# Fangyu Ding 丁方玉

fangyu.ding@outlook.com | dingfangyu.github.io | dingfangyu

## Education

**Hong Kong University of Science and Technology** 2024 –
PhD student, Computer Science, advised by Prof. Binhang Yuan

**Shanghai Jiao Tong University** 2021 – 2024
M.Eng., Computer Science, 3.70/4.0 GPA, advised by Prof. Junchi Yan

**Shanghai Jiao Tong University** 2017 – 2021
B.Eng., Computer Science, 88.9/100 GPA

## Technical Skills

**Mathematics**: Matrix Theory, Optimization, Probability Theory, Machine Learning, Complex Analysis, Discrete Mathematics

**Programming Languages and Tools**: Python, C/C++, CUDA, SQL, Shell, Julia, PyTorch, PyTorch-Geometric, Hugging Face, TensorRT-LLM

## Research Interests

- Machine Learning Systems, Especially for Large Language Models
  - (Distributed) Training: Parallelisms, Kernel (Fusion), Hardware Efficiency for Computation and Communication
  - (Efficient) Fine-Tuning: Parameter Efficiency, Memory Efficiency
  - Inference (Systems and Algorithms): Serving, Offloading, Speculative Decoding, Retrieval Augmented Generation
- Model Architecture Design for Computation/Memory Efficient Training/Inference
  - Sparse Computation: Mixture-of-Experts, Elastic Models, etc.
  - Low-Rank Computation and Interesting Mathematics: Linformer, Performer, etc.

## Publications

- **Fangyu Ding**, Haiyang Wang, Zhixuan Chu, Tianming Li, Zhaoping, Hu, Junchi Yan.
  *GSINA: Improving Subgraph Extraction for Graph Invariant Learning via Graph Sinkhorn Attention.* (Arxiv:2402.07191)
- **Fangyu Ding**, Junchi Yan, Haiyang Wang.
  *c-NTPP: Learning Cluster-Aware Neural Temporal Point Process.* (AAAI 2023)

## Work Experiences

**UCloud** April 2024 – July 2024
Algorithm Intern

- Performed multiple rounds of seminars to share knowledge on LLM training, fine-tuning, inference, serving, algorithms, models, systems, and hardware with the UCloud AI team.
- Benchmarked LLM inference throughput with different task settings with TensorRT-LLM framework, and profiled hardware computation/memory/communication bandwidths on 8xA100 and 8x4090 servers.
- Derived the cost models of LLM training/fine-tuning/inference, and used the PuLP package to solve the linear programming problem for throughput-maximized policy of LLM inference.

**AntGroup** October 2022 – May 2023

Research Intern

- Developed a subgraph-mining-based Graph Neural Network (GNN) model for robust graph classification and node classification, the Optimal Transport (OT) theory is leveraged to (sparsely, softly, and differentiably) score the importance of nodes and edges in the graph structure for a robustly weighted GNN message-passing computation.
- Conducted extensive experiments on both graph-level and node-level classification benchmarks, and our approach can outperform the baselines by large margins (up to ≈15% for graph classification and up to ≈20% for node classification).

**Alibaba Damo Academy** July 2021 – September 2021

Research Intern

- Developed a novel variational autoencoder (VAE) based deep sequential clustering framework to mine the latent clusters (i.e. subsequences) in event sequence, and a Transformer temporal point process (TPP) model is leveraged to model the temporal dynamics of the sequential data.