YUAN, BINHANG 袁彬航

Phone: +86 186 5320 6625
Email: biyuan@ust.hk
Site: binhangyuan.github.io

PROFILE

I am an Assistant Professor at the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. My main research focuses are large-scale distributed machine learning systems and data management with foundation models.

EXPERIENCE

Assistant Professor, the Hong Kong University of Science and Technology **2023/08 - Present**

Department of Computer Science and Engineering

Postdoc Research Scientist, ETH Zurich **2021/05 - 2023/05**

System Group, supervised by Dr. Ce Zhang

EDUCATION

Rice University, Houston, TX, USA **2016/08 - 2020/12**

Ph.D., computer science, supervised by Dr. Chris Jermaine

Rice University, Houston, TX, USA **2013/08 - 2016/05**

M.S., computer science, supervised by Dr. Ron Goldman

Fudan University, Shanghai, China **2009/09 - 2013/07**

B.S., Computer Science

AWARDS

2024 TMLR Outstanding Certification

2020 SIGMOD Research Highlight Award

2019 VLDB Best Paper Honorable Mention Award

2011, 2012 Chinese National Scholarship

SELECTED PUBLICATIONS

- Tianyi Bai, Ling Yang, ZhenHao Wong, Fupeng Sun, Xinlin Zhuang, Jiahui Peng, Chi Zhang, Lijun Wu, Jiantao Qiu, Wentao Zhang, **Binhang Yuan**, and Conghui He. "Efficient Pretraining Data Selection for Language Models via Multi-Actor Collaboration." To Appear in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025).
- Youhe Jiang, Fangcheng Fu, Xiaozhe Yao, Guoliang HE, Xupeng Miao, Ana Klimovic, Bin Cui, **Binhang Yuan**, and Eiko Yoneki. "Demystifying Cost-Efficiency in LLM Serving over Heterogeneous GPUs." To Appear in International Conference on Machine Learning. (ICML 2025)
- Suyi Li, Hanfeng Lu, Tianyuan Wu, Minchen Yu, Qizhen Weng, Xusheng Chen, Yizhou Shan, **Binhang Yuan**, and Wei Wang. "Toppings: CPU-Assisted, Rank-Aware Adapter Serving for LLM Inference." To Appear in the 2025 USENIX Annual Technical Conference (ATC 2025).
- Wenjie Qu*, Yuguang Zhou*, Yongji Wu, Tingsong Xiao, **Binhang Yuan**, Yiming Li, and Jiaheng Zhang. "Prompt Inversion Attack against Collaborative Inference of Large Language Models." In IEEE Symposium on Security and Privacy, pp. 1695-1712, 2025 (S&P 2025).
- Jinwei Yao, Kaiqi Chen, Kexun Zhang, Jiaxuan You, **Binhang Yuan**, Zeke Wang, Tao Lin. "DeFT: Decoding with Flash Tree-attention for Efficient Tree-structured LLM Inference" In the 13th International Conference on Learning Representations 2025. (ICLR 2025 *Selected as Spotlight*)
- Youhe Jiang*, Ran Yan*, and **Binhang Yuan**. "HexGen-2: Disaggregated Generative Inference of LLMs in Heterogeneous Environment" In the 13th International Conference on Learning Representations 2025. (ICLR 2025)

- Changyue Liao, Mo Sun, Zihan Yang, Jun Xie, Kaiqi Chen, **Binhang Yuan**, Fei Wu, and Zeke Wang. "Ratel: Optimizing Holistic Data Movement to Fine-tune 100B Model on a Consumer GPU." In the 41st IEEE International Conference on Data Engineering, pp. 292-306, 2025. (ICDE 2025)
- Yongjun He, Roger Waleffe, Zhichao Han, Johnu George, **Binhang Yuan**, Zitao Zhang, Yinan Shan, Yang Zhao, Debojyoti Dutta, Theodoros Rekatsinas, and Ce Zhang. "MLKV: Efficiently Scaling up Large Embedding Model Training with Disk-based Key-Value Storage." In the 41st IEEE International Conference on Data Engineering, pp. 4134-4141. 2025. (ICDE 2025)
- Xinyu Zhao*, Guoheng Sun*, Ruisi Cai*, Yukun Zhou*, Pingzhi Li*, Peihao Wang, Bowen Tan, Yexiao He, Li Chen, Yi Liang, Beidi Chen, **Binhang Yuan**, Hongyi Wang, Ang Li, Zhangyang Wang, Tianlong Chen. “Model-Glue: Democratized LLM Scaling for A Large Model Zoo in the Wild.” In Advances in Neural Information Processing Systems 37 (2024). (NeurIPS 2024)
- Youhe Jiang*, Ran Yan*, Xiaozhe Yao*, Yang Zhou, Beidi Chen, and **Binhang Yuan**. “HexGen: Generative Inference of Large-Scale Foundation Model over Heterogeneous Decentralized Environment.“ In International Conference on Machine Learning (pp. 21946-21961). PMLR. (ICML 2024)
- Lin Lu*, Chenxi Dai*, Wangcheng Tao, **Binhang Yuan**, Yanan Sun, and Pan Zhou “Exploring the Robustness of Pipeline-Parallelism-Based Decentralized Training.” In International Conference on Machine Learning (pp. 32978-32989). PMLR. (ICML 2024)
- Alexandre E Eichenberger, Qi Lin, Saif Masood; Hong Min, Alex Sim, Yida Wang, Kesheng Wu, **Binhang Yuan**, Lixi Zhou, and Jia Zou. “Serving Deep Learning Model in Relational Databases.” In 27th International Conference on Extending Database Technology 2024. EDBT, pp. 717-724. (EDBT 2024)
- Ying Sheng, Lianmin Zheng, **Binhang Yuan**, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. “High-throughput Generative Inference of Large Language Models with a Single GPU.” In International Conference on Machine Learning (pp. 31094-31116). PMLR. (ICML 2023 *Selected as Oral*).
- Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, **Binhang Yuan**, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. “Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time.” In International Conference on Machine Learning (pp. 22137-22176). PMLR. (ICML 2023 *Selected as Oral*).
- Jue Wang, Yucheng Lu, **Binhang Yuan**, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Ré, and Ce Zhang. “CocktailSGD: Fine-tuning Foundation Models over 500Mbps Networks.“ In International Conference on Machine Learning (pp. 36058-36076). PMLR. (ICML 2023)
- Yuxin Tang, Zhimin Ding, Dimitrije Jankov, **Binhang Yuan**, Daniel Bourgeois, and Chris Jermaine. “Auto-Differentiation of Relational Computations for Very Large Scale Machine Learning.” In International Conference on Machine Learning (pp. 33581-33598). PMLR. (ICML 2023)
- **Binhang Yuan***, Yongjun He*, Jared Quincy Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang, Christopher Re, and Ce Zhang. “Decentralized Training of Foundation Models in Heterogeneous Environments.” In Advances in Neural Information Processing Systems 35 (2022), 25464-25477. (NeurIPS 2022 *Selected as Oral*)
- Jue Wang*, **Binhang Yuan***, Luka Rimanic*, Yongjun He, Tri Dao, Beidi Chen, Christopher Re, and Ce Zhang. "Fine-tuning Language Models over Slow Networks using Activation Compression with Guarantees.” In Advances in Neural Information Processing Systems 35 (2022), 19215-19230. (NeurIPS 2022)
- Rui Pan, Yiming Lei, Jialong Li, Zhiqiang Xie, **Binhang Yuan**, and Yiting Xia. "Efficient Flow Scheduling in Distributed Deep Learning Training with Echelon Formation.” In Proceedings of the twenty first ACM Workshop on Hot Topics in Networks (2022). (HotNets 2022)
- Xiangru Lian, **Binhang Yuan**, Xuefeng Zhu, Yulong Wang, Yongjun He, Honghuan Wu, Lei Sun, Haodong Lyu, Chengjun Liu, Xing Dong, Yiqiao Liao, Mingnan Luo, Congfei Zhang, Jingru Xie, Haonan Li, Lei Chen, Renjie Huang, Jianying Lin, Chengchun Shu, Xuezhong Qiu, Zhishan Liu, Dongying Kong, Lei Yuan, Hai Yu, Sen Yang, Ce Zhang, and Ji Liu. “Persia: An Open, Hybrid System Scaling Deep Learning-based Recommenders up to 100 Trillion Parameters.” In Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 3288-3298. 2022 (SIGKDD 2022)
- Lijie Xu, Shuang Qiu, **Binhang Yuan**, Jiawei Jiang, Cedric Renggli, Shaoduo Gan, Kaan Kara, Guoliang Li, Ji Liu, Wentao Wu, Jieping Ye, and Ce Zhang. “In-Database Machine Learning with CorgiPile: Stochastic Gradient Descent without Full Data Shuffle.” In Proceedings of the 2022 International Conference on Management of Data, pp. 1286-1300. 2022. (SIGMOD 2022)
- **Binhang Yuan**, Cameron R. Wolfe, Chen Dun, Yuxin Tang, Anastasios Kyrillidis, and Chris Jermaine. “Distributed Learning of Deep Neural Networks using Independent Subnet Training.” In Proceedings of the VLDB Endowment, 15(8). (VLDB 2022)

- Shaoduo Gan, Xiangru Lian, Rui Wang, Jianbin Chang, Chengjun Liu, Hongmei Shi, Shengzhuo Zhang, Xianghong Li, Tengxu Sun, Jiawei Jiang, **Binhang Yuan**, Sen Yang, Ji Liu, and Ce Zhang. "BAGUA: Scaling up Distributed Learning with System Relaxations." In Proceedings of the VLDB Endowment, 15(4). (VLDB 2022)
- Shangyu Luo, Dimitrije Jankov, **Binhang Yuan**, and Chris Jermaine. "Automatic Optimization of Matrix Implementations for Distributed Machine Learning and Linear Algebra." In Proceedings of the 2021 International Conference on Management of Data(pp. 1222-1234). ACM. (SIGMOD 2021)
- **Binhang Yuan**, Dimitrije Jankov, Jia Zou, Yuxin Tang, Daniel Bourgeois, and Chris Jermaine. "Tensor Relational Algebra for Machine Learning System Design." In Proceedings of the VLDB Endowment, 14(8), 1338-1350 (VLDB 2021)
- Jia Zou, Pratik Barhate, Amitabh Das, Arun Iyengar, **Binhang Yuan**, Dimitrije Jankov, and Chris Jermaine. "Lachesis: Automatic Partitioning for UDF-Centric Analytics." In Proceedings of the VLDB Endowment, 14(8), 1262-1275 (VLDB 2021)
- Dimitrije Jankov, **Binhang Yuan**, Shangyu Luo, and Chris Jermaine. "Distributed Numerical and Machine Learning Computations via Two-Phase Execution of Aggregated Join Trees." In Proceeding of VLDB Endowment, 14(7), 1228-1240. (VLDB 2021)
- Dimitrije Jankov, Shangyu Luo, **Binhang Yuan**, Zhuhua Cai, Jia Zou, Chris Jermaine, and Zekai J Gao. "Declarative recursive computation on an RDBMS: or, why you should use a database for distributed machine learning." In Proceedings of the VLDB Endowment, 12(7), 822-835. (VLDB 2019, *VLDB Best Paper Honorable Mention Award*)
- Jia Zou, R Matthew Barnett, Tania Lorido-Botran, Shangyu Luo, Carlos Monroy, Sourav Sikdar, Kia Teymourian, **Binhang Yuan**, and Chris Jermaine. "PlinyCompute: A platform for high-performance, distributed, data-intensive tool development." In Proceedings of the 2018 International Conference on Management of Data(pp. 1189-1204). ACM. (SIGMOD 2018)
- **Binhang Yuan**, Vijayaraghavan Murali, and Chris Jermaine. "Abridging source code." In Proceedings of the ACM on Programming Languages 1.OOPSLA (2017): 58. (OOPSLA 2017)
- Bo Yan, **Binhang Yuan**, and Bo Yang. "Effective video retargeting with jittery assessment." In IEEE Transactions on Multimedia, Vol. 16, Issue 1, pp. 272-277, Jan. 2014. (TMM 2014)

ACADEMIC SERVICE

- Conference reviewer:
  - AAAI: 2020, 2021;
  - ICLR: 2022, 2023, 2024, 2025;
  - ICML: 2021, 2022, 2023, 2024, 2025;
  - NeurIPS: 2020, 2021, 2022, 2023;
  - MLSys: 2024, 2025.
- Conference area chair:
  - NeurIPS: 2024
- Journal reviewer:
  - IEEE Access: 2020;
  - IEEE TKDE: 2022, 2023;
  - IEEE BigData: 2023;
  - JMLR: 2023;
  - PVLDB: 2023.

OTHER EXPERIENCE

Research Intern, Microsoft Research Asia, Beijing, China **2017/07 - 2017/12**

SDE Intern, Tableau Software, Seattle, WA, USA **2016/05 - 2016/08**

SDE Intern, Isilon EMC2, Santa Clara, CA, USA **2015/05 - 2015/08**