates. Work is published in HPCA 2026.

GPU Virtualization System Engineer

Aug. 2023 – Feb. 2025

VirtAi Technology

Beijing, China

- Design a playback-based hot migration module for NCCL-based deep learning programs, which allows deep learning tasks to be migrated to different GPUs within a single machine without interrupting the training process
- Design a GPU computing resource control module, it hooks the CUDA launch kernel APIs, then monitors and controls the CUDA kernel rate for deep learning jobs to balance GPU computing resources or set computing resource priorities for different tasks in the OS layer
- Design a fault-tolerant GPU virtualization mechanism that hooks CUDA APIs and uses Prometheus-based health monitoring to detect GPU and node failures and automatically migrate workloads to healthy GPUs or nodes, minimizing service disruption.

Quantitative Trading System Developer Outstanding Intern

Apr. 2022 – Sep. 2022

ZhuoShi Private Equity Fund

Beijing, China

- Develop an exchange's seat selection module based on the LambdaMart algorithm to select the seat with the lowest NDCG latency
- Develop a millisecond-level Kafka-based index calculation service

AWARDS & HONORS

RPg Research Performance Award of Hong Kong Baptist University (\$20,000 HKD)

Merit Student of Dalian University of Technology

National Scholarship of Dalian University of Technology

First Class Scholarship (Top 5%) of Dalian University of Technology

SKILLS

Techniques: C++, Cuda, Python

Languages: Mandarin(native), English(IELTS-7 CET 6-592)