Junhao Hu

No.5 Yiheyuan Road – Beijing – China

junhaohu@stu.pku.edu.cn • derekhjh.com • DerekHJH

Research Summary

My research centers on building high-performance Large Language Model (LLM) systems, aimed at maximizing both **system efficiency** and **model capabilities** through the principle of algorithm-infrastructure co-design.

In the initial stage of my doctoral studies, I developed intelligent software engineering frameworks to optimize adaptive compilation and code generation (ASE'23, ESEC/FSE'23). Subsequently, my focus shifted to AI infrastructure, where I proposed novel mechanisms for Position-Independent Caching and Reasoning-Aware Attention Sparsity—embodied in systems such as EPIC, CacheSlide, and RaaS—which significantly enhance inference throughput and memory utilization.

My academic research is deeply integrated with industrial practice; at Huawei Cloud, I co-led the development of the serverless LLM serving system DeepServe (ATC'25). Currently, at Xiaomi, I co-led the **co-design of model architecture and infrastructure**, contributing critical optimizations to the MiMo-V2-Flash foundation model regarding **kernel profiling and hardware estimation**.

Looking ahead, I aim to explore the **intersection of system efficiency and model architecture**, specifically targeting optimization for **long-context and reasoning-heavy workloads**.

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

Published Conference/Journal Papers:

[1] Yang Liu, Yunfei Gu, Liqiang Zhang, Chentao Wu, Guangtao Xue, Jie Li, Minyi Guo, **Junhao Hu**, Jie Meng. CacheSlide: Unlocking Cross Position-Aware KV Cache Reuse for Accelerating LLM Serving. In: *Proceedings of the 24th USENIX Conference on File and Storage Technologies, 2026.* (FAST 2026) **CCF-A**

- I contribute to the original idea, experiment design, and writing.

[2] **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**

[3] **Junhao Hu**, Wenrui Huang, Weidong Wang, Haoyi 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**

[4] **Junhao Hu**, Jiang Xu, Zhixia Liu, Yulong He, Yuetao Chen, Hao Xu, Jiang Liu, Jie Meng, Baoquan Zhang, Shining Wan, Gengyuan Dan, Zhiyu Dong, Zhihao Ren, Changhong Liu, Tao Xie, Dayun Lin, Qin Zhang, Yue Yu, Hao Feng, Xusheng Chen, Yizhou Shan. DeepServe: Serverless Large Language Model Serving at Scales In: *Proceedings of the 2025 USENIX Conference on Usenix Annual Technical Conference, 2025.* (ATC 2025) **CCF-A**

[5] 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**

- I contribute to the experiments and writing.

[6] Qingyuan Liang, Zeyu Sun, Qihao Zhu, **Junhao Hu**, Yifan Zhao, Lu Zhang. Cupcleaner: A data cleaning approach for comment updating. In: *Transactions of SCIENCE CHINA Information Sciences, 2025.* (SCIS 2025) **CCF-A**

- I contribute to the experiments and writing.

[7] 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**

- I contribute to half of the experiments and writing.

[8] **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**

Technical Reports:

[1] **Junhao Hu (Core contributor in author list)**, LLM-Core @ Xiaomi. MiMo-V2-Flash Technical Report. In: *GitHub*

- I contribute to the design/profiling/hardware resource estimation of MiMo-V2-Flash model.

[2] **Junhao Hu (Contributor in author list)**, LLM-Core @ Xiaomi. MiMo-Audio: Audio Language Models are Few-Shot Learners. In: *GitHub*

- I contribute to the profiling/optimizations of the underlying inference engine.

[3] **Junhao Hu (Core contributor in author list)**, xDeepServe Team @ Huawei. xDeepServe: Model-as-a-Service on Huawei CloudMatrix384. In: *arXiv preprint arXiv:2508.02520*

- I contribute to the underlying system—DeepServe (ATC 2025) and writing.

Unpublished High-Impact Papers:

[1] **Junhao Hu**, Fangze Li, Mingtao Xu, Feifan Meng, Shiju Zhao, Tiancheng Hu, Ting Peng, Anmin Liu, Wenrui Huang, Chenxu Liu, Ziyue Hua, Tao Xie. Lil: Less is Less When Applying Post-Training Sparse-Attention Algorithms in Long-Decode Stage. In: *arXiv preprint arXiv:2601.03043*

[2] 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*

- I contribute to the KV cache management, experiments, and writing.

[3] 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*

- I contribute to the original idea, experiments and writing. I am the corresponding author.

[4] 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*

- I contribute to the experiments and whole paper writing.

Internship Experiences

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

- Co-lead *Model-Arch and Infra Co-Design* Project, which leads to key sections of the MiMo-V2-Flash Technical Report on kernel profiling, MTP layer selection, and inference hardware estimation.
- Lead the *Efficiency-Aware Context Engineering* Project, which leads to key sections of the MiMo-V2-Flash Technical Report on context management and agent memory.

Huawei Cloud **July 2024 – Aug. 2025**
Research intern mentored by Xusheng Chen

- Lead the *Positional-Independent Caching (PIC)* Project, which leads to a paper accepted by ICML 2025.
- Co-Lead the *DeepServe: Serverless LLM at scale* Project, which leads to a paper accepted by ATC 2025.
- Lead the *Reasoning-Aware Attention Sparsity (RaaS)* Project, which 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, which leads to a paper accepted by ASE 2023.
- Co-lead the *AI-assisted code generation* Project, which leads to a paper accepted by ESEC/FSE 2023.

Data/AML of ByteDance **Dec. 2021 – Feb. 2022**
Engineer intern mentored by Xuan Zou

- Rewrite TensorFlow models using ByteDance's internal framework.

Services

[1] Co-reviewer of ICSE 2025, ATC 2025, ACL 2025, NSDI 2025, Eurosys 2025, ICSE 2023/2024, ASE 2022, ICLR 2026, ACL 2026.

[2] Deputy Director of New Media Center (新媒体中心学生副主任), School of Computer Science, Peking University. 2024

[3] Director of the Inter-communication Department (对外联络组组长), Graduate Student Union, School of Computer Science, Peking University. 2023

[4] 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

[5] 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

[1] 中国科协青年科技人才培育工程博士生专项计划（资助金额4万元），中国科学技术协会 2025

[2] 国家自然科学基金委员会青年学生基础研究项目（博士研究生）（资助金额30万元，北京大学软件方向当年唯一），国家自然科学基金委 2025

[3] Nomination for the Top Ten Academic Achievers Award (学术十杰提名), Peking University 2025

[4] Merit Student (三好学生), Peking University 2025

[5] China National Scholarship (国家奖学金) 2025

[6] Hunyuan Fellowship (混元学者，又名2025年度中国电子学会-腾讯博士生科研激励计划（混元大模型专项），全国仅23人，资助金额10万元) 2025

[7] First Prize of The Challenge Cup (挑战杯), Peking University 2024

[8] 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，全国入围55人，全国唯一包揽三项奖) 2024

[9] Ubiquant Scholarship (九坤奖学金), Peking University 2024

[10] Excellent Research Award (优秀科研奖), Peking University 2024

[11] Excellent Social Work Award (社会工作奖), Peking University 2023

[12] Outstanding Graduates (优秀毕业生), Nanjing University 2022

[13] Pacemaker to Excellent League Member (优秀共青团员标兵), Nanjing University 2022

[14] Merit Student (江苏省三好学生，全院唯一), Jiangsu Province 2021

[15] Huawei Scholarship (华为奖学金) 2021

[16] Special Scholarship for Undergraduates in Basic Science (基础学科奖学金), Nanjing University 2021

[17] Merit Student (优秀学生), Nanjing University 2019/2020/2021

[18] China National Scholarship (国家奖学金) 2020

[19] Elite Program First-class Scholarship (拔尖计划奖学金一等奖), Nanjing University 2019/2020

[20] Chow Tai Fook Scholarship, First Prize (周大福奖学金一等奖) 2019

Selected Open-Source Projects

[1] **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.

[2] **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.