

ZIHAO XU

Homepage, Google Scholar, Github
Amazon Web Service (AWS)
shsjxzh98@gmail.com
+1 848 313 5495

EDUCATION

- **Rutgers University** New Brunswick, NJ
Ph.D. in Computer Science and Technology
Advised by Prof. [Hao Wang](#) Sep. 2021 – May. 2026
- **Rutgers University** New Brunswick, NJ
M.S. in Computer Science and Technology
Advised by Prof. [Hao Wang](#) Sep. 2021 – Feb. 2024
- **Shanghai Jiao Tong University** Shanghai, China
B.E. in Computer Science and Technology
[ACM Honored Class, Zhiyuan College](#)
Advised by Prof. [Hongtao Lu](#) Sep. 2016 – Jun. 2020

RESEARCH INTEREST

- My research interests span Large Language Models (LLMs), Domain Adaptation, Recommendation Systems, and Bayesian Deep Learning. My current focus is on token compression and exploring the cognitive capabilities of LLMs.

PUBLICATION

Domain-Indexing Variational Bayes: Interpretable Domain Index for Domain Adaptation

Zihao Xu*, Guang-Yuan Hao*, Hao He, Hao Wang

International Conference on Learning Representations (ICLR), 2023 (Spotlight)

Implicit In-context Learning

Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, Dimitris N. Metaxas

International Conference on Learning Representations (ICLR), 2025

Continual Learning of Large Language Models: A Comprehensive Survey

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wen Yuan Wang, Yibin Wang, Hao Wang

ACM Computing Surveys, 2024

Taxonomy-Structured Domain Adaptation

Tianyi Liu*, Zihao Xu*, Hao He, Guang-Yuan Hao, Guang-He Lee, Hao Wang

International Conference on Machine Learning (ICML), 2023

Probabilistic Residual Learning for Online Recommendations

Wen Yuan Wang, Yusong Zhao, Zihao Xu, Hengyi Wang, Qi Xu, Zhigang Hua, Yan Xie, Yi Wang, Zihao Zhao, Bo Long, Chengzhi Mao, Shuang Yang, Hengguan Huang, Hao Wang

ACM Conference on Recommender Systems (RecSys), 2026

Graph-Relational Domain Adaptation

Zihao Xu, He Hao, Guang-He Lee, Yuyang Wang, Hao Wang

International Conference on Learning Representations (ICLR), 2022

Domain Indexing Collaborative Filtering for Recommender Systems

Rohit Amarnath*, Zihao Xu*, Qi Xu, Zhigang Hua, Yan Xie, Shuang Yang, Bo Long, Hao Wang

Transactions on Machine Learning Research (TMLR), 2026

Rate-My-LoRA: Efficient and Adaptive Federated Model Tuning for Cardiac MRI Segmentation

Xiaoxiao He, Haizhou Shi, Ligong Han, Chaowei Tan, Bo Liu, **Zihao Xu**, Meng Ye, Leon Axel, Kang Li, Dimitris Metaxas

International Symposium on Biomedical Imaging (ISBI), 2025

GenVP: Generating Visual Puzzles with Contrastive Hierarchical VAEs

Kalliopi Basioti, Pritish Sahu, Tony Qingze Liu, **Zihao Xu**, Hao Wang, Vladimir Pavlovic

International Conference on Learning Representations (ICLR), 2025

Knowledge Graph-Enhanced Retrieval Augmented Generation for E-Commerce

Zihao Xu, Petar Ristoski, Qunzhi Zhou

RAGE-KG workshop for the 23rd International Semantic Web Conference

Towards a Generalized Bayesian Model of Reconstructive Memory

Zihao Xu, Pernille Hemmer, and Qiong Zhang

Computational Brain & Behavior, 2024

“*” indicates equal contribution.

PROFESSIONAL SERVICE

- Reviewer for European Conference on Computer Vision (ECCV) 2026
- Reviewer for Transactions on Machine Learning Research (TMLR) 2026
- Reviewer for Conference on Computer Vision and Pattern Recognition (CVPR) 2025, 2026
- Reviewer for AAAI Conference on Artificial Intelligence (AAAI) 2025, 2026
- Reviewer for International Conference on Machine Learning (ICML) 2026, 2025, 2024
- Reviewer for International Joint Conference on Artificial Intelligence (IJCAI) 2024
- Reviewer for International Conference on Learning Representations (ICLR) 2026, 2025, 2024
- Reviewer for Conference on Neural Information Processing Systems (NeurIPS) 2023
- Reviewer for International Conference on Computer Vision (ICCV) 2023
- Reviewer for ICLR 2022 Workshop PAIR2Struct 2022

RESEARCH EXPERIENCE

Amazon Web Service (AWS) **San Jose, CA**

Applied Research Scientist June. 2025 – Present

- Companies have an emerging need to modernize their IT services, and evaluating such migrations is key to ensuring that the process is robust, sustainable, and professional. The development of large language models (LLMs) has created new opportunities for this type of evaluation. My work focuses on developing a systematic evaluation framework for modern database modernization, leveraging LLMs and agentic workflows.

Rutgers University **New Brunswick, NJ**

Graduate Research Assistant Sep. 2021 – May. 2026

- Conduct research in Large Language Models (LLMs), domain adaptation, recommendation systems, and Bayesian deep learning. Supervised by Prof. Hao Wang.

Magnit @ Meta **Remote**

Academic Collaborator Sep. 2025 – Apr. 2026

- Coevolving the code generation and code evaluation abilities of large language models (LLMs) has attracted increasing interest. In this project, we propose a coevolutionary reinforcement learning framework that jointly enhances both capabilities, enabling LLMs to more effectively solve optimization problems such as the Traveling Salesman Problem (TSP). This work is supervised by Zhigang Hua and Qi Xu.

Amazon Web Service (AWS)**San Jose, CA**

Applied Research Intern

May. 2025 – Aug. 2025

- Long contexts pose significant challenges for Large Language Models (LLMs). We propose an (Almost) Lossless Token Sequence Compression method that reduces input token length by up to 75% without sacrificing performance. Supervised by John Harvill and Ziwei Fan.

Magnit @ Meta**Remote**

Academic Collaborator

Sep. 2024 – Mar. 2025

- Modern recommender systems often use deep learning (DL) models to encode user and item data. However, the complexity and opacity of these models can obscure their workings and limit enhancements. To overcome these issues, we introduce Probabilistic Residual User Clustering (PRUC), a causal Bayesian model that clusters users to improve recommendations. Supervised by Zhigang Hua and Qi Xu.

eBay**San Jose, CA**

Applied Research Intern

May. 2024 – Aug. 2024

- Large Language Models (LLMs) face challenges in e-commerce tasks due to inaccessible, frequently updated proprietary data. Fine-tuning with this data is costly and time-consuming. We propose a knowledge graph-enhanced retrieval augmented generation approach for e-commerce. Our KG, derived from eBay's inventory data, captures product relationships, enabling our framework to outperform baseline LLMs on three e-Commerce tasks in zero-shot and instruction-tuned settings. Supervised by Petar Ristoski and Qunzhi Zhou.

Amazon AI Lab**Shanghai, China**

Research Intern

Aug. 2020 – Aug. 2021

Traditional Domain Adaptation (DA) methods treat all domains equally, ignoring their inherent heterogeneity that can be modeled using graphs. In this project, we proposed a graph-based domain adaptation method, resulting in the publication "Graph-relational Domain Adaptation" at ICLR 2022. Our approach extends adversarial learning by introducing a novel discriminator that models encoding-conditioned graph embeddings, theoretically yielding graph-invariant features. Experiments on AWS SageMaker validated our method on both synthetic and real datasets. Supervised by Prof. Hao Wang and Bernie Wang.

Pennsylvania State University – College of Information Sciences & Technology**University Park, PA**

Research Intern

Jun. 2019 – Dec. 2019

- An Imitation-Learning-based method is adopted for the training of 3D object localizer, to see if action feedback can serve as a supervised signal. In a virtual environment, we trained the robot agent to navigate to certain objects (like chair) in the fewest steps. We made this pipeline differentiable, thus incorporating an imitation learning framework where agents are trained by expert trajectory. Directed by Prof. [Zihan Zhou](#).

CODING LANGUAGE

- Python: Proficient
- C++: Familiar
- Java: Familiar
- Matlab: Familiar

SELECTED AWARDS AND HONORS

- University candidate of Apple Scholarship

2023

- ICLR Travel Award 2023
- SMC Scholarship 2018
- Eleme (饿了么) Scholarship 2017
- Zhiyuan Honors Scholarship 2018, 2017
- Academic Excellence Scholarship (Third-Class) 2018, 2017