# Junhao Hu

No.5 Yiheyuan Road – Beijing – China
junhaohu@stu.pku.edu.cn • derekhjh.com • DerekHJH

## Education

**Peking University** **Sept. 2022 – June 2027 (expected)**
*Ph.D. student in Computer Science*

- Advisor: *Prof.* Tao Xie

**Nanjing University** **Sept. 2018 – June 2022**
*Bachelor of Computer Science and Technology*

- Advisor: *Prof.* Yanyan Jiang
- GPA: 92.1/100
- Nanjing University Outstanding Undergraduate, Class of 2022
- Excellent Undergraduate, Kuangyaming Honors School

## Publications

- xDeepServe Team @ Huawei. xDeepServe: Model-as-a-Service on Huawei CloudMatrix384. In: *arXiv preprint arXiv:2508.02520*
- Qingyuan Liang, Zeyu Sun, Qihao Zhu, **Junhao Hu**, Yifan Zhao, Yizhou Chen, Mingxuan Zhu, Guoqing Wang, Lu Zhang. Directional Diffusion-Style Code Editing Pre-training. In: *Transactions of Software Engineering, 2025.* (TSE 2025) **CCF-A**
- **Junhao Hu**, Wenrui Huang, Weidong Wang, Zhenwen Li, Tiancheng Hu, Zhixia Liu, Xusheng Chen, Tao Xie, Yizhou Shan. RaaS: Reasoning-Aware Attention Sparsity for Efficient Long-Decoding Inference. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025.* (ACL 2025) **CCF-A**
- **Junhao Hu**, Wenrui Huang, Haoyi Wang, Weidong Wang, Tiancheng Hu, Qin Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, Tao Xie. EPIC: Efficient Position-Independent Context Caching for Serving Large Language Models In: *Proceedings of the 42nd International Conference on Machine Learning, 2025* (ICML 2025.) **CCF-A**
- **Junhao Hu**, Jiang Xu, Zhixia Liu, Yulong He, Yuetao Chen, Gengyuan Dan, Baoquan Zhang, Shining Wan, Zhiyu Dong, Hao Xu, Zhihao Ren, Jiang Liu, Jie Meng, Changhong Liu, Tao Xie, Dayun Lin, Qin Zhang, Yue Yu, Hao Feng, Xusheng Chen, Yizhou Shan. DeepFlow: Serverless Large Language Model Serving at Scales In: *Proceedings of the 2025 USENIX Conference on Usenix Annual Technical Conference, 2025.* (ATC 2025) **CCF-A**
- Chaozheng Wang, **Junhao Hu**, Cuiyun Gao, Yu Jin, Tao Xie, Hailiang Huang, Zhenyu Lei, Yuetang Deng. How Practitioners Expect Code Completion? In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pages 1294-1306, 2023.* (ESEC/FSE 2023) **CCF-A**
- **Junhao Hu**, Chaozheng Wang, Hailiang Huang, Huang Luo, Yu Jin, Yuetang Deng, Tao Xie. Predicting Compilation Resources for Adaptive Build in an Industrial Setting. In: *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering, pages 1808-1813, 2023.* (ASE 2023) **CCF-A**
- Cunchen Hu, Heyang Huang, **Junhao Hu**, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, Yizhou Shan. MemServe: Context Caching for Disaggregated LLM Serving with Elastic Memory Pool. In: *arXiv preprint arXiv:2406.17565*
- Shiju Zhao, **Junhao Hu**, Rongxiao Huang, Jiaqi Zheng, Guihai Chen. MPIC: Position-Independent Multimodal Context Caching System for Efficient MLLM Serving. In: *arXiv preprint arXiv:2502.01960*
- Yuchen Liu, **Junhao Hu**, Yingdi Shan, Ge Li, Yanzhen Zou, Yihong Dong, Tao Xie. LLMigrate: Transforming "Lazy" Large Language Models into Efficient Source Code Migrators In: *arXiv preprint arXiv:2503.23791*

## Internship Experiences

**Xiaomi Top Talent (on par with ByteDance TopSeed, Tencent Qingyun)** **Sep. 2025 – Now**
*Research intern mentored by Fuli Luo*

- Model architecture and infra co-design.

**Huawei Cloud** **July 2024 – Aug. 2025**
*Research intern mentored by Yizhou Shan and Chenxi Wang*

- Lead the *Positional-Independent Caching (PIC)* Project. This project leads to a paper accepted by ICML 2025.
- Co-Lead the *DeepFlow: Serverless LLM at scale* Project. This project leads to a paper accepted by ATC 2025.
- Lead the *Reasoning-Aware Attention Sparsity (RaaS)* Project. This project leads to a paper accepted by ACL 2025.

**WXG of Tencent** **Mar. 2022 – June 2024**
*Research intern mentored by Yuetang Deng and Hailiang Huang*

- Lead the *adaptive distributed build* Project. This project reduces 30% of build costs (of time and hardware resources) and leads to a paper accepted by ASE 2023 industry track.
- Co-lead the *AI-assisted code generation* Project. This project leads to a survey paper accepted by ESEC/FSE 2023.

**Data/AML of ByteDance** **Dec. 2021 – Feb. 2022**
*Intern engineer*

- Rewrite TensorFlow models using ByteDance's internal framework.

## Services

- Co-reviewer of ICSE 2025, ATC 2025, ACL 2025, NSDI 2025, Eurosys 2025, ICSE 2023/2024, ASE 2022
- Deputy Director of New Media Center (新媒体中心学生副主任), School of Computer Science, Peking University. 2024
- Director of the Inter-communication Department (对外联络组组长), Graduate Student Union, School of Computer Science, Peking University. 2023
- Vice Director of the Scientific Research and Innovation Department (学术创新组副组长), Graduate Student Union, School of Computer Science, Peking University. 2022
- Vice President of the Student Union and Director of the Academic Center (学生会副主席兼任学术中心主任), Kuang Yaming Honors School, Nanjing University. 2020

## Selected Awards

- Hunyuan Fellowship (混元学者) 2025
- First Prize of The Challenge Cup (挑战杯), Peking University 2024
- Best Presentation Award, Self-Breakthrough Award, Recognition Award of 2023 Tencent Rhino-Bird Research Elite Program (腾讯犀牛鸟精英人才个人风采奖/突破进取奖/优秀奖), Tencent Inc. 2024
- Ubiquant Scholarship (九坤奖学金), Peking University 2024
- Excellent Research Award (优秀科研奖), Peking University 2024
- Excellent Social Work Award (社会工作奖), Peking University 2023
- Outstanding Graduates (优秀毕业生), Nanjing University 2022
- Pacemaker to Excellent League Member (优秀共青团员标兵), Nanjing University 2022
- Merit Student (江苏省三好学生), Jiangsu Province 2021
- Huawei Scholarship (华为奖学金) 2021
- Special Scholarship for Undergraduates in Basic Science (基础学科奖学金), Nanjing University 2021
- Merit Student (优秀学生), Nanjing University 2019/2020/2021
- China National Scholarship (国家奖学金) 2020
- Elite Program First-class Scholarship (拔尖计划奖学金一等奖) 2019/2020
- Chow Tai Fook Scholarship, First Prize (周大福奖学金一等奖) 2019

## Selected Open-Source Projects

- **C minus compiler** `https://github.com/DerekHJH/compilePA`
  - Write a compiler that translates a subset of the C language into MIPS machine code.
  - Include syntax/semantic error analysis, intermediate code generation, and optimization.
- **miniOS** `https://github.com/DerekHJH/os-workbench`
  - Write a thread-based mini operating system.
  - Include memory allocation, spinning locks, semaphores, a file system, and a FIFO scheduling policy.
  - Include a reconstructor FAT files, and a simple but crash-consistent key-value database.