

Xiaochen Wang

Ph.D. Candidate

College of Information Sciences and Technology

The Pennsylvania State University

E348 Westgate Building, University Park, PA 16802

<https://xiaochenwang-psu.github.io/>

EDUCATION

The Pennsylvania State University

Aug. 2022 – May. 2027

Ph.D. in Informatics, supervised by Dr. Fenglong Ma

University Park, PA

University of North Carolina at Chapel Hill

Aug. 2020 – May 2022

M.S. in Information Science, supervised by Dr. Yue Wang

Chapel Hill, NC

Central South University

Sep. 2016 – Jun. 2020

B.E. in Automation (Honors Program), supervised by Dr. Fan Guo

Changsha, China

RESEARCH INTERESTS

My research interests include but not limited to:

- *Multimodal Learning for Healthcare*: [1][6][8][9][10][17][2][23]
- *Multimodal Self-supervised Learning*: [3][5][10][17]
- *Knowledge Grounded AI4Healthcare*: [2][8][9][16][20][22]
- *Natural Language Processing & Large Language Model*: [3][4][5][8][9][11][16][19]
- *Information Retrieval*: [3][4][5][11][14][21]

EXPERIENCE

Amazon

Jun. 2026 – Sep. 2026

Applied Scientist Intern

Seattle, WA

IQVIA

Aug. 2025 – May 2026

Generative AI Scientist Intern

Durham, NC

TikTok

May 2025 – Aug. 2025

Machine Learning Engineer Intern

San Jose, CA

Instacart

May 2023 – Aug. 2023

Machine Learning Scientist Intern

San Francisco, CA

Renaissance Computing Institute

May 2022 – Aug. 2022

Research Scientist

Chapel Hill, NC

PUBLICATIONS

TUTORIAL

AAAI'25 [1] Jiaqi Wang, **Xiaochen Wang**, Yuan Zhong, Ziyi Yin, Aofei Chang, Cao Xiao, Fenglong Ma (2025). *Multimodal Artificial Intelligence in Healthcare*. AAAI 2025 Tutorial and Lab Forum.

- MM'26 [2] **Xiaochen Wang**, Yuan Zhong, Lingwei Zhang, Lisong Dai, Ting Wang, Fenglong Ma (2026). *MedMKG: Benchmarking Medical Knowledge Exploitation with Multimodal Knowledge Graph*. Accepted by ACM Multimedia (MM'26), to appear.
- COLM'26 [3] **Xiaochen Wang**, Yuan Zhong, Haoyu Wang, Ting Wang, Fenglong Ma (2026). *KAMR: Grounding Generation via Knowledge-Aligned Multi-hop Retrieval*. Accepted by Conference on Language Modeling (COLM'26), to appear.
- KDD'26 [4] **Xiaochen Wang**, Bao Hoang, Han Liu, Ting Wang, Fenglong Ma (2026). *MKG-RAG-Bench: Benchmarking Retrieval in Multimodal Knowledge Graph Augmented Generation*. Accepted by KDD'26, to appear.
- WWW'26 [5] **Xiaochen Wang**, Zongyu Wu, Yuan Zhong, Xiang Zhang, Suhang Wang, Fenglong Ma (2026). *GPR: Empowering Generation with Graph-Pretrained Retriever*. In Proceedings of the ACM Web Conference 2026 (pp. 8349-8352).
- KDD'25 [6] Yuan Zhong, **Xiaochen Wang**, Jiaqi Wang, Xiaokun Zhang, Fenglong Ma (2025). *MedDiTPro: A Prompt-Guided Diffusion Transformer for Multimodal Longitudinal Medical Data Synthesis*. In 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025) (pp. 4086-4097).
- AAAI'25 [7] **Xiaochen Wang** (2025). *Developing Multimodal Healthcare Foundation Model: From Data-driven to Knowledge-enhanced*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 28, pp. 29305-29306).
- EMNLP'24 [8] **Xiaochen Wang**¹, Jiaqi Wang*, Houping Xiao, Jinghui Chen, Fenglong Ma (2024). *FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8141-8154).
- NeurIPS'24 (Spotlight) [9] Jiaqi Wang*, **Xiaochen Wang***, Lingjuan Lyu, Jinghui Chen, Fenglong Ma (2024). *FEDMEKI: A Benchmark for Scaling Medical Foundation Models via Federated Knowledge Injection*. Advances in Neural Information Processing Systems 37 (2024): 1082-1116.
- ACL'24 [10] **Xiaochen Wang**, Junyu Luo, Jiaqi Wang, Yuan Zhong, Xiaokun Zhang, Yaqing Wang, Parminder Bhatia, Cao Xiao, Fenglong Ma (2024). *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers) (pp. 3644-3656).
- KDD'24 [11] **Xiaochen Wang**, Xiao Xiao, Ruhan Zhang, Xuan Zhang, Taesik Na, Tejaswi Tenneti, Haixun Wang, Fenglong Ma (2024). *Mitigating Pooling Bias in E-commerce Search via False Negative Estimation*. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (pp. 5917-5925).
- KDD'24 [12] Yuan Zhong, **Xiaochen Wang**, Jiaqi Wang, Xiaokun Zhang, Yaqing Wang, Mengdi Huai, Cao Xiao, Fenglong Ma (2024). *Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models*. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (pp. 4607-4618).

¹Equal Contribution

- IJCAI'24 [13] Jiaqi Wang, Junyu Luo, Muchao Ye, **Xiaochen Wang**, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, Fenglong Ma (2024). *Recent Advances in Predictive Modeling with Electronic Health Records*. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (pp. 8272-8280).
- SIGIR'24 [14] Xiaokun Zhang, Bo Xu, Zhaochun Ren, **Xiaochen Wang**, Hongfei Lin and Fenglong Ma (2024). *Disentangling ID and Modality Effects for Session-based Recommendation*. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (pp. 1883-1892).
- SDM'24 [15] Yuan Zhong, Suhan Cui, Jiaqi Wang, **Xiaochen Wang**, Ziyi Yin, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang, Fenglong Ma (2024). *MedDiffusion: Boosting Health Risk Prediction via Diffusion-based Data Augmentation*. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM) (pp. 499-507). Society for Industrial and Applied Mathematics.
- COLING'24 [16] Luo, Junyu, **Xiaochen Wang**, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma (2024). *CoRelation: Boosting Automatic ICD Coding Through Contextualized Code Relation Learning*. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 3997-4007).
- EMNLP'23 [17] **Xiaochen Wang**, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, Fenglong Ma (2023). *Hierarchical Pretraining on Multimodal Electronic Health Records*. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2839-2852).
- CIKM'23 [18] Junyu Luo, Cheng Qian, **Xiaochen Wang**, Lucas Glass and Fenglong Ma (2023). *pADR: Towards Personalized Adverse Drug Reaction Prediction*. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM) (pp. 4724-4730).
- NAACL'22 [19] **Xiaochen Wang**, Yue Wang (2022). *Sentence-Level Resampling for Named Entity Recognition*. The 2022 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL) (pp. 2151-2165).
- AMIA'22 [20] **Xiaochen Wang**, Yue Wang, José-Luis Ambite, Abhishek Appaji, Howard Lander, Stephen M. Moore, Arcot K. Rajasekar, Jessica A. Turner, Matthew D. Turner, Lei Wang, Satya S. Sahoo (2022). *Enabling Scientific Reproducibility through FAIR Data Management: An ontology-driven deep learning approach in the NeuroBridge Project*. American Medical Informatics Association 2022 Annual Symposium (AMIA) (Vol. 2022, p. 1135).
- CIKM'22 [21] Jingwen Hou*, **Xiaochen Wang***, Jean-Jacques Dubois, R. Byron Rice, Amanda Haddock, Yue Wang (2022). *Resource: Extreme Systematic Reviews: A Large Literature Screening Dataset to Support Environmental Policymaking*. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM) (pp. 4029-4033).

PEER-REVIEWED JOURNAL PAPER

- [22] Lei Wang, Jose Luis Ambite, Abhishek M. Appaji, Janine Bijsterbosch, Jérôme Dockès, Rick Herrick, Alex Kogan, Howard M. Lander, Petra Lenzini, Daniel Marcus, Stephen M. Moore, J- B Poline, Arcot Rajasekar, Satya S. Sahoo, Matthew D. Turner, **Xiaochen Wang**, Yue Wang, Jessica A. Turner(2023). *NeuroBridge: A Prototype Platform for Discovery of The Long-Tail Neuroimaging Data Authors*. Frontiers in Neuroinformatics, 17, 1215261.

SUBMISSIONS

[23] Yuan Zhong, Ziyi Yin, **Xiaochen Wang**, Yuning Xie, Jing Gao, Fenglong Ma (2026). *Synthesizing Longitudinal Multimodal Clinical Data via Staged Adaptation of Any-to-Any Models*. Submitted to AAAI'26.

[24] Yi-Hsiang Chang, Xiaoze Liu, Yi-Chung Chen, Wentao Chen, **Xiaochen Wang**, Junyu Luo, Sipeng Zhang, Tianle Wang, Zilong Wang, Jingbo Shang, Fenglong Ma, Daniel C. Semenza, Jing Gao (2026). *Proto-OKN NIKG: A Neural-Symbolic Knowledge Warehouse for Neighborhood Information*. Submitted to BigData'26.

[25] Zongyu Wu, Yilong Wang, **Xiaochen Wang**, Minhua Lin, Zhichao Xu, Fenglong Ma, Xiang Zhang, Suhang Wang (2026). *Multi-Granularity Context-Enhanced RAG over Multimodal Knowledge Graphs*. Submitted to WSDM'26.

PATENT

Xiaochen Wang, Taesik Na, Xiao Xiao, Ruhan Zhang, Xuan Zhang, Tejaswi Tenneti, Haixun Wang, *False Negative Prediction for Training a Machine-Learning Model*. U.S. Patent No. 12,450,277, granted October 21, 2025.

TEACHING

Teaching Assistant

DS 340W: Applied Data Science

Fall 2022 - Spring 2025

Instructor: Dr. Kaamran Raahemifar

MENTORSHIP

Bao Hoang

B.S. at the Michigan State University

Fall 2025 - Summer 2026

Now Machine Learning Engineer at TikTok

PROFESSIONAL SERVICE

CONFERENCE PROGRAM COMMITTEE MEMBER | REVIEWER

- 2026: ACL, KDD, COLM, ICML, EMNLP
- 2025: AAAI, AISTATS, ICLR, WWW, PAKDD, ACL, KDD, IJCAI, EMNLP, IEEE BigData
- 2024: AAAI, AISTATS, EMNLP, AMIA, NAACL
- 2023: CIKM, AAAI

JOURNAL ARTICLE REVIEWER

- Progress in Artificial Intelligence
- Heliyon
- Human-centric Computing and Information Sciences
- Artificial Intelligence In Medicine

VOLUNTEER

- 2024: KDD, EMNLP

TALKS & PRESENTATIONS

- [1] *GPR: Empowering Generation with Graph-Pretrained Retriever*
Poster Presentation at WWW 2026
- [2] *Tuning LLM for Enhanced ICF Generation*
Oral Presentation at IQVIA
- [3] *Machine Learning-Based Prediction of HIV Care Loss to Follow-Up*
Oral Presentation at Penn State Institute of Computational and Data Sciences
- [4] *Adaptive Federated Knowledge Injection into Medical Foundation Models*
Invited Talk at Stanford University
- [5] *Developing Multimodal Healthcare Foundation Model: From Data-driven to Knowledge-enhanced*
Poster Presentation at AAAI 2025
- [6] *Multimodal Artificial Intelligence in Healthcare*
Tutorial at AAAI 2025
- [7] *FEDKIM: Adaptive Federated Knowledge Injection into Medical Foundation Models*
Poster Presentation at EMNLP 2024
- [8] *Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources*
Poster Presentation at ACL 2024
- [9] *Mitigating Pooling Bias in E-commerce Search via False Negative Estimation*
Oral & Poster Presentation at KDD 2024
- [10] *Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models*
Oral & Poster Presentation at KDD 2024
- [11] *Hierarchical Pretraining on Multimodal Electronic Health Records*
Poster Presentation at EMNLP 2023
- [12] *Boosting Search Embedding via Cross Encoder*
Oral Presentation at Instacart
- [13] *Sentence-Level Resampling for Named Entity Recognition*
Poster presentation at NAACL 2022
- [14] *Enabling Scientific Reproducibility through FAIR Data Management: An ontology-driven deep learning approach in the NeuroBridge Project*
Oral Presentation at AMIA 2022

SELECTED HONORS

GRANT

- *Machine Learning-Based Prediction of HIV Care Loss to Follow-Up*
2025-2026 ICDS Rising Researcher Grant at Penn State (RRID:SCR_02515) \$13,000

AWARDS

- Student Travel Award for KDD 2024
- Student Travel Award for EMNLP 2024

- SIGIR Student Travel Grant for CIKM 2022
- IST Graduate Programs Travel Award for AMIA 2022

MEDIA COVERAGE

- *Can Machine Learning Improve Air Quality?*

SILS, UNC