HAOYUE BAI

University of Wisconsin-Madison
Computer Sciences Department
Homepage, Google Scholar (1.5k+ citations)
Twitter (1.1k+ followers in AI/ML)
Mobile: (608) 867-0117
Email: baihaoyue@cs.wisc.edu
U.S. Work Authorization: EB-1A (I-140) approved
Priority Date: [Dec 2023]

RESEARCH INTEREST

My research focuses on developing theoretical foundations and algorithms for reliable machine learning, trustworthy foundation models, and related applications:

• Out-of-distribution (OOD) learning and open-world robustness: Designing adaptive and interpretable learning principles that help ML models detect and generalize under distribution shifts, such as semantic and covariate shifts, correlation and diversity shifts.

• Reliable algorithms with provable guarantees: Developing machine learning techniques with statistical and algorithmic guarantees to ensure reliable deployment of AI systems under real-world distribution shifts.

• Safety and reliability of foundation models: Understanding the failure modes and boundaries of large language models (LLMs) and vision language models (VLMs) through systematic diagnostics, such as failure analysis in open-world reasoning, robustness under distribution shift, and generalization. Developing methods to strengthen their reliability and improve quality of learned representations and embeddings.

• Data-efficient machine learning: Selecting the most informative data and signals (e.g., human or model feedback) to improve robustness, calibration, and coverage under tight annotation and compute budgets.

EDUCATION

Ph.D. in Computer Sciences, University of Wisconsin–Madison — Aug. 2022 – Present
• Advisor: Prof. Robert Nowak — Madison, WI

• Ph.D. research in reliable machine learning

• Committee Members: Profs. Yong Jae Lee, Fred Sala, Tengyang Xie

Research Visiting Student, EECS, University of California–Berkeley — May. 2025 – Oct. 2025
• Advisor: Prof. Dawn Song — Berkeley, CA

• Research in fine-grained analysis of LLM reasoning generalization

MPhil, CSE, The Hong Kong University of Science and Technology — Aug. 2018 – Jun. 2022
• Advisor: Prof. Shueng-Han Gary Chan — Hong Kong

• Thesis: Towards Reliable Out-of-Distribution Generalization

• Committee Members: Profs. Andrew Horner, Dan Xu

B.Eng., Information Science & Electronic Engineering, Zhejiang University — Sep. 2014 – Jun. 2018
• Overall GPA: 3.87/4.0 — Hangzhou, China

• Honors: First-class Academic Scholarship, Outstanding Graduates

AWARDS

- **OpenAI Superalignment Fellowship** – US$150,000 personal research funding
  (∼**1% acceptance rate**; **50 fellows worldwide** across faculty and PhD students) 2024-present
- **OpenAI Researcher Access Program Recipient** – US$5,000 API credits support 2024
- **CVPR Travel Support Award** 2025
- **Research Travel Grant**, The Hong Kong University of Science and Technology 2019
- **Outstanding Graduates Award**, Zhejiang University 2018
- **Alibaba–Zhejiang News Scholarship** 2017
- **Zhejiang University Scholarship & Awards**: First-Prize Academic Scholarship (Top 3%), First-Class Scholarship for Outstanding Merits (Top 3%), Outstanding Social Worker Scholarship (Top 2%), Outstanding Student Leader Award (Top 6%) 2016-2017

PUBLICATION LIST

(See the full list on my Google Scholar page; * indicates equal contribution.)

[1] DELTA: How Does RL Unlock and Transfer New Algorithms in LLMs
Yiyou Sun, Yuhan Cao, Pohao Huang, **Haoyue Bai**, Hannaneh Hajishirzi, Houha Dziri, Dawn Song
International Conference on Learning Representations (**ICLR**), 2026

[2] Towards Text-Guided Attribute-Disentangled Multimodal Representation Learning
Yibing Wei, Sudeep Katakol, Manuel Brack, Jinhong Lin, **Haoyue Bai**, Yu-Teng Li, Richard Zhang, Eli Shechtman, Hareesh Ravi, Ajinkya Kale
IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) findings 2025

[3] OoDBench+: Quantifying and Understanding Two Dimensions of Out-of-Distribution Generalization
Nanyang Ye, Kaican Li, **Haoyue Bai**, Lanqing Hong, Fengwei Zhou, Zhenguo Li, Jun Zhu.
IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2025

[4] Where is the Liability in the Generative Era? Recovery-based AI Content Generation Detection
Haoyue Bai, Yiyou Sun, Cheng Wei, Haifeng Wang
IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2025

[5] Climbing the Ladder of Reasoning: What LLMs Can—and Still Can't—Solve after SFT?
Yiyou Sun, Georgia Zhou, **Haoyue Bai**, Hao Wang, Dacheng Li, Nouha Dziri, Dawn Song
Neural Information Processing Systems (**NeurIPS**) MATH-AI, 2025

[6] Deep Active Learning in the Open World
Tian Xie, Jifan Zhang, **Haoyue Bai**, Robert Nowak
Transactions on Machine Learning Research (**TMLR**), 2025

[7] Out-of-Distribution Learning with Human Feedback
Haoyue Bai, Xuefeng Du, Katie Rainey, Shibin Parameswaran, Yixuan Li
Transactions on Machine Learning Research (**TMLR**), 2025

[8] Adaptive Human-Assisted Out-of-Distribution Generalization and Detection
Haoyue Bai, Jifan Zhang, Robert Nowak
Neural Information Processing Systems (**NeurIPS**), 2024

[9] Provable Out-of-Distribution Generalization in Hypersphere
Haoyue Bai*, Yifei Ming*, Julian Katz-Samuels, Yixuan Li.
International Conference on Learning Representations (**ICLR**), 2024

[10] Feed Two Birds with One Scone: Exploiting Wild Data for Both OOD Generalization and Detection
Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert Nowak, Yixuan Li.
International Conference on Machine Learning (**ICML**), 2023

[11] Benchmarking and Understanding Out-of-Distribution Generalization Datasets and Algorithms
Nanyang Ye, Kaican Li, **Haoyue Bai**, Lanqing Hong, Fengwei Zhou, Zhenguo Li, Jun Zhu.
IEEE Conference on Computer Vision and Pattern Recognition (**CVPR oral**), 2022

[12] Out-of-Distribution Generalization via Feature Decomposition and Semantic Augmentation
Haoyue Bai*, Rui Sun*, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S.-H. Gary Chan, Zhenguo Li
AAAI Conference on Artificial Intelligence (**AAAI**), 2021

[13] NAS-OoD: Neural Architecture Search for Out-of-Distribution Generalization
Haoyue Bai, Fengwei Zhou, Lanqing Hong, Nanyang Ye, S.-H. Gary Chan, Zhenguo Li
International Conference on Computer Vision (**ICCV**), 2021

[14] Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection
Jiageng Mao, Minzhe Niu, **Haoyue Bai**, Xiaodan Liang, Hang Xu, Chunjing Xu
International Conference on Computer Vision (**ICCV**), 2021

[15] Voxel Transformer for 3D Object Detection
Jiageng Mao, Yujing Xue, Minzhe Niu, **Haoyue Bai**, Jiashi Feng, Xiaodan Liang, Hang Xu, Chunjing Xu
International Conference on Computer Vision (**ICCV oral**), 2021

[16] A Survey on Single Image Crowd Counting: Network Design, Loss Function and Supervisory Signal
Haoyue Bai, S.-H. Gary Chan
Neurocomputing, 2022

[17] CounTr: A Novel Transformer Approach for Image-based Crowd Counting
Haoyue Bai*, Hao He*, Zhuoxuan Peng, Tianyuan Dai, S.-H. Gary Chan
European Conference on Computer Vision (**ECCV**) *workshop*, 2022

[18] Crowd Counting on Images with Scale Variation and Isolated Clusters
Haoyue Bai, Song Wen, S.-H. Gary Chan
International Conference on Computer Vision (**ICCV**) *workshop*, 2019

PREPRINTS

[1] Why Do Long-Horizon LLM Agents Fail? A Systematic Study of Failures, Patterns, and Limits
Haoyue Bai*, Xinyu Jessica Wang*, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, Robert D Nowak
under review, *Conference on Language Modeling* (**COLM**), 2026

[2] Expert-Choice Routing Enables Adaptive Computation in Diffusion Language Models
Shuibai Zhang, Caspian Zhuang, Chihan Cui, Zhihan Yang, Fred, Yanxin Zhang, **Haoyue Bai**, Zack Jia, Yang Zhou, Guanhua Chen, Ming Liu
under review, *Conference on Language Modeling* (**COLM**), 2026

[3] How and Why LLMs Generalize: LLM Reasoning from Cognitive Behaviors to Low-Level Patterns
Haoyue Bai, Yiyou Sun, Wenjie Hu, Shi Qiu, Maggie Ziyu Huan, Peiyang Song, Robert D Nowak, Dawn Song
under review, ACL Rolling Review, 2026

[4] Unknown Aware AI-Generated Content Detection
Ellie Thieu, Jifan Zhang, **Haoyue Bai** (*correspondence author)
arXiv 2025

[5] Self-Authentication Mechanism for Generative AI: Can Models Identify Their Own Outputs?
Wei Cheng, Xianjun Yang, **Haoyue Bai**, Yiyou Sun, Cong Zeng, Shengkun Tang, Yuanzhou Chen, Yue Wu, Xiaojie Zhao, Wenchao Yu, Dongkuan Xu, William Yang Wang, Linda Ruth Petzold, Haifeng Chen
under review, Nature Machine Intelligence 2025

[6] Improving Out-of-Distribution Robustness of Classifiers through Interpolated Generative Models
Haoyue Bai, Ceyuan Yang, Yinghao Xu, S.-H. Gary Chan, Bolei Zhou
arXiv, 2022

Mentored Publications and Student Collaborations

(Student papers where I served in a mentoring or advising role.)

[1] Why Do Long-Horizon LLM Agents Fail? A Systematic Study of Failures, Patterns, and Limits
Haoyue Bai*, Xinyu Jessica Wang*, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, Robert D Nowak
under review, *International Conference on Machine Learning* (**ICML**), 2026

[2] Deep Active Learning in the Open World
Tian Xie, Jifan Zhang, **Haoyue Bai**, Robert Nowak
Transactions on Machine Learning Research (**TMLR**), 2025

[3] CounTr: A Novel Transformer Approach for Image-based Crowd Counting
Haoyue Bai*, Hao He*, Zhuoxuan Peng, Tianyuan Dai, S.-H. Gary Chan
European Conference on Computer Vision (**ECCV**) *workshop*, 2022

[4] Unknown Aware AI-Generated Content Detection
Ellie Thieu, Jifan Zhang, **Haoyue Bai**
under review, IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) 2025

[5] Towards Text-Guided Attribute-Disentangled Multimodal Representation Learning
Yibing Wei, Sudeep Katakol, Manuel Brack, Jinhong Lin, **Haoyue Bai**, Yu-Teng Li, Richard Zhang, Eli Shechtman, Hareesh Ravi, Ajinkya Kale
IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**) findings 2025

Service

- **Reviewer for**

ICML, NeurIPS, ICLR

CVPR, ECCV, ICCV

AAAI, IJCAI, WACV

TPAMI, TNNLS, TIP

- **University Service**

Graduate Student Mentor at UW-Madison

ACM's Women in Computing Mentor at UW-Madison

- **Seminar Organization**

Core Student Organizer of the Systems, Information, Learning and Optimization (SILO) Seminar

- **Workshop Organization**

Co-organizer of NeurIPS 2025 Socially Responsible and Trustworthy Foundation Models Workshop.

Students and Mentees

Xinyu Wang, PhD student at UW-Madison, now PhD student at UW-Madison.

Yibing Wei, PhD student at UW-Madison, now PhD student at UW-Madison.

Ellie Thieu, PhD student at UW-Madison, now PhD student at UW-Madison.

Tian Xie, master student at UW-Madison, 2024, now PhD student at ASU.

Tianyuan Dai, undergraduate student at HKUST, 2022, now master student at Stanford.

Jiachen Zhao, undergraduate student at HKUST, 2021, now PhD student at Northeastern University.

Xiaoyuan Ni, undergraduate student at HKUST, 2021, now master student at Stanford.

TALKS

Invited Talk at SJTU University.

Invited Talk at MLOPT Seminar, UW-Madison.

Talk at NEC Laboratories America, 07/2024.

Talk at Genentech, 02/2024.

Talk at Snowflake, 01/2024.

TEACHING

University of Wisconsin–Madison
COMP SCI CS540: Introduction to Artificial Intelligence — Fall 2025

University of Wisconsin–Madison
COMP SCI CS400: Data Science Programming III — Fall 2024

University of Wisconsin–Madison
COMP SCI CS220: Data Science Programming I — Fall 2022

The Hong Kong University of Science and Technology
COMP2611: Computer Organization — Fall 2019, Spring 2019

RESEARCH EXPERIENCE

Meta Superintelligence Labs — May 2026 – Aug. 2026
• Research Scientist Intern, Host: Dr. Andrey Zhmoginov — Seattle

• Research on memory in agent systems

University of California, Berkeley — May 2025 – Sep. 2025
• Graduate Visiting Researcher, Host: Prof. Dawn Song — Berkeley

• Research on fine-grained analysis of LLM reasoning generalization

NEC Laboratories America — May 2024 – Aug. 2024
Research Scientist Intern, Mentor: Dr. Yiyou Sun and Dr. Wei Cheng — Princeton

• Research on recovery-based black-box AI-generated content detection

University of Wisconsin-Madison — Sep 2022 – Sep. 2023
Research Assistant, Mentor: Prof. Sharon Li — Madison

• Research on provable out-of-distribution generalization and detection

The Chinese University of Hong Kong — Jun. 2021 – Sep. 2021
Research Assistant, Advisor: Prof. Bolei Zhou — Hong Kong

• Research on improving OOD robustness of classifiers via interpolated generative models

AI Theory Group, Noah's Ark Lab — Jun. 2020 – May 2021
Research Scientist Intern, Mentor: Dr. Nanyang Ye, Dr, Fengwei Zhou, Dr. Lanqing Hong — Hong Kong

• Research on neural architecture search for out-of-distribution generalization

Hong Kong University of Science and Technology — Jan. 2018 – May. 2018
Visiting Student, Advisor: Prof. Shueng-Han Gary Chan — Hong Kong

• Research on pedestrian tracking and behavior analysis in ultra-crowded scenes

Hong Kong University — Jul. 2017 – Aug. 2017
Research Assistant, Advisor: Prof. Lucas C.K. Hui — Hong Kong

• Research on exploring blockchain security through cryptographic techniques