

|                     |  |   |
|---------------------|--|---|
| Contact Information | No. 38 Tongyan Road, Jinnan District<br>Tianjin, P.R. China 300350   | tnyang2000@gmail.com<br><a href="https://tiannuo-yang.github.io/">https://tiannuo-yang.github.io/</a> |
| Research Interests  | <p>As Moore’s Law fades, what will be the next-generation AI system that can overcome the conflict between surging computational needs and scarce hardware resources? Inspired by this question, my current research focuses on building low-cost, high-efficiency machine learning systems that can be deployed to serve a wide range of scenarios.</p> <p>Specifically, I identify performance bottlenecks, opportunities, and challenges in the system, and leverage technologies such as online learning to address the performance- and resource-oriented problems. My works aim to harness hardware resources, automate the operation of complex systems, and enhance the efficiency of systems (e.g., retrieval augmented generation).</p>  |   |
| Education           | <p><b>Nankai University</b>, Tianjin, China<br/> <i>Master in Computer Science</i> (Exempted from Entrance Exam) August 2022 – Present<br/>         Advisor: Professor Yusen Li GPA: 3.63/4.0<br/>         Thesis: Automated Performance Tuning Techniques for Parallel Applications</p> <p><b>University of Science and Technology Beijing</b>, Beijing, China<br/> <i>Bachelor in Information Management and Information System</i> August 2018 – June 2022<br/>         Major GPA: 3.97/4.0 Cumulative GPA: 3.89/4.0</p> <p><b>Southern Taiwan University of Science and Technology</b>, Taiwan, China<br/> <i>Major in Information Management</i> (Exchange Program) September 2019 – January 2020<br/>         Cumulative GPA: 4.3/4.3</p>  |   |
| Research Experience | <p><b>University of Illinois at Urbana-Champaign</b>, Urbana, IL, United States<br/> <i>Retrieval-Augmented Generation, Text/Video Foundation Model</i> May 2024 – Present<br/>         Working with Professor Minjia Zhang on GPU-enhanced retrieval augmented generation, shedding lights on key concerns like batching strategies and latency-quality tradeoffs.</p> <p><b>Nankai University</b>, Tianjin, China<br/> <i>Datacenter, System, Machine Learning for System</i> August 2022 – Present<br/>         Working with Professor Yusen Li on automatic performance tuning and hardware resource isolation for job collocations within multi-core systems.</p> <p><b>Ant Group</b>, Beijing, China<br/> <i>Vector Retrieval, Vector Database Optimization</i> June 2023 – January 2024<br/>         Worked as a research intern under Dr. Jianguo Li and Wen Hu on optimizing AI infrastructure - vector database, enhancing CodeFuse services (a coding large language model platform).</p> <p><b>University of Chinese Academy of Sciences</b>, Beijing, China<br/> <i>Mixed Integer Programming, Heuristic Algorithm</i> September 2020 – September 2021<br/>         Worked as an undergraduate research assistant under Professor Guanghui Zhou to develop data-driven combinatorial optimization problem (vehicle routing optimization).</p> |   |
| Publications        | <p><b>Performance Tuning</b></p> <p>*Denotes equal contribution. †Denotes corresponding authorship.</p> <p>K. Cheng, Z. Wang, W. Hu, <u>T. Yang</u>, J. Li, S. Zhang. “SCOOT: Towards SLO-Optimized LLM Serving via Automatic Inference Engine Tuning.” Under Review at <i>The Web Conference (WWW)</i>, 2025.</p>   |   |

T. Yang, W. Hu, W. Peng, Y. Li, J. Li, X. Liu, G. Wang. “VDTuner: Automated Performance Tuning for Vector Data Management Systems.” *International Conference on Data Engineering (ICDE)*, 2024.

T. Yang, R. Chen, Y. Li, X. Liu, G. Wang. “CoTuner: A Hierarchical Learning Framework for Coordinately Optimizing Resource Partitioning and Parameter Tuning.” *International Conference on Parallel Processing (ICPP)*, 2023.

### Research Survey

Y. Zhou\*, X. Lin\*, X. Zhang\*, M. Wang\*, G. Jiang\*, H. Lu\*, Y. Wu\*, K. Zhang\*, Z. Yang\*, K. Wang\*, Y. Sui\*, F. Jia\*, Z. Tang\*, Y. Zhao\*, H. Zhang\*, T. Yang\*, W. Chen\*, Y. Mao\*, Y. Li\*, D. Bao\*, Y. Li\*, H. Liao\*, T. Liu\*, J. Liu\*, J. Guo\*, X. Zhao, Y. WEI, H. Qian, Q. Liu, X. Wang, W.K. Chan, C. Li, Y. Li, S. Yang, J. Yan, C. Mou, S. Han, W. Jin, G. Zhang, X. Zeng. “On the Opportunities of Green Computing: A Survey.” *arXiv*, 2023  
(Writing Section: 6.4 Resource Optimization).

### Operations Research (Undergraduate Thesis)

T. Yang<sup>†</sup>, Z. Chu, B. Wang. “Feasibility on the Integration of Passenger and Freight Transportation in Rural Areas: A Service Mode and an Optimization Model.” *Socio-Economic Planning Sciences* (SCI JCR Q1), 2023.

### Research Projects

|   |                               |
|---|-------------------------------|
| Qiyuan Laboratory Innovation Fund   | November 2023 – November 2024 |
| Worked as a core member on resource isolation mechanism for a multi-tenant cache system (i.e., Cachelib by Facebook).                                 |                               |
| CCF-Ant Research Fund on Green Computing  | January 2023 – January 2024   |
| Worked as a leader to write project proposal, conduct research on AI infrastructure (vector database optimization) and practical platform deployment. |                               |
| National Natural Science Foundation (NSF) of China  | January 2023 – Present        |
| Working as a core member on improving resource utilization in cloud with QoS guarantee.   |                               |
| Major Project of National NSF of China  | December 2022 – Present       |
| Working on real-time scheduling of cluster robots for major equipment manufacturing.  |                               |

### Honors And Awards

|   |            |
|---|------------|
| 1st-Class Gongneng Scholarship, Nankai University   | 2023, 2024 |
| National 3rd Prize, Massive Storage Competition   | 2022       |
| Provincial Gold Prize, 'Internet+' Innovation and Entrepreneurship Competition                          | 2021       |
| Excellent Student, University of Science and Technology Beijing (USTB)                                  | 2020       |
| National 2nd Prize (1493/41826 = top 3.6%), Contemporary Undergraduate Mathematical Contest in Modeling | 2020       |
| 2nd Place, Marketing Competition (turnover RMB 20,000+), USTB   | 2020       |
| Top Ten Singers on Campus, USTB   | 2020       |
| Professional Certification in E-Commerce, Taiwan Computer Skills Foundation (CSF)                       | 2019       |
| Certification of Enterprise Electronic Assistant Planner, Taiwan CSF                                    | 2019       |
| Team Gold Award, College Students' Social Practice, USTB  | 2019       |

### Talks And Services

**Reviewer** for  
the 2025 ACM Web Conference.

**Conference Talk** at  
the 40th ICDE at Utrecht, the Netherlands, May 2024;  
the 52nd ICPP, Online, August 2023.

**Teaching Assistant** for

Computer Architecture (Fall 2023); C++ (Spring 2024).

**Reading Group Founder and Leader** for

Machine learning system research at Nankai-Baidu Joint Lab (from October 2024).

## Open Source

VDTuner

<https://github.com/tiannuo-yang/VDTuner>

An automated performance tuning system for vector data management systems.

G-VRP-IPD-TW

<https://github.com/tiannuo-yang/G-VRP-IPD-TW>

Mathematical model and real-world scenario-based dataset for green vehicle routing problem with integrated pickup and delivery and time windows (G-VRP-IPD-TW).