

# Yiwei ZHAO | 赵奕炜

## Curriculum Vitae

Carnegie Mellon University

Office: 41D6 Collaborative Innovation Center, 4720 Forbes Ave, Pittsburgh, PA, 15213

Email: [yiwei3@andrew.cmu.edu](mailto:yiwei3@andrew.cmu.edu) / [yiweizhao@cmu.edu](mailto:yiweizhao@cmu.edu) | Website: <https://zhaoyw007.github.io/>

Last modified in December 2025

## RESEARCH INTERESTS

Cross-Stack Design for Emerging Hardware Systems.

Inter-discipline of: Algorithmic Theory + Computer System Design + Computer Architecture and Hardware.

## EDUCATION

**Carnegie Mellon University**, Pittsburgh, Pennsylvania, 2021-present.

- Ph.D. candidate in Computer Engineering.
- Advisor: Prof. Phillip B. Gibbons.

**Tsinghua University**, Beijing, China, 2017-2021.

- B.E. in Electronic Engineering. Graduated with highest honor.
- Double major in Economics & Finance.

## FELLOWSHIPS, HONORS & AWARDS

- Michel and Kathy Doreau Graduate Fellowship (2024 - present).
- Lee-Stanziale Ohana Fellowship (2023 - 2024).
- Qualcomm PhD Fellowship Finalist (2024).
- Best Paper Runner-up, VLDB (2023) [6].
- Carnegie Mellon Institute of Technology Dean's Fellow (2021 - 2022).
- Undergraduate with Highest Honor, Tsinghua University & Beijing (2021).
- China National Fellowship for Undergrads (2018 - 2020).

## RESEARCH EXPERIENCE

**Carnegie Mellon University (Graduate Research Assistant)**, Pittsburgh, PA. August 2021 – present.

- Thesis Title: Algorithm-System Co-Design of Processing-in-Memory Systems.
- **Foundation of Processing-in-Memory**: Multi-year effort to design provably guaranteed, practically efficient PIM and emerging-hardware systems. Developed index structures [2,5,6] with provably optimal utilization (throughput) and minimal data movement (power) and high real-world performance. Spanning applications across exact nearest-neighbor search in vector databases [1], end-to-end OLTP systems [3], and distributed graph processing [12].

**Meta (Research Scientist Intern)**, Redmond, WA. May 2025 – August 2025; May 2023 – November 2023.

- **AdaVFM** [13] (May 2025 – August 2025): An adaptive on-device inference framework for language-aligned vision foundation models, combining task-aware model scaling, NAS-guided subnet selection, and cloud-assisted multimodal control. State-of-the-art accuracy-efficiency trade-offs for open-vocabulary vision tasks.
- **H4H** [4,10] (May 2023 – November 2023): A two-stage NAS framework that automatically co-optimizes hybrid CNN/ViT models for heterogeneous edge systems combining NPUs and PIM, leveraging architectural heterogeneity and further optimizes hardware and circuits accordingly. State of the art in small on-device vision models.

**Tsinghua University (Undergraduate Researcher)**, Beijing, China, September 2017 – June 2021.

- Thesis: Error Tolerant Designs for ReRAM based Compute-In-Memory Accelerators.

## PUBLICATIONS

### Full Publications

[1] **Yiwei Zhao**, Hongbo Kang, Ziyang Men, Yan Gu, Guy E. Blelloch, Laxman Dhulipala, Charles McGuffey, and Phillip B. Gibbons. 2025. “**PIM-zd-tree: A Fast Space-Partitioning Index Leveraging Processing-in-Memory**”. Accepted by 31st ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (**PPoPP** '26).

[2] **Yiwei Zhao**, Hongbo Kang, Yan Gu, Guy E. Blelloch, Laxman Dhulipala, Charles McGuffey, and Phillip B. Gibbons. 2025. “**Optimal Batch-Dynamic kd-trees for Processing-In-Memory with Applications**”. In Proceedings of the 37th ACM Symposium on Parallelism in Algorithms and Architectures (**SPAA** '25). Association for Computing Machinery, New York, NY, USA, 350–366. [doi:10.1145/3694906.3743318](https://doi.org/10.1145/3694906.3743318).

[3] Hyoungjoo Kim, **Yiwei Zhao**, Andrew Pavlo, and Phillip B. Gibbons. 2025. “**No Cap, This Memory Slaps: Breaking Through the Memory Wall of Transactional Database Systems with Processing-in-Memory**”. In Proceedings of the VLDB Endowment (**PVLDB**), 18(11): 4241-4254, July 2025. [doi:10.14778/3749646.3749690](https://doi.org/10.14778/3749646.3749690).

[4] **Yiwei Zhao**, Jinhui Chen, Sai Qian Zhang, Syed Shakib Sarwar, Kleber Hugo Stangherlin, Jorge Tomas Gomez, Jae Sun Seo, Barbara De Salvo, Chiao Liu, Phillip B. Gibbons, Ziyun Li. 2025. “**H4H: Hybrid Convolution-Transformer Architecture Search for NPU-CIM Heterogeneous Systems for AR/VR Applications**”. In Proceedings of the 30th Asia and South Pacific Design Automation Conference (**ASPDAC** '25). Association for Computing Machinery, New York, NY, USA, 1133–1141. [doi:10.1145/3658617.3697627](https://doi.org/10.1145/3658617.3697627).

[5] Hongbo Kang, **Yiwei Zhao**, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, and Phillip B. Gibbons. 2023. “**PIM-trie: A Skew-Resistant Trie for Processing-in-Memory**”. In Proceedings of the 35th ACM Symposium on Parallelism in Algorithms and Architectures (**SPAA** '23). Association for Computing Machinery, New York, NY, USA, pp. 1–14. [doi:10.1145/3558481.3591070](https://doi.org/10.1145/3558481.3591070).

[6] Hongbo Kang, **Yiwei Zhao**, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, and Phillip B. Gibbons. 2022. “**PIM-tree: A Skew-resistant Index for Processing-in-Memory**”. In Proceedings of the VLDB Endowment (**PVLDB**), 16(4): 946-958, December 2022. [doi:10.14778/3574245.3574275](https://doi.org/10.14778/3574245.3574275). [arXiv:2211.10516](https://arxiv.org/abs/2211.10516). *Best Research Paper Runner-up in VLDB 2023*.

[7] Zeyan Li, Junjie Chen, Yihao Chen, Chengyang Luo, **Yiwei Zhao**, Yongqian Sun, Kaixin Sui, Xiping Wang, Dapeng Liu, Xing Jin, Qi Wang, and Dan Pei. 2023. “**Generic and Robust Root Cause Localization for Multi-Dimensional Data in Online Service Systems**”. In Journal of Systems and Software (**JSS**), Vol. 203, (2023), 111748. [doi:10.1016/j.jss.2023.111748](https://doi.org/10.1016/j.jss.2023.111748).

[8] Zeyan Li, Chengyang Luo, **Yiwei Zhao**, Yongqian Sun, Kaixin Sui, Xiping Wang, Dapeng Liu, Xing Jin, Qi Wang, and Dan Pei. 2019. “**Generic and Robust Localization of Multi-Dimensional Root Cause**”. In the 30th International Symposium on Software Reliability Engineering (**ISSRE** '19). Oct. 28-31, 2019, Berlin. [doi:10.1109/ISSRE.2019.00015](https://doi.org/10.1109/ISSRE.2019.00015).

### Short Publications, Posters & Workshops

[9] **Yiwei Zhao**, Jinhui Chen, Sai Qian Zhang, Syed Shakib Sarwar, Kleber Hugo Stangherlin, Jorge Tomas Gomez, Jae Sun Seo, Phillip B. Gibbons, Barbara De Salvo, Chiao Liu, Ziyun Li. 2025. “**H4H: Hybrid Convolution-Transformer Architecture Search for NPU-CIM Heterogeneous Systems for AR/VR Applications (Abstract)**”. In Proceedings of the 3rd Highlights of Parallel Computing Workshop (**HOPC** '25), July 28, 2025, Portland, OR, USA. [doi:10.1145/3746238.3746241](https://doi.org/10.1145/3746238.3746241).

[10] **Yiwei Zhao**, Ziyun Li, Win-San Khwa, Xiaoyu Sun, Sai Qian Zhang, Syed Shakib Sarwar, Kleber Hugo Stangherlin, Yi-Lun Lu, Jorge Tomas Gomez, Jae Sun Seo, Phillip B. Gibbons, Barbara De Salvo, Chiao Liu. 2024. “**Neural Architecture Search of Hybrid Models for NPU-CIM Heterogeneous AR/VR Devices**”. In 61th ACM/IEEE Design Automation Conference (**DAC '24**), Poster Session, San Francisco, CA, USA, 2024. [arXiv:2410.08326](https://arxiv.org/abs/2410.08326).

[11] Hongbo Kang, **Yiwei Zhao**, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, and Phillip B. Gibbons. 2023. “**PIM-tree: A Skew-resistant Index for Processing-in-Memory (Abstract)**”. In Proceedings of the 2023 ACM Workshop on Highlights of Parallel Computing (**HOPC '23**), June 16, 2023, Orlando, FL, USA. [doi:10.1145/3597635.3598029](https://doi.org/10.1145/3597635.3598029).

## **Preprints**

[12] **Yiwei Zhao**, Qiushi Lin, Hongbo Kang, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, and Phillip B. Gibbons. “**TD-Orch: Scalable Load-Balancing for Distributed Systems with Applications to Graph Processing**”. [arXiv:2511.11843](https://arxiv.org/abs/2511.11843). Under review for OSDI '26.

[13] **Yiwei Zhao**, Yi Zheng, Huapeng Su, Jieyu Lin, Stefano Ambrogio, Cijo Jose, Michael Ramamonjisoa, Patrick Labatut, Barbara De Salvo, Chiao Liu, Phillip B. Gibbons, Ziyun Li. “**AdaVFM: Adaptive Vision Foundation Models for Edge Intelligence via LLM-Guided Runtime Execution**”. Under review for CVPR '26.

[14] Hongbo Kang, Xiangyun Ding, Yingdi Shan, **Yiwei Zhao**, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, Mingxing Zhang, Yongwei Wu, and Phillip B. Gibbons. “**FH-index: A Finger-Hinted Learned Index**”. Under review for VLDB '26.

[15] Mohammad Bakhshalipour, **Yiwei Zhao**, Valerie Choung, Phillip B. Gibbons. “**Benchmarking Memory Systems with a Diverse Zoo**”.

## **TEACHING EXPERIENCES**

- **18-751 Applied Stochastic Processes, with Applications to AI/ML**: Teaching Assistant, Fall 2024, CMU.
- **18-742 Computer Architecture and Systems**: Teaching Assistant, Spring 2024, CMU.

## **INVITED TALKS**

(All conference paper presentations are excluded due to space limit; please refer to the *Publications* for the complete list.)

- **October 2025, UC Berkeley, Simons Institute for the Theory of Computing. In Algorithmic Foundations for Emerging Computing Technologies**: Optimal Semi-Balanced Trees for Processing-in-Memory.
- **October 2025, UC Berkeley. In Programming System Seminar**: Building HPC Systems for Near-Data-Processing: Theory and Practice.
- **October 2025, Stanford University. In Software Research Seminar**: Building Programming Systems for Near-Data-Processing: Theory and Practice.
- **October 2024, CMU. In Parallel Data Lab (PDL) Retreat**: Fast and Principled Techniques for Heterogeneous Compute and Memory.
- **November 2023, CMU. In Parallel Data Lab (PDL) Retreat**: System Design on Processing-In-Memory: Starting from Database Systems.
- **November 2022, CMU. In Parallel Data Lab (PDL) Retreat**: PIM-tree: A Theoretically and Practically Efficient Index for Processing-In-Memory.

## SERVICES

- **SIAM Symposium on Algorithm Engineering and Experiments (ALENEX26)**: Artifact Evaluation Committee.
- **37th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'25)**: Junior Program Committee.
- **2025 International Conference on Parallel Architectures and Compilation Techniques (PACT'25)**: External Review Committee.
- **Student Council for Departmental Faculty Hiring**: Chair, January 2024 – June 2025, CMU.
- **Student Council for Departmental Faculty Hiring**: Member, January 2023 – January 2024, CMU.

## REFERENCES

**Phillip B. Gibbons** (Ph.D. Advisor)

Computer Science Department

Carnegie Mellon University

Pittsburgh, PA 15213

[gibbons@cs.cmu.edu](mailto:gibbons@cs.cmu.edu)

**Guy E. Blelloch**

Computer Science Department

Carnegie Mellon University

Pittsburgh, PA 15213

[guyb@cs.cmu.edu](mailto:guyb@cs.cmu.edu)

**Laxman Dhulipala**

Department of Computer Science

University of Maryland, College Park

College Park, MD 20742

[laxman@umd.edu](mailto:laxman@umd.edu)

**Yan Gu**

Computer Science and Engineering Department

University of California, Riverside

Riverside, CA 92521

[ygu@cs.ucr.edu](mailto:ygu@cs.ucr.edu)

**Andrew Pavlo**

Computer Science Department

Carnegie Mellon University

Pittsburgh, PA 15213

[pavlo@cs.cmu.edu](mailto:pavlo@cs.cmu.edu)