424-440-4678
qinzongyue@cs.ucla.edu
scholar.google.com/citations?user=8QDAVAgAAAA

Zongyue Qin

Education

2020–now **Ph.D.**, *Computer Science Department*, University of California, Los Angeles, US
- Advisor: Prof. Yizhou Sun
- GPA: 4.0/4.0

2016–2020 **B.A**, *Yuanpei College*, Peking University, China
- Major: Data Science and Big Data Technology
- GPA: 3.69/4 (top 5)

2018.9-2018.12 **Exchange Student**, *Computer Science Department*, University of British Columbia, Canada
- GPA: 96.5/100

Internship

2023.06–2023.9 **Applied Scientist Intern**, *Search, Query Understanding*, Amazon Inc, Palo Alto, CA, USA
Developed a framework than improves LLM's ability to answer questions whose answers are stored in relational databases by 50% by utilizing in-context learning and retrieval augmentation with Pytorch.

2022.06–2022.11 **Applied Scientist Intern**, *Search, Query Understanding*, Amazon Inc, Palo Alto, CA, USA
Developed a framework that accelerates similarity search with million-scale session data by over $7\times$ with $100\times$ less memory cost, and improves the downstream recommendation effectiveness by 5% via a novel pre-training learning-to-hash framework with Pytorch.

2020.02–2020.05 **Internship**, *Data, Ad System, Ad Data Platform Department*, ByteDance, Beijing, China
Worked with cross-functional teams and developed 5 new routine programs that generate data for downstream applications that improves the analysis by introducing smartly-designed data management programs with C/C++ and SQL.

Selective Projects

2024 **UCLA**, *Optimized Multi-Token Joint Decoding with Auxiliary Model for LLM Inference*
Developed an innovative decoding algorithm that enhances LLM inference perplexity by 20.2%, boosts downstream task performance, and achieves a 1.4 $\times$ increase in speed and 1.5 $\times$ reduction in energy consumption compared to speculative decoding, by sampling multiple tokens from their joint probability with the assistance of an auxiliary model.

2024 **UCLA**, *Dynamic-Width Speculative Beam Decoding for Efficient LLM Inference*
Proposed a novel speculative beam decoding algorithm for LLM inference that achieves 1.5-1.9$\times$ speed up and 1.8-2.5$\times$ lower energy consumption compared to beam sampling, with no loss in downstream perfrmance. Moreover, it can improve downstream scores by 15-50% compared to conventional speculative decoding, while achieving similar time, memory, and energy consumption. These are achieved by introducing a novel draft and verification scheme.

2023 **UCLA**, *Cross-Modality Program Representation Learning for Electronic Design Automation with High-Level Synthesis*
Designed a multi-modality model that combines GNN and LLM to predict HLS design's performance and achieved 22% less error and $1.26\times$ better design space exploration results by fusing graph modality and source code modality of programs.

2020 **UCLA**, *Graph Hashing via Graph Neural Network for Similarity Search*
Developed a GNN-based algorithm in tensorflow that enhances graph similarity search speed by $6.23\times$ by using GNN to estimate the lower bound of graph edit distance and writing optimized C/C++ code to accelerate search.

Publications

Zongyue Qin, Yunsheng Bai, Atefeh Sohrabizadeh, Zijian Ding, Yizhou Sun, Jason Cong. "*Cross-Modality Program Representation Learning for Electronic Design Automation with High-Level Synthesis*". MLCAD2024, github

ZongyueQin, Ziniu Hu, Zifan He, Neha Prakriya, Jason Cong, Yizhou Sun"*Optimized Multi-Token Joint Decoding with Auxiliary Model for LLM Inference*" arxiv preprint, github, submitted to ICLR25

ZongyueQin, Zifan He, Neha Prakriya, Jason Cong, Yizhou Sun"*Dynamic-Width Speculative Beam Decoding for LLM Inference*" arxiv preprint, submitted to AAAI25

Weikai Li, Ding Wang, Zijian Ding, Atefeh Sohrabizadeh, **Zongyue Qin**, Jason Cong, Yizhou Sun"*Hierarchical Mixture of Experts: Generalizable Learning for High-Level Synthesis*", submitted to AAAI25

Zifan He, **Zongyue Qin**, Neha Prakriya, Jason Cong, Yizhou Sun "*HMT: Hierarchical Memory Transformer for Long Context Language Processing*" arxiv preprint, github, submitted to Neurips24

Ding, Z., Sohrabizadeh, A., Li, W., **Qin, Z.**, Sun, Y., & Cong, J. (2024). Efficient Task Transfer for HLS DSE. ICCAD 2024.

Zongyue Qin, Chen Luo, Haoming Jiang, Zhengyang Wang, Yizhou Sun. "*Relational Database Augmented Large Language Models*" arxiv preprint

Yunsheng Bai, Atefeh Sohrabizadeh, **Zongyue Qin**, Ziniu Hu, Yizhou Sun, Jason Cong. "*Towards a Comprehensive Benchmark for High-Level Synthesis Targeted to FPGAs.*", Neurips2023, github

Zongyue Qin, Yunsheng Bai, Yizhou Sun. "*GHashing: Semantic Graph Hashing for Approximate Similarity Search in Graph Databases.*", SIGKDD2020.

Yuyan Chen, Lei Zou, **Zongyue Qin**. "*Gated Relational Graph Neural Network for Semi-supervised Learning on Knowledge Graphs*", WISE'19.

Teaching Experience

Teaching Assistant — **University of California, Los Angeles**, *Computer Science Department*, 2023-2024
- "*Introduction to Data Science*", Prof. Yizhou Sun
- "*Introduction to Data Science*", Prof. Yizhou Sun, Dr. Shichang Zhang

Teaching Assistant — **University of California, Los Angeles**, *Computer Science Department*, 2021-2022
- "*Introduction to Data Mining*", Prof. Yizhou Sun
- "*Advanced Data Mining*", Prof. Yizhou Sun
- "*Introduction to Data Mining*", Prof. Wei Wang

Honors and Awards

- Peking University Scholarship, 2019
- Peking University Freshman Scholarship, 2016

Leadership

2018–2019 **Vice-chairman**, *Presidium*, Yuanpei Academic Societies of Students, Beijing

2017–2018 **Director**, *Academic Department*, Yuanpei Academic Societies of Students, Beijing