Furong Huang

Associate Professor, University of Maryland

furongh@umd.edu • +1 301 405 8010 • furong-huang.com • 4124 Brendan Iribe Center

I certify that this curriculum vitae is a current and accurate statement of my professional record. Date: August 1, 2026

Signature: 

PERSONAL INFORMATION

Education

Doctor of Philosophy, Electrical and Computer Engineering
University of California, Irvine 09/2010-06/2016
Advisor: Anima Anandkumar

Bachelor of Science, Electrical Engineering and Computer Science
Zhejiang University, Zhejiang, China 09/2006-06/2010

Professional Experience

University of Maryland 07/2024-present
Associate Professor, Department of Computer Science (CS)
Associate Professor, Institute for Advanced Computer Studies (UMIACS)
Associate Professor, Maryland Robotics Center (MRC)
Affiliate, Applied Mathematics, Statistics, and Computation (AMSC)
Affiliate, Department of Electrical and Computer Engineering (ECE)

University of Maryland 07/2017-06/2024
Assistant Professor, Department of Computer Science (CS)
Assistant Professor, Institute for Advanced Computer Studies (UMIACS)
Affiliate, Applied Mathematics, Statistics, and Computation (AMSC)
Affiliate, Department of Electrical and Computer Engineering (ECE)

Microsoft Research, New York City 07/2016-07/2017
Postdoctoral Researcher
Mentors: John Langford, Robert Schapire

Microsoft Research, New England 06/2014-12/2014
Research Intern
Mentors: Jennifer Chayes, Christian Borgs

Microsoft Research, Cloud and Information Services Lab, Seattle 03/2014-05/2014
Research Intern
Mentors: Paul Mineiro, Nikos Karampatziakis

SCHOLARLY ACTIVITIES

Preprints and Manuscripts

P24 Liu, Haowen, Xirui Li, Shaoxiong Yao, Peng Shi, Tianyi Zhou, Jia-Bin Huang, Furong Huang, and Jiayuan Mao. “Guava: An Effective and Universal Harness for Embodied Manipulation.” arXiv, 2026.

P23 Lee, Seungjae, Yoonkyo Jung, Jusuk Lee, Jonghun Shin, Amir Hossein Shahidzadeh, Yao-Chih Lee, H. Jin Kim, Jia-Bin Huang, and Furong Huang. “μ0: A Scalable 3D Interaction-Trace World Model.” arXiv. Project page. Code. Models, 2026.

P22 Yuan, Lingzhi, Chenghao Deng, Fangxu Yu, Souradip Chakraborty, Mohammad Rostami, and Furong Huang. “FlowBank: Query-Adaptive Agentic Workflows Optimization through Precompute-and-Reuse.” arXiv, 2026.

P21 Kini, Prajakta, Avinash Reddy, Souradip Chakraborty, Satya Sai Srinath Namburi GNVV, Furong Huang, Amrit Singh Bedi, and Alvaro Velasquez. “Does Reasoning Preserve Alignment? On the Trustworthiness of Large Reasoning

Models." arXiv, 2026.

P20 Agrawal, Aakriti, Souradip Chakraborty, Armin Saghafian, Nihal Sharma, Rizal Fathony, Nam H. Nguyen, C. Bayan Bruss, Amrit Singh Bedi, and Furong Huang. "The Hidden Bias of Process Reward Models: PRISM for Rewarding the Right Reasoning." arXiv, 2026.

P19 Agrawal, Aakriti, Minghui Liu, and Furong Huang. "VeriGate: Verifier-Gated Step-Level Supervision for GRPO." arXiv, 2026.

P18 Lee, Jusuk, Seungjae Lee, Jonghun Shin, Hoseong Jung, Sungha Kim, Daesol Cho, H. Jin Kim, Jia-Bin Huang, and Furong Huang. "DynaFLIP: Rethinking Robotics Perception via Tri-Modal-Dynamics Guided Representation." arXiv, 2026.

P17 Ho, Sy-Tuyen, Minghui Liu, Huy Nghiem, and Furong Huang. "SoundnessBench: Can Your AI Scientist Really Tell Good Research Ideas from Bad Ones?" arXiv, 2026.

P16 Ankireddy, Sravan Kumar, Nikita Seleznev, Nam H. Nguyen, Yulun Wu, Senthil Kumar, Furong Huang, and C. Bayan Bruss. "TimeSqueeze: Dynamic Patching for Efficient Time Series Forecasting." arXiv, 2026.

P15 Liu, Weize, Minghui Liu, Sy-Tuyen Ho, Souradip Chakraborty, Xiyao Wang, and Furong Huang. "Agentic Critical Training." arXiv. Project page, 2026.

P14 Stein, Alex, Furong Huang, and Tom Goldstein. "GATES: Self-Distillation under Privileged Context with Consensus Gating." arXiv, 2026.

P13 Zheng, Ruijie, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. "EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data." arXiv, 2026.

P12 Satheesh, Anirudh, Ziyi Chen, Furong Huang, and Heng Huang. "Provably Efficient Algorithms for S- and Non-Rectangular Robust MDPs with General Parameterization." arXiv, 2026.

P11 Yuan, Dehao, Tyler Farnan, Stefan Tesliuc, Doron L. Bergman, Yulun Wu, Xiaoyu Liu, Minghui Liu, James Montgomery, Nam H. Nguyen, C. Bayan Bruss, and Furong Huang. "PersonaLedger: Generating Realistic Financial Transactions with Persona Conditioned LLMs and Rule Grounded Feedback." arXiv, 2026.

P10 Pathmanathan, Pankayaraj, Michael-Andrei Panaitescu-Liess, Cho-Yu Jason Chiang, and Furong Huang. "RAGPart & RAGMask: Retrieval-Stage Defenses Against Corpus Poisoning in Retrieval-Augmented Generation." arXiv, 2025.

P9 Liu, Minghui, Aadi Palnitkar, Tahseen Rabbani, Hyunwoo Jae, Kyle Rui Sang, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang. "Hold Onto That Thought: Assessing KV Cache Compression on Reasoning." arXiv, 2025.

P8 Chen, Yuen, Yulun Wu, Samuel Sharpe, Igor Melnyk, Nam H. Nguyen, Furong Huang, C. Bayan Bruss, and Rizal Fathony. "Bridging the Divide: End-to-End Sequence-Graph Learning." arXiv, 2025.

P7 Zhai, Kevin, Utsav Singh, Anirudh Thatipelli, Souradip Chakraborty, Anit Kumar Sahu, Furong Huang, Amrit Singh Bedi, and Mubarak Shah. "MIRA: Towards Mitigating Reward Hacking in Inference-Time Alignment of T2I Diffusion Models." arXiv, 2025.

P6 Wang, Xiyao, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. "LLaVA-Critic-R1: Your Critic Model Is Secretly a Strong Policy Model." arXiv, 2025.

P5 Cai, Zikui, Andrew Wang, Anirudh Satheesh, Ankit Nakhawa, Hyunwoo Jae, Keenan Powell, Minghui Liu, Neel Jay, Sungbin Oh, Xiyao Wang, Yongyuan Liang, Tom Goldstein, and Furong Huang. "MORSE-500: A Programmatically Controllable Video Benchmark to Stress-Test Multimodal Reasoning." arXiv, 2025.

P4 Chakraborty, Souradip, Mohammadreza Pourreza, Ruoxi Sun, Yiwen Song, Nino Scherrer, Furong Huang, Amrit Singh Bedi, Ahmad Beirami, Jindong Gu, Hamid Palangi, and Tomas Pfister. "On the Role of Feedback in Test-Time Scaling of Agentic AI Workflows." arXiv, 2025.

P3 Liu, Minghui, Tahseen Rabbani, Tony O'Halloran, Ananth Sankaralingam, Mary-Anne Hartley, Furong Huang, Cornelia Fermüller, and Yiannis Aloimonos. "HashEvict: A Pre-Attention KV Cache Eviction Strategy using Locality-Sensitive Hashing." arXiv, 2024.

P2 Li, Lincan, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, Alex Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijian Du, Fangzhou Mu, Jiaxin Pei, Jieyu Zhao, Swabha Swayamdipta, Willie Neiswanger, Hua Wei, Xiyang Hu, Shixiang Zhu, Tianlong Chen, Yingzhou Lu, Yang Shi, Lianhui Qin, Tianfan Fu, Zhengzhong Tu, Yuzhe Yang, Jaemin Yoo, Jiaheng Zhang, Ryan Rossi, Liang Zhan, Liang Zhao, Emilio Ferrara, Yan Liu, Furong Huang, Xiangliang Zhang, Lawrence Rothenberg, Shuiwang Ji, Philip S. Yu, Yue Zhao, and Yushun Dong. "Political-LLM: Large Language Models in Political Science." arXiv, 2024.

P1 Yang, Yifan, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W. John Wilbur, Xiaojun Li, Jiaxin Yuan, Bang An, Kelvin S. Castro, Francisco Erramuspe Álvarez, Matías Stockle, Aidong Zhang, Furong Huang, and Zhiyong Lu. "Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine." arXiv, 2024.

Conference Publications

130 items

C130 Xu, Jiawei, Minghui Liu, Aakriti Agrawal, Yifan Chen, and Furong Huang. “Scheduling Thoughts: Learning the Order of Thought in Diffusion Language Models.” openreview. Forty-third International Conference on Machine Learning (ICML), 2026.

C129 Zheng, Tong, Chengsong Huang, Runpeng Dai, Yun He, Rui Liu, Xin Ni, Huiwen Bao, Kaishen Wang, Hongtu Zhu, Jiaxin Huang, Furong Huang, and Heng Huang. “Parallel-Probe: Towards Efficient Parallel Thinking via 2D Probing.” arXiv. Forty-third International Conference on Machine Learning (ICML), 2026.

C128 Yu, Fangxu, Xingang Guo, Lingzhi Yuan, Haoqiang Kang, Hongyu Zhao, Lianhui Qin, Furong Huang, Bin Hu, and Tianyi Zhou. “TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models.” arXiv. Project page. Forty-third International Conference on Machine Learning (ICML), 2026.

C127 Lazri, Zachary McBride, Anirudh Nakra, Ivan Brugere, Danial Dervovic, Antigoni Polychroniadou, Furong Huang, Dana Dachman-Soled, and Min Wu. “MAFE: Enabling Equitable Algorithm Design in Multi-Agent Multi-Stage Decision-Making Systems.” arXiv. Forty-third International Conference on Machine Learning (ICML), 2026.

C126 Ghosal, Soumya Suvra, Souradip Chakraborty, Vaibhav Singh, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. “Safety Recovery in Reasoning Models Is Only a Few Early Steering Steps Away.” arXiv. Forty-third International Conference on Machine Learning (ICML), 2026.

C125 Xiong, Nuoya, Yuhang Zhou, Hanqing Zeng, Zhaorun Chen, Furong Huang, Shuchao Bi, Lizhu Zhang, and Zhuokai Zhao. “Token-Level LLM Collaboration via FusionRoute.” arXiv. Forty-third International Conference on Machine Learning (ICML), 2026.

C124 Sehwag, Udari Madhushani, Elaine Lau, Haniyeh Ehsani Oskouie, Shayan Shabihi, Erich Liang, Andrea Sarai Echeverria Toledo, Guillermo A. Mangialardi, Sergio Fonrouge, Ed-Yeremai Hernandez-Cardona, Paula Vergara, Utkarsh Tyagi, Chen Bo Calvin Zhang, Pavi Bhatter, Nicholas E. Johnson, Furong Huang, Ernesto Gabriel Hernández Montoya, and Bing Liu. “SciPredict: Can LLMs Predict the Outcomes of Scientific Experiments in Natural Sciences?” arXiv. Code. Forty-third International Conference on Machine Learning (ICML), 2026.

C123 Bornstein, Marco, Zora Che, Suhas Julapalli, Abdirisak Mohamed, Amrit Singh Bedi, and Furong Huang. “Advancing Regulation in Artificial Intelligence: An Auction-Based Approach.” arXiv. The Ninth Annual ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2026.

C122 Pathmanathan, Pankayaraj, and Furong Huang. “Teach a Reward Model to Correct Itself: Reward Guided Adversarial Failure Discovery for Robust Reward Modeling.” arXiv. Code. Main Conference, The 64th Annual Meeting of the Association for Computational Linguistics (ACL), **Oral**, 2026.

C121 Agrawal, Aakriti, Gouthaman KV, Rohith Aralikatti, Gauri Jagatap, Jiaxin Yuan, Sarvesh Baskar, Vijay Kamarshi, Andrea Fanelli, and Furong Huang. “Towards Mitigating Hallucinations in Large Vision-Language Models by Refining Textual Embeddings.” arXiv. Findings, The 64th Annual Meeting of the Association for Computational Linguistics (ACL), 2026.

C120 Agrawal, Aakriti, Mucong Ding, Chenghao Deng, Zora Che, Anirudh Satheesh, Arjun Rajaram, Bang An, C. Bayan Bruss, John Langford, and Furong Huang. “EnsemW2S: Enhancing Weak-to-Strong Generalization with Large Language Model Ensembles.” arXiv. Findings, The 64th Annual Meeting of the Association for Computational Linguistics (ACL), 2026.

C119 Lee, Seungjae, Yoonkyo Jung, Inkook Chun, Yao-Chih Lee, Zikui Cai, Hongjia Huang, Aayush Talreja, Tan Dat Dao, Yongyuan Liang, Jia-Bin Huang and Furong Huang. “TraceGen: World Modeling in 3D Trace Space Enables Learning from Cross-Embodiment Videos.” arXiv. Project page. Code. Conference on Computer Vision and Pattern Recognition (CVPR), 2026.

C118 Ju, Yuanchen, Yongyuan Liang, Yen-Jen Wang, Nandiraju Gireesh, Yuanliang Ju, Seungjae Lee, Qiao Gu, Elvis Hsieh, Furong Huang*, and Koushil Sreenath*. “MomaGraph: State-Aware Unified Scene Graphs with Vision-Language Model for Embodied Task Planning.” arXiv. Project page. Code. The Fourteenth International Conference on Learning Representations (ICLR), **Oral**, 2026.

C117 Liang, Yongyuan, Wei Chow, Feng Li, Ziqiao Ma, Xiyao Wang, Jiageng Mao, Jiuhai Chen, Jiatao Gu, Yue Wang, and Furong Huang. “ROVER: Benchmarking Reciprocal Cross-Modal Reasoning for Omnimodal Generation.” arXiv. Project page. Code. The Fourteenth International Conference on Learning Representations (ICLR), 2026.

C116 Sehwag, Udari Madhushani, Shayan Shabihi, Alex McAvoy, Vikash Sehwag, Yuancheng Xu, Dalton Towers, and Furong Huang. “PropensityBench: Evaluating Latent Safety Risks in Large Language Models via an Agentic Approach.” arXiv. Code. The Fourteenth International Conference on Learning Representations (ICLR), 2026.

C115 Li, Ang, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein and Micah Goldblum. “Zebra-CoT: A Dataset for Interleaved Vision-Language Reasoning.” arXiv. Dataset. The Fourteenth International Conference on Learning Representations (ICLR), 2026.

C114 Huang, Yue, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Jiayi Ye, Yujun Zhou, Yanbo Wang, Jiawen Shi, Qihui Zhang, Han Bao, Zhaoyi Liu, Yuan Li, Tianrui Guan, Peiran Wang, Haomin Zhuang, Dongping Chen, Kehan Guo„y Zou, Bryan Hooi, Caiming Xiong, Elias Stengel-Eskin, Hongyang Zhang, Hongzhi Yin, Huan Zhang, Huaxiu Yao, Jieyu Zhang, Jaehong Yoon, Kai Shu, Ranjay Krishna, Swabha Swayamdipta, Weijia Shi, Xiang Li, Yuexing

Hao, Zhihao Jia, Zhize Li, Xiuying Chen, Zhengzhong Tu, Xiyang Hu, Tianyi Zhou, Jieyu Zhao, Lichao Sun, Furong Huang, Or Cohen-Sasson, Prasanna Sattigeri, Anka Reuel, Max Lamparth, Yue Zhao, Nouha Dziri, Yu Su, Huan Sun, Heng Ji, Chaowei Xiao, Mohit Bansal, Nitesh V Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S. Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, Dawn Song, Xiangliang Zhang. "TrustGen: A Platform of Dynamic Benchmarking on the Trustworthiness of Generative Foundation Models." arXiv. The Fourteenth International Conference on Learning Representations (ICLR), 2026.

C113 Pathmanathan, Pankayaraj, Udari Madhushani Sehwag, Michael-Andrei Panaitescu-Liess, Cho-Yu Jason Chiang and Furong Huang. "AdvBDGen: A Robust Framework for Generating Adaptive and Stealthy Backdoors in LLM Alignment Attacks." arXiv. Code. Project page. AAAI 2026 AI Alignment Track (AAAI), **Oral**, 2026.

C112 Beetham, James, Souradip Chakraborty, Mengdi Wang, Furong Huang, Amrit Singh Bedi and Mubarak Shah. "Jailbreaks as Inference-Time Alignment: A Framework for Understanding Safety Failures in LLMs." paper. 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2026.

C111 Wang, Xiyao, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang*, and Lijuan Wang*. "SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement." arXiv. Code. The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS), **Spotlight**, 2025.

C110 Wang, Xiyao, Zhengyuan Yang, Chao Feng, Yuhang Zhou, Xiaoyu Liu, Yongyuan Liang, Ming Li, Ziyi Zang, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang* and Lijuan Wang*. "ViCrit: A Verifiable Reinforcement Learning Proxy Task for Visual Perception in VLMs." arXiv. Code. The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.

C109 Ghosal, Soumya Suvra, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh and Amrit Singh Bedi. "Does Thinking More Always Help? Mirage of Test-Time Scaling in Reasoning Models." arXiv. The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.

C108 Ding, Mucong, Bang An, Tahseen Rabbani, Chenghao Deng, Anirudh Satheesh, Souradip Chakraborty, Mehrdad Saberi, Yuxin Wen, Kyle Rui Sang, Aakriti Agrawal, Xuandong Zhao, Mo Zhou, Mary-Anne Hartley, Lei Li, Yu-Xiang Wang, Vishal M. Patel, Soheil Feizi, Tom Goldstein and Furong Huang. "A Technical Report on "Erasing the Invisible": The 2024 NeurIPS Competition on Stress Testing Image Watermarks." openreview. Competition website (beige box). Competition website (black box). Project page. The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS), 2025.

C107 Agrawal, Aakriti, Rohith Aralikatti, Anirudh Satheesh, Souradip Chakraborty, Amrit Singh Bedi, Furong Huang. "Uncertainty-Aware Answer Selection for Improved Reasoning in Multi-LLM Systems." arXiv. The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025.

C106 Zhou, Yuhang, Jing Zhu, Shengyi Qian, Zhuokai Zhao, Xiyao Wang, Xiaoyu Liu, Ming Li, Paiheng Xu, Wei Ai, Furong Huang. "DISCO Balances the Scales: Adaptive Domain- and Difficulty-Aware Reinforcement Learning on Imbalanced Data." arXiv. Code. The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025.

C105 Lee, Seungjae[†], Daniel Ekpo[†], Haowen Liu, Furong Huang[*], Abhinav Shrivastava[*], Jia-Bin Huang[*]. "Imagine, Verify, Execute: Agentic Exploration with Vision-Language Models." arXiv. Code. Project page. 9th Annual Conference on Robot Learning (CoRL), 2025.

C104 Zheng, Ruijie, Jing Wang, Scott Reed, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, Linxi Fan. "FLARE: Robot Learning with Implicit World Modeling." arXiv. 9th Annual Conference on Robot Learning (CoRL), 2025.

C103 Yu, Kelin, Sheng Zhang, Harshit Soora, Furong Huang, Heng Huang, Pratap Tokekar, and Ruohan Gao. "GenFlowRL: Shaping Rewards with Generative Object-Centric Flow in Visual Reinforcement Learning." arXiv. Project page. International Conference on Computer Vision (ICCV), 2025.

C102 Wang, Xiyao, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang[*], and Lijuan Wang[*]. "VisVM: Scaling Inference-time Search with Vision Value Model for Improved Visual Comprehension." arXiv. Code. Project page. International Conference on Computer Vision (ICCV), 2025.

C101 Yue, Kaiyu, Vasu Singla, Menglin Jia, John Kirchenbauer, Rifaa Qadri, Zikui Cai, Abhinav Bhatele, Furong Huang, and Tom Goldstein. "Zero-Shot Vision Encoder Grafting via LLM Surrogates." arXiv. International Conference on Computer Vision (ICCV), 2025.

C100 Joshua McClellan, Greyson Brothers, Furong Huang, and Pratap Tokekar. "PEnGUiN: Partially Equivariant Graph NeUral Networks for Sample Efficient MARL." arXiv. Reinforcement Learning Conference (RLC), 2025.

C99 Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, and Amrit Singh Bedi. "Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment." arXiv. Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

C98 Zheng, Ruijie, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daume III, Andrey Kolobov, Furong Huang, and

Jianwei Yang. "TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies." arXiv. Code. Project page. The Thirteenth International Conference on Learning Representations (ICLR), 2025.

C97 Xu, Yuancheng, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. "GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-Time Alignment." arXiv. Project page. Code. The Thirteenth International Conference on Learning Representations (ICLR), 2025.

C96 Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. "Benchmarking Vision Language Model Unlearning via Fictitious Facial Identity Dataset." arXiv. Project page. The Thirteenth International Conference on Learning Representations (ICLR), 2025.

C95 Chakraborty, Souradip, Sujay Bhatt, Udari Madhushani Sehwag, Soumya Suvra Ghosal, Jiahao Qiu, Mengdi Wang, Dinesh Manocha, Furong Huang, Alec Koppel, and Sumitra Ganesh. "Collab: Controlled Decoding using Mixture of Agents for LLM Alignment." arXiv. Project page. The Thirteenth International Conference on Learning Representations (ICLR), 2025.

C94 Liu, Xiaoyu, Jiaxin Yuan, Yuhang Zhou, Jingling Li, Furong Huang, and Wei Ai. "CSRec: Rethinking Sequential Recommendation from A Causal Perspective." arXiv. The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2025.

C93 Liu, Zeyuan, Maggie Z. Huan, Xiyao Wang, Jiafei Lyu, Jian Tao, Xiu Li, Furong Huang, and Huazhe Xu. "World Models with Hints of Large Language Models for Goal Achieving." arXiv. Project page. The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL), 2025.

C92 Wang, Xiyao, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang*, Cao Xiao*. "SIMA: Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement." arXiv. Code. The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL), 2025.

C91 Zhou, Yuhang, Giannis Karamanolakis, Victor Soto, Anna Rumshisky, Mayank Kulkarni, Furong Huang, Wei Ai, Jianhua Lu. "MergeME: Model Merging Techniques for Homogeneous and Heterogeneous MoEs." arXiv. Project page. The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL), 2025.

C90 Panaitescu-Liess, Michael-Andrei, Pankayaraj Pathmanathan, Yigitcan Kaya, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, Furong Huang. "PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models." arXiv. The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL), 2025.

C89 Liu, Xiaoyu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, Furong Huang. "Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey." arXiv. The 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL), 2025.

C88 Zhang, Zhi, Chris Chow, Yasi Zhang, Yanchao Sun, Haochen Zhang, Eric Hanchen Jiang, Han Liu, Furong Huang, Yuchen Cui, and Oscar Hernan Madrid Padilla. "Statistical Guarantees for Lifelong Reinforcement Learning using PAC-Bayesian Theory." arXiv. Project page. The 28th International Conference on Artificial Intelligence and Statistics (AISTATS), 2025.

C87 Xu, Paiheng, Yuhang Zhou, Bang An, Wei Ai, and Furong Huang. "GFairHint: Improving Individual Fairness for Graph Neural Networks via Fairness Hint." arXiv. Code. ACM Transactions on Knowledge Discovery from Data (TKDD), 2025.

C86 Zhang, Zhili, H M Sabbir Ahmad, Ehsan Sabouni, Yanchao Sun, Furong Huang, Wenchao Li, and Fei Miao. "Safety Guaranteed Robust Multi-Agent Reinforcement Learning with Hierarchical Control for Connected and Automated Vehicles." arXiv. Project page. IEEE International Conference on Robotics and Automation (ICRA), 2025.

C85 Pathmanathan, Pankayaraj, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, and Furong Huang. "Is poisoning a real threat to DPO? Maybe more so than you think." arXiv. Code. AAAI 2025 AI Alignment Track (AAAI), 2025.

C84 Panaitescu-Liess, Michael-Andrei, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. "Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?" arXiv. Code. **Best Paper Award** at the Third Workshop on New Frontiers in Adversarial Machine Learning, NeurIPS 2024. The 39th Annual AAAI Conference on Artificial Intelligence (AAAI), 2025.

C83 Ding, Mucong, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. "Easy2Hard-Bench: Standardized Difficulty Labels for Profiling LLM Performance and Generalization." arXiv. Project page. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2024.

C82 Bornstein, Marco, Amrit Bedi, Abdirisak Mohamed, and Furong Huang. "FACT or Fiction: Can Truthful Mechanisms Eliminate Federated Free Riding?" arXiv. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

C81 Liang, Yongyuan, Tingqiang Xu, Kaizhe Hu, Guangqi Jiang, Furong Huang, and Huazhe Xu. "Make-An-Agent: A

Generalizable Policy Network Generator with Behavior-Prompted Diffusion." arXiv. Project page. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

C80 Xu, Yuancheng, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. "Shadowcast: Stealthy Data Poisoning Attacks Against Vision-Language Models." arXiv. Project page. Code. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

C79 Chakraborty, Souradip, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Bedi, and Furong Huang. "Transfer Q-star: Principled Decoding for LLM Alignment." arXiv. Code. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

C78 McClellan, Joshua, Naveed Haghani, John Winder, Furong Huang, and Pratap Tokekar. "Boosting Sample Efficiency and Generalization in Multi-agent Reinforcement Learning via Equivariance." arXiv. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.

C77 Zhou, Yuhang, Jing Zhu, Paiheng Xu, Xiaoyu Liu, Xiyao Wang, Danai Koutra, Wei Ai, and Furong Huang. "Multi-Stage Balanced Distillation: Addressing Long-Tail Challenges in Sequence-Level Knowledge Distillation." arXiv. The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

C76 Wu, Xiyang, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. "AUTOHALLUSION: Automatic Generation of Hallucination Benchmarks for Vision-Language Models." arXiv. Code. The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

C75 Zhu, Sicheng, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. "AutoDAN: Interpretable Gradient-Based Adversarial Attacks on Large Language Models." arXiv. First Conference on Language Modeling (COLM), 2024.

C74 Zhu, Sicheng, Bang An, Ruiyi Zhang, Michael-Andrei Panaitescu-Liess, Yuancheng Xu, and Furong Huang. "Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models." arXiv. Code. First Conference on Language Modeling (COLM), 2024.

C73 Zhou, Yuhang, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. "Explore Spurious Correlations at the Concept Level in Language Models for Text Classification." arXiv. The 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

C72 Wang, Xiyao, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. "Mementos: A Comprehensive Benchmark for Multimodal Large Language Model Reasoning over Image Sequences." arXiv. Code. The 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

C71 Sang, Kyle, Tahseen Rabbani, and Furong Huang. "Balancing Label Imbalance in Federated Environments Using Only Mixup and Artificially-Labeled Noise." arXiv. 4th International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI), 2024.

C70 Chakraborty, Souradip, Amrit Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. "Position: On the Possibilities of AI-Generated Text Detection." arXiv. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C69 Yuan, Dehao, Furong Huang, Tahseen Rabbani, Cornelia Fermuller, and Yiannis Aloimonos. "Inherently Efficient and Noise-Robust Local Point Cloud Geometry Encoder via Vectorized Kernel Mixture." arXiv. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C68 Huang,Yue, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. "Position: TrustLLM: Trustworthiness in Large Language Models." arXiv. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C67 An, Bang, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. "WAVES: Benchmarking the Robustness of Image Watermarks." arXiv. Project page. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C66 Xu, Yuancheng, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. "Adapting Static Fairness to Sequential Decision-Making: Bias Mitigation Strategies towards Equal Long-term Benefit Rate." arXiv. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C65 Zheng, Ruijie, Ching-An Cheng, Hal Daume III, Furong Huang, and Andrey Kolobov. "PRISE: LLM-Style Sequence Compression for Learning Temporal Action Abstractions in Control." arXiv. Code. Proceedings of the 41st International

Conference on Machine Learning (ICML), 2024.

C64 Zheng, Ruijie, Yongyuan Liang, Xiyao Wang, Shuang Ma, Hal Daume III, Huazhe Xu, John Langford, Praveen Palanisamy, Kalyan Shankar Basu, and Furong Huang. "Premier-TACO is a Few-Shot Policy Learner: Pretraining Multitask Representation via Temporal Action-Driven Contrastive Loss." arXiv. Project page. Code. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C63 Ji, Tianying, Yongyuan Liang, Yan Zeng, Yu Luo, Guowei Xu, Jiawei Guo, Ruijie Zheng, Furong Huang, Fuchun Sun, and Huazhe Xu. "ACE: Off-Policy Actor-Critic with Causality-Aware Entropy Regularization." arXiv. Project page. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C62 Chakraborty, Souradip, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. "MaxMin-RLHF: Alignment with Diverse Human Preferences." arXiv. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

C61 Guan, Tianrui, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. "HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination & Visual Illusion in Large Vision-Language Models." arXiv. Code. Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

C60 Liu, Xiangyu, Chenghao Deng, Yanchao Sun, Yongyuan Liang, and Furong Huang. "Beyond Worst-case Attacks: Robust RL with Adaptive Defense via Non-dominated Policies." **Spotlight**. arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C59 Liu, Xiangyu, Souradip Chakraborty, Yanchao Sun, and Furong Huang. "Rethinking Adversarial Policies: A Generalized Attack Formulation and Provable Defense in RL." arXiv. A workshop version of it, "Controllable Attack and Improved Adversarial Training in Multi-Agent Reinforcement Learning," won the **Outstanding Paper Award** at the Workshop on Trustworthy and Socially Responsible Machine Learning NeurIPS 2022. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C58 Chakraborty, Souradip, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. "PARL: A Unified Framework for Policy Alignment in Reinforcement Learning." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C57 Ding, Mucong, Bang An, Yuancheng Xu, Anirudh Satheesh, and Furong Huang. "SAFLEX: Self-Adaptive Augmentation via Feature Label Extrapolation." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C56 Wang, Xiyao, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. "COPlanner: Plan to Roll Out Conservatively but to Explore Optimistically for Model-Based RL." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C55 Xu, Guowei, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, Shuzhen Li, Yanjie Ze, Hal Daume III, Furong Huang, and Huazhe Xu. "DrM: Mastering Visual Reinforcement Learning through Dormant Ratio Minimization." **Spotlight**. arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C54 An, Bang, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. "More Context, Less Distraction: Zero-shot Visual Classification by Inferring and Conditioning on Contextual Attributes." arXiv. Code. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C53 Liang, Yongyuan, Yanchao Sun, Ruijie Zheng, Xiangyu Liu, Benjamin Eysenbach, Tuomas Sandholm, Furong Huang, and Stephen Marcus McAleer. "Game-Theoretic Robust Reinforcement Learning Handles Temporally-Coupled Perturbations." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C52 Panaitescu-Liess, Michael-Andrei, Yigitcan Kaya, Sicheng Zhu, Furong Huang, and Tudor Dumitras. "Like Oil and Water: Group Robustness Methods and Poisoning Defenses Don't Mix." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C51 Yuan, Dehao, Furong Huang, Cornelia Fermuller, and Yiannis Aloimonos. "Decodable and Sample Invariant Continuous Object Encoder." arXiv. The Twelfth International Conference on Learning Representations (ICLR), 2024.

C50 Liu, Xiaoyu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. "C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confounder." arXiv. The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023.

C49 Rabbani, Tahseen, Marco Bornstein, and Furong Huang. "Large-Scale Distributed Learning via Private On-Device LSH." Paper. arXiv. The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023.

C48 Zheng, Ruijie, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daume III, and Furong Huang. "TACO: Temporal Latent Action-Driven Contrastive Loss for Visual Reinforcement Learning." arXiv. Code. The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023.

C47 Bansal, Arpit, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise." arXiv. Code. The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023.

C46 Wang, Xiyao, Wichayaporn Wongkamjan, and Furong Huang. "Live in the Moment: Learning Dynamics Model Adapted to Evolving Policy." arXiv. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

C45 Zhu, Sicheng, Bang An, Furong Huang, and Sanghyun Hong. "Learning Unforeseen Robustness from Out-of-distribution Data Using Equivariant Domain Translator." Openreview. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

C44 Chakraborty, Souradip, Amrit Singh Bedi, Alec Koppel, Mengdi Wang, Furong Huang, and Dinesh Manocha. "STEERING : Stein Information Directed Exploration for Model-Based Reinforcement Learning." arXiv. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

C43 Sun, Yanchao, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. "Certifiably Robust Multi-Agent Reinforcement Learning against Adversarial Communication." arXiv. The Eleventh International Conference on Learning Representations (ICLR), 2023.

C42 Zheng, Ruijie, Xiyao Wang, Huazhe Xu, and Furong Huang. "Is Model Ensemble Necessary? Model-based RL via a Single Model with Lipschitz Regularized Value Function." arXiv. The Eleventh International Conference on Learning Representations (ICLR), 2023.

C41 Bornstein, Marco, Tahseen Rabbani, Evan Wang, Amrit Singh Bedi, and Furong Huang. "SWIFT: Rapid Decentralized Federated Learning via Wait-Free Model Communication." arXiv. The Eleventh International Conference on Learning Representations (ICLR), 2023.

C40 Xu, Yuancheng, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang. "Exploring and Exploiting Decision Boundary Dynamics for Adversarial Robustness." arXiv. Code. The Eleventh International Conference on Learning Representations (ICLR), 2023.

C39 Sun, Yanchao, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, and Ashish Kapoor. "SMART: Self-supervised Multi-task pretrAining with contRol Transformers." arXiv. Code. The Eleventh International Conference on Learning Representations (ICLR), 2023.

C38 Chakraborty, Souradip, Amrit Bedi, Pratap Tokekar, Alec Koppel, Brian Sadler, Furong Huang, and Dinesh Manocha. "Posterior Coreset Construction with Kernelized Stein Discrepancy for Model-Based Reinforcement Learning." arXiv. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI), 2023.

C37 Liang, Yongyuan, Yanchao Sun, Ruijie Zheng, and Furong Huang. "Efficient Adversarial Training without Attacking: Worst-Case-Aware Robust Reinforcement Learning." arXiv. Code. Neural Information Processing System (NeurIPS), 2022.

C36 An, Bang, Zora Che, Mucong Ding, and Furong Huang. "Transferring Fairness under Distribution Shifts via Fair Consistency Regularization." arXiv. Code. Neural Information Processing System (NeurIPS), 2022.

C35 Ding, Mucong, Tahseen Rabbani, Bang An, Evan Wang, and Furong Huang. "Sketch-GNN: Efficient GNNs with Graph Size-Independent Scalability." arXiv. Neural Information Processing System (NeurIPS), 2022.

C34 Yang, Kaiwen, Yanchao Sun, Jiahao Su, Fengxiang He, Xinmei Tian, Furong Huang, Tianyi Zhou, and Dacheng Tao. "Adversarial Auto-Augment with Label Preservation: A Representation Learning Principle Guided Approach." arXiv. Code. Neural Information Processing System (NeurIPS), 2022.

C33 Bansal, Arpit, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. "End-to-end Algorithm Synthesis with Recurrent Networks: Logical Extrapolation Without Overthinking." arXiv. Neural Information Processing System (NeurIPS), 2022.

C32 Levin, Roman, Manli Shu, Eitan Borgnia, Furong Huang, Micah Goldblum, and Tom Goldstein. "Where do Models go Wrong? Parameter-Space Saliency Maps for Explainability." arXiv. Neural Information Processing System (NeurIPS), 2022.

C31 Su, Jiahao, Wonmin Byeon, and Furong Huang. "Scaling-up Diverse Orthogonal Convolutional Networks by a Paraunitary Framework." arXiv. Proceedings of the 39th International Conference on Machine Learning (ICML), 2022.

C30 Zhang, Zhi, Zhuoran Yang, Han Liu, Pratap Tokekar, and Furong Huang. "Reinforcement Learning under a Multi-agent Predictive State Representation Model: Method and Theory." Openreview. Tenth International Conference on Learning Representations (ICLR), 2022.

C29 Sun, Yanchao, Ruijie Zheng, Xiyao Wang, Andrew Cohen, and Furong Huang. "Transfer RL across Observation Feature Spaces via Model-Based Regularization." arXiv. Tenth International Conference on Learning Representations (ICLR), 2022.

C28 Sun, Yanchao, Ruijie Zheng, Yongyuan Liang, and Furong Huang. "Who is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL." arXiv. Code. **Best Paper Award** at the Workshop on Safe and Robust Control of Uncertain Systems (SafeRL) NeurIPS 2021. Tenth International Conference on Learning Representations (ICLR), 2022.

C27 Liu, Xiaoyu, Jiahao Su, and Furong Huang. "Tuformer: Data-driven Design of Transformers for Improved Generalization or Efficiency." Openreview. Code. Tenth International Conference on Learning Representations (ICLR), 2022.

C26 Zhu, Sicheng, Bang An, and Furong Huang. "Understanding the Generalization Benefit of Model Invariance from a

Data Perspective." arXiv. Neural Information Processing System (NeurIPS), 2021.

C25 Ding, Mucong, Kezhi Kong, Jingling Li, Chen Zhu, John Dickerson, Furong Huang, and Tom Goldstein. "VQ-GNN: A Universal Framework to Scale up Graph Neural Networks using Vector Quantization." arXiv. Neural Information Processing System (NeurIPS), 2021.

C24 Schwarzschild, Avi, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. "Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks." arXiv. Neural Information Processing System (NeurIPS), 2021.

C23 Sun, Yanchao, Da Huo, and Furong Huang. "Vulnerability-Aware Poisoning Mechanism for Online RL with Unknown Dynamics." arXiv. Ninth International Conference on Learning Representations (ICLR), 2021.

C22 Sun, Yanchao, Xiangyu Yin, and Furong Huang. "TempLe: Learning Template of Transitions for Sample Efficient Multi-Task RL." arXiv. 35th AAAI conference on Artificial Intelligence (AAAI), 2021.

C21 Zhu, Chen, Yu Cheng, Zhe Gan, Furong Huang, Jingjing Liu, and Tom Goldstein. "MaxVA: Fast Adaptation of Stepsizes by Maximizing Observed Variance of Gradients." arXiv. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2021.

C20 Zeng, Huimin, Chen Zhu, Tom Goldstein, and Furong Huang. "Are Adversarial Examples Created Equal? A Learnable Weighted Minimax Risk for Robustness under Non-Uniform Attacks." arXiv. 35th AAAI conference on Artificial Intelligence (AAAI), 2021.

C19 Rabbani, Tahseen, Apollo Jain, Arjun Rajkumar, and Furong Huang. "Practical and Fast Momentum-Based Power Methods." Paper. arXiv. 2nd Annual Conference on Mathematical and Scientific Machine Learning (MSML), Proceedings of Machine Learning Research vol 145:1-36, 2021.

C18 Su, Jiahao, Shiqi Wang, and Furong Huang. "ARMA Nets: Expanding Receptive Field for Dense Prediction." arXiv. Neural Information Processing System (NeurIPS), 2020.

C17 Su, Jiahao, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. "Convolutional Tensor-Train LSTM for Spatio-Temporal Learning." arXiv. Neural Information Processing System (NeurIPS), 2020.

C16 Decarolis, Chris, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. "An end-to-end Differentially Private Latent Dirichlet Allocation Using a Spectral Algorithm." arXiv. Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

C15 Reustle, Alexander, Tahseen Rabbani, and Furong Huang. "Fast GPU Convolution for CP-decomposed Tensorial Neural Networks." Paper. Intelligent Systems Conference (IntelliSys), 2020.

C14 Sun, Yanchao, and Furong Huang. "Can Agents Learn by Analogy? An Inferable Model for PAC Reinforcement Learning." arXiv. International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2020.

C13 Li, Jingling, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. "Understanding Generalization in Deep Learning via Tensor Methods." arXiv. The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

C12 Su, Jiahao, Milan Cvitkovic, and Furong Huang. "Sampling-Free Learning of Bayesian Quantized Neural Networks." arXiv. The International Conference on Learning Representations (ICLR), 2020.

C11 Shafahi, Ali, Amin Ghiasi, Mahyar Najibi, Furong Huang, John P. Dickerson, and Tom Goldstein. "Label Smoothing and Logit Squeezing: A Replacement for Adversarial Training?" arXiv. 30th British Machine Vision Conference (BMVC), 2019.

C10 Curry, Michael J., Duncan McElfresh, Xuchen You, Cameron Moy, Furong Huang, Tom Goldstein, and John P. Dickerson, "Reinforcement Learning for Dynamic Set Packing." Conference on Reinforcement Learning and Decision Making (RLDM), 2019.

C9 Huang, Furong, Niranjan Uma Naresh, Ioakeim Perros, Robert Chen, Jimeng Sun, and Anima Anandkumar. "Guaranteed Scalable Learning of Latent Tree Models." arXiv. The Conference on Uncertainty in Artificial Intelligence (UAI), 2019.

C8 Huang, Furong, Jordan Ash, John Langford, and Robert Schapire. "Learning Deep ResNet Blocks Sequentially using Boosting Theory." arXiv. The 35th International Conference on Machine Learning (ICML), 2018.

C7 Huang, Furong, and Animashree Anandkumar. "Convolutional Dictionary Learning through Tensor Factorization." arXiv. Journal of Machine Learning Research conference and workshop proceedings 2015.

C6 Arabshahi, Forough, Furong Huang, Animashree Anandkumar, Carter T. Butts, and Sean M. Fitzhugh. "Are you going to the party: depends, who else is coming? –Learning hidden group dynamics via conditional latent tree models." arXiv. IEEE International Conference on Data Mining 2015.

C5 Ge, Rong, Furong Huang, Chi Jin, and Yang Yuan. (Alphabetic Order) "Escaping From Saddle Points – Online Stochastic Gradient for Tensor Decomposition." arXiv. Conference of Learning Theory (COLT) 2015.

C4 Huang, Furong, and Anima Anandkumar. "FCD: Fast-Concurrent-Distributed Load Balancing under Switching Costs and Imperfect Observations." Paper. In Proc. of the 32nd Annual IEEE International Conference on Computer Communications (INFOCOM), Turin, Italy, Apr. 2013.

C3 Anandkumar, Animashree, Daniel Hsu, Furong Huang, and Sham M. Kakade. "Learning High-Dimensional Mixtures of Graphical Models." arXiv. Conference on Neural Information Processing Systems (NIPS), 2012.

C2 Anandkumar, Animashree, Vincent YF Tan, Furong Huang, and Alan S. Willsky. “High-Dimensional Gaussian Graphical Model Selection: Walk-Summability and Local Separation Criterion.” Paper. arXiv. Conference on Neural Information Processing Systems (NIPS), 2011.

C1 Anandkumar, Animashree, Vincent YF Tan, Furong Huang, and Alan S. Willsky. “High-Dimensional Structure Learning of Ising Models: Local Separation Criterion.” arXiv. Conference on Neural Information Processing Systems (NIPS), 2011.

Journal Publications

10 items

J10 Yang, Yifan, Qiao Jin, Furong Huang, and Zhiyong Lu. “Adversarial Prompt and Fine-Tuning Attacks Threaten Medical Large Language Models.” Nature Communications, 16:9011, 2025. DOI. arXiv.

J9 Che, Zora, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, Zikui Cai, Bilal Chughtai, Yarin Gal, Furong Huang, and Dylan Hadfield-Menell. “Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities.” arXiv. Transactions on Machine Learning Research (TMLR), 2025.

J8 Yang, Yifan, Mingquan Lin, Han Zhao, Yifan Peng, Furong Huang, and Zhiyong Lu. ”A survey of recent methods for addressing AI fairness and bias in biomedicine.” Journal of Biomedical Informatics (2024): 104646.

J7 Ghosal, Soumya Suvra, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit S. Bedi. “A Survey on the Possibilities & Impossibilities of AI-generated Text Detection.” arXiv. Transactions on Machine Learning Research (TMLR), 2024.

J6 Su, Jiahao, Jingling Li, Xiaoyu Liu, Teresa Ranadive, Christopher Coley, Tai-Ching Tuan, and Furong Huang. “Compact Neural Architecture Designs by Tensor Representations.” Frontiers in Artificial Intelligence-Machine Learning and Artificial Intelligence, 2022.

J5 Giraud, Luc, Ulrich Rüde, Linda Stals, Paolo Bientinesi, David Ham, Furong Huang, Paul HJ Kelly, Christian Lengauer, and Saday Sadayappan. “Dagstuhl Reports, Vol. 10, Issue 3 ISSN 2192-5283.” (2020).

J4 Huang, Furong, and Animashree Anandkumar. “Convolutional Dictionary Learning through Tensor Factorization.” Journal of Machine Learning Research conference and workshop proceedings 2015.

J3 Huang, Furong, U. N. Niranjan, Mohammad Umar Hakeem, and Animashree Anandkumar. “Fast Detection of Overlapping Communities via Online Tensor Methods/ Online Tensor Methods for Learning Latent Variable Models.” Journal of Machine Learning Research 2014.

J2 Anandkumar, Animashree, Vincent YF Tan, Furong Huang, and Alan S. Willsky. “High-Dimensional Gaussian Graphical Model Selection: Walk-Summability and Local Separation Criterion.” Journal of Machine Learning Research, Aug. 2012. An abridged version appears in the Conference on Neural Information Processing Systems, Dec. 2011.

J1 Anandkumar, Animashree, Vincent YF Tan, Furong Huang, and Alan S. Willsky. “High-Dimensional Structure Learning of Ising Models: Local Separation Criterion.” Annals of Statistics, Volume 40, Number 3 (2012), 1346-1375. An abridged version appears in the Proc. of NIPS, Dec. 2011.

Refereed Workshop Publications

84 items

W84 Pathmanathan,Pankayaraj, and Furong Huang. “Deliberative Alignment is Deep, but Uncertainty Remains: Inference time safety improvement in reasoning via attribution of unsafe behavior to base model.” Actionable Interpretability Workshop (AIW), COLM 2026.

W83 Wang, Zhi, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, Yiannis Aloimonos. “Interaction-Centric Tokens: A Representation for Cross-Embodiment Manipulation.” Workshop on From Perception to Action: Representation-Centric Robot Autonomy, RSS 2026.

W82 Wang, Zhi, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. “HumanEgo: Zero-Shot Robot Learning from Minutes of Human Egocentric Videos.” Workshop on Data-Centric Robotics: What Data Do Robots Really Need, RSS 2026.

W81 Satheesh, Anirudh, Michael-Andrei Panaitescu-Liess, Andrew Ye Xu, Georgios Milis, Heng Huang, Zikui Cai, and Furong Huang. “Compositional Adversarial Training for Robust Visual Watermarking.” 2nd Workshop on Compositional Learning: Safety, Interpretability, and Agents, ICML 2026.

W80 Panaitescu-Liess, Michael-Andrei, Yigitcan Kaya, Fiza Mulla, Anirudh Satheesh, and Furong Huang. “When Detection Tools Collide: On the Impact of Watermarking on AI-Generated Text Detectors.” 4th Workshop on Towards Knowledgeable Foundation Models (KnowFM), ACL 2026.

W79 Panaitescu-Liess, Michael-Andrei, Aadi Palnitkar, Archit Kambhamettu, Yigitcan Kaya, Daniel Brown, Sungbin Oh, Sean Michael McLeish, Marco Bornstein, Furong Huang, and Tom Goldstein. “Practical Memorization Tests for Detecting Copyrighted Data in Large Language Models.” Seventh Workshop on Privacy in Natural Language Processing (PrivateNLP), ACL 2026.

W78 Ding, Mucong Sean Michael McLeish, Kazem Meidani, Igor Melnyk, Nam H Nguyen, C. Bayan Bruss, and Furong Huang. "You Only Train Once: Efficient Tokenizer Selection for Arithmetic in Language Models." Tokenization Workshop (TokShop), ICML 2025.

W77 Yu, Kelin, Sheng Zhang, Harshit Soora, Furong Huang, Heng Huang, Pratap Tokekar, and Ruohan Gao. "GENFLOWRL: Generative Object-Centric Flow Matching for Reward Shaping in Visual Reinforcement Learning." ICRA 2025 Workshop on Foundation Models and Neuro-Symbolic AI for Robotics, spotlight, ICRA 2025.

W76 Liang, Yongyuan, Xiyao Wang, Yuanchen Ju, Jianwei Yang, and Furong Huang. "LEMON: A Unified and Scalable 3D Multimodal Model for Universal Spatial Understanding." The 4th Workshop on Computer Vision in the Wild, spotlight, CVPR 2025.

W75 Zheng, Ruijie, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhe Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. "FLARE: Robot Learning with Implicit World Modeling." The 4th Workshop on Computer Vision in the Wild, spotlight, CVPR 2025.

W74 Zheng, Ruijie, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhe Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. "FLARE: Robot Learning with Implicit World Modeling." Structured World Models for Robotic Manipulation (SWOMO), oral, RSS 2025.

W73 Chiang, Jeffrey Yang Fan, Seungjae Lee, Jia-Bin Huang, Furong Huang, and Yizheng Chen. "Why Are Web AI Agents More Vulnerable Than Standalone LLMs? A Security Analysis." Workshop BuildingTrust Workshop, ICLR 2025.

W72 Cai, Zikui, Shayan Shabihi, Bang An, Zora Che, Brian R. Bartoldson, Bhavya Kailkhura, Tom Goldstein, and Furong Huang. "AegisLLM: Scaling Agentic Systems for Self-Reflective Defense in LLM Security." Workshop BuildingTrust Workshop, ICLR 2025.

W71 Pathmanathan, Pankayaraj, Udari Madhushani Sehwag, Michael-Andrei Panaitescu-Liess, and Furong Huang. "AdvB-DGen: A Robust Framework for Generating Adaptive and Stealthy Backdoors in LLM Alignment Attacks." Workshop BuildingTrust Workshop, ICLR 2025.

W70 Wang, Xiyao, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang*, and Lijuan Wang*. "VisVM: Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension." Scaling Self-Improving Foundation Models without Human Supervision (SSI-FM), ICLR 2025.

W69 Wang, Xiyao, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang*, and Lijuan Wang*. "VisVM: Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension." Workshop on World Models, ICLR 2025.

W68 Wang, Xiyao, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang*, and Lijuan Wang*. "VisVM: Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension." Reasoning and Planning for LLMs, ICLR 2025.

W67 Zheng, Ruijie, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. "TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies." Workshop on Generative Models for Robot Learning (GenBot), ICLR 2025.

W66 Pathmanathan, Pankayaraj, Udari Madhushani Sehwag, Michael-Andrei Panaitescu-Liess, and Furong Huang. "AdvBDGen: Adversarially Fortified Prompt-Specific Fuzzy Backdoor Generator Against LLM Alignment." The AAAI-25 Workshop on Artificial Intelligence for Cyber Security, AICS 2025.

W65 Panaitescu-Liess, Michael-Andrei, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. "Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?" **Best Paper Award**, Oral, The Third Workshop on New Frontiers in Adversarial Machine Learning, NeurIPS 2024.

W64 Panaitescu-Liess, Michael-Andrei, Pankayaraj Pathmanathan, Yigitcan Kaya, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, and Furong Huang. "PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models." Safe Generative AI Workshop, NeurIPS 2024.

W63 Agrawal, Aakriti, Mucong Ding, Zora Che, Chenghao Deng, Anirudh Satheesh, John Langford, and Furong Huang. "EnsemW2S: Can an Ensemble of LLMs be Leveraged to Obtain a Stronger LLM?" Safe Generative AI Workshop, NeurIPS 2024.

W62 Che, Zora, Stephen Casper, Anirudh Satheesh, Rohit Gandikota, Domenic Rosati, Stewart Slocum, Lev E McKinney, Zichu Wu, Zikui Cai, Bilal Chughtai, Furong Huang, and Dylan Hadfield-Menell. "Model Manipulation Attacks Enable More Rigorous Evaluations of LLM Unlearning." Safe Generative AI Workshop, NeurIPS 2024.

W61 Pathmanathan, Pankayaraj, Udari Madhushani Sehwag, Michael-Andrei Panaitescu-Liess, and Furong Huang. "AdvB-DGen: Adversarially Fortified Prompt-Specific Fuzzy Backdoor Generator Against LLM Alignment." Safe Generative AI Workshop, NeurIPS 2024.

W60 Liang, Yongyuan, Tingqiang Xu, Kaizhe Hu, Guangqi Jiang, Furong Huang, and Huazhe Xu. "Make-An-Agent: A Generalizable Policy Network Generator with Behavior-Prompted Diffusion." Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning, oral, NeurIPS 2024.

W59 Liu, Xiaoyu, Jiaxin Yuan, Yuhang Zhou, Jingling Li, Furong Huang, and Wei Ai. "CSRec: Rethinking Sequential Recommendation from A Causal Perspective." Causal Representation Learning Workshop, NeurIPS 2024.

W58 Bornstein, Marco, Tahseen Rabbani, Brian Joseph Gravelle, and Furong Huang. "Shrinking the Size of Extreme Multi-Label Classification." Workshop on Machine Learning and Compression, NeurIPS 2024.

W57 Rabbani, Tahseen, Minghui Liu, Tony O'Halloran, Ananth Sankaralingam, Mary-Anne Hartley, and Furong Huang. "LSH-E Tells You What To Discard: An Adaptive Locality-Sensitive Strategy for KV Cache Compression." Workshop on Machine Learning and Compression, NeurIPS 2024.

W56 Chakraborty, Souradip, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. "MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences." Oral, ICML 2024 Workshop on Models of Human Feedback for AI Alignment, ICML 2024.

W55 Pathmanathan, Pankayaraj, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, and Furong Huang. "Is poisoning a real threat to LLM alignment? Maybe more so than you think." ICML 2024 Workshop on Models of Human Feedback for AI Alignment, ICML 2024.

W54 Panaitescu-Liess, Michael-Andrei, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. "Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?" ICML 2024 Next Generation of AI Safety Workshop, ICML 2024.

W53 An, Bang, Sicheng Zhu, Ruiyi Zhang, Michael-Andrei Panaitescu-Liess, Yuancheng Xu, and Furong Huang. "Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models." ICML 2024 Next Generation of AI Safety Workshop, ICML 2024.

W52 Ding, Mucong, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, and Furong Huang. "SAIL: Self-improving Efficient Online Alignment of Large Language Models." ICML 2024 Workshop on Theoretical Foundations of Foundation Models, ICML 2024.

W51 Chakraborty, Souradip, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Bedi, and Furong Huang. "Transfer Q-star: Principled Decoding for LLM Alignment." Adaptive Learning in Complex Environments TTIC Workshop 2024.

W50 Ding, Mucong, Tahseen Rabbani, Bang An, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. "WAVES: Benchmarking the Robustness of Image Watermarks." Workshop on Reliable and Responsible Foundation Models, ICLR 2024.

W49 Ding, Mucong, Tahseen Rabbani, Bang An, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. "WAVES: Benchmarking the Robustness of Image Watermarks." Workshop on Privacy Regulation and Protection in Machine Learning, ICLR 2024.

W48 Xu, Yuancheng, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. "Shadowcast: Stealthy Data Poisoning Attacks Against Vision-Language Models." Workshop on Secure and Trustworthy Large Language Models, ICLR 2024.

W47 Xu, Yuancheng, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. "Shadowcast: Stealthy Data Poisoning Attacks Against Vision-Language Models." Workshop on Navigating and Addressing Data Problems for Foundation Models, ICLR 2024.

W46 Bornstein, Marco, Amrit Bedi, Anit Kumar Sahu, Furqan Khan, and Furong Huang. "RealFM: A Realistic Mechanism to Incentivize Data Contribution and Device Participation." Workshop on Federated Learning in the Age of Foundation Models (FL@FM-NeurIPS'23), NeurIPS 2023.

W45 Wang, Xiyao, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. "COPlanner: Plan to Roll Out Conservatively but to Explore Optimistically for Model-Based RL." Workshop on Generalization in Planning (GenPlan '23), NeurIPS 2023.

W44 Nguyen, Khanh, Ruijie Zheng, Hal Daume III, Furong Huang, and Karthik Narasimhan. "Progressively Efficient Communication." Intrinsically Motivated Open-ended Learning (IMOL) Workshop, NeurIPS 2023.

W43 Zheng, Ruijie, Yongyuan Liang, Xiyao Wang, Shuang Ma, Hal Daume III, Huazhe Xu, John Langford, Praveen Palanisamy, Kalyan Basu, and Furong Huang. "PREMIER-TACO is a Few-Shot Policy Learner: Pretraining Multitask Representation via Temporal Action-Driven Contrastive Loss." 6th Robot Learning Workshop: Pretraining, Fine-Tuning, and Generalization with Large Scale Models, NeurIPS 2023.

W42 Zheng, Ruijie, Yongyuan Liang, Xiyao Wang, Shuang Ma, Hal Daume III, Huazhe Xu, John Langford, Praveen Palanisamy, Kalyan Basu, and Furong Huang. "PREMIER-TACO is a Few-Shot Policy Learner: Pretraining Multitask Representation via Temporal Action-Driven Contrastive Loss." Workshop on Foundation Models for Decision Making, oral, NeurIPS 2023.

W41 Liu, Xiangyu, Chenghao Deng, Yanchao Sun, Yongyuan Liang, and Furong Huang. "Beyond Worst-case Attacks: Robust RL with Adaptive Defense via Non-dominated Policies." Workshop on Multi-Agent Security: Security as Key to AI Safety, spotlight, NeurIPS 2023.

W40 Zhu, Sicheng, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. "AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models." Workshop on Socially

Responsible Language Modelling Research (SoLaR), NeurIPS 2023.

W39 Agrawal, Aakriti, Rohith Aralikatti, Yanchao Sun, and Furong Huang. "Robustness to Multi-Modal Environment Uncertainty in MARL using Curriculum Learning." Workshop Multi-Agent Security: Security as Key to AI Safety, spotlight, NeurIPS 2023.

W38 Xu, Yuancheng, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. "Equal Long-term Benefit Rate: Adapting Static Fairness Notions to Sequential Decision Making." AdvML-Frontiers workshop, ICML 2023.

W37 Liang, Yongyuan, Yanchao Sun, Ruijie Zheng, Xiangyu Liu, Tuomas Sandholm, Furong Huang, and Stephen McAleer. "Adapting Robust Reinforcement Learning to Handle Temporally-Coupled Perturbations." AdvML-Frontiers workshop, ICML 2023.

W36 An, Bang, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. "Mental Calibration: Discovering and Adjusting for Latent Factors Improves Zero-Shot Inference of CLIP." Workshop on Efficient Systems for Foundation Models (ES-FoMO), ICML 2023.

W35 Chakraborty, Souradip, Amrit Singh Bedi, Alec Koppel, Furong Huang, and Mengdi Wang. "Principal-Driven Reward Design and Agent Policy Alignment via Bilevel-RL." Interactive Learning with Implicit Human Feedback Workshop (ILHF), ICML 2023.

W34 Liu, Xiaoyu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. "C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confounder." The Second Workshop on Spurious Correlations, Invariance and Stability (SCIS), ICML 2023.

W33 Liu, Xiaoyu, Jiaxin Yuan, Bang An, Yuancheng Xu, Yifan Yang, and Furong Huang. "C-Disentanglement: Discovering Causally-Independent Generative Factors under an Inductive Bias of Confounder." Workshop on Structured Probabilistic Inference & Generative Modeling (SPIGM), ICML 2023.

W32 Ding, Peijian, Davit Soselia, Thomas Armstrong, Jiahao Su, and Furong Huang. "Reviving Shift Equivariance in Vision Transformers." The Second Workshop on Spurious Correlations, Invariance and Stability (SCIS), ICML 2023.

W31 Zhu, Sicheng, Bang An, Furong Huang, and Sanghyun Hong. "Learning Unforeseen Robustness from Out-of-distribution Data Using Equivariant Domain Translator." Workshop on Trustworthy ML, ICLR 2023.

W30 Rabbani, Tahseen, Marco Bornstein, and Furong Huang. "PGHash: Large-Scale Distributed Learning via Private On-Device Locally Sensitive Hashing." Sparsity in Neural Networks 2023 workshop, ICLR 2023.

W29 Sun, Yanchao, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. "Certifiably Robust Policy Learning against Adversarial Multi-Agent Communication." Rebellion and Disobedience in AI (RaD-AI) Workshop, AAMAS 2023.

W28 Chakraborty, Souradip, Amrit Singh Bedi, Alec Koppel, Furong Huang, and Dinesh Manocha. "Stein Information Directed Sampling for Efficient Exploration in Model-Based Reinforcement Learning." Oral, AAAI Reinforcement Learning Ready for Production Workshop, 2023.

W27 Bornstein, Marco, Tahseen Rabbani, Evan Wang, Amrit Singh Bedi, and Furong Huang. "SWIFT: Rapid Decentralized Federated Learning via Wait-Free Model Communication." FL-NeurIPS Workshop, oral (11%), NeurIPS 2022.

W26 Zheng, Ruije, Xiyao Wang, Huazhe Xu, and Furong Huang. "Is Model Ensemble Necessary? Model-based RL via a Single Model with Lipschitz Regularized Value Function." Deep Reinforcement Learning Workshop, oral, NeurIPS 2022.

W25 Ding, Mucong, Xiaoyu Liu, Tahseen Rabbani, and Furong Huang. "Faster Hyperparameter Search on Graphs via Calibrated Dataset Condensation." GLFrontiers Workshop, NeurIPS 2022.

W24 Ding, Mucong, Tahseen Rabbani, Bang An, Evan Wang, and Furong Huang. "Sketch-GNN: Scalable Graph Neural Networks with Sublinear Training Complexity." GLFrontiers Workshop, NeurIPS 2022.

W23 Sun, Yanchao, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, and Ashish Kapoor. "SMART: Self-supervised Multi-task pretrAining with contRol Transformers." Foundation Models for Decision Making Workshop, NeurIPS 2022.

W22 Chakraborty, Souradip, Amrit Singh Bedi, Alec Koppel, Furong Huang, Pratap Tokekar, and Dinesh Manocha. "Posterior Coreset Construction with Kernelized Stein Discrepancy for Model-Based Reinforcement Learning." Workshop on Score-Based Methods, NeurIPS 2022.

W21 Liu, Xiangyu, Souradip Chakraborty, and Furong Huang. "Controllable Attack and Improved Adversarial Training in Multi-Agent Reinforcement Learning." Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML), **Outstanding Paper Award**, Oral, NeurIPS 2022.

W20 Xu, Paiheng, Yuhang Zhou, Bang An, Wei Ai, and Furong Huang. "GFairHint: Improving Individual Fairness for Graph Neural Networks via Fairness Hint." workshop on Trustworthy and Socially Responsible Machine Learning (TSRML), NeurIPS 2022.

W19 Borgnia, Eitan, Jonas Geiping, Valeriia Cherepanova, Liam H. Fowl, Arjun Gupta, Amin Ghiasi, Furong Huang, Micah Goldblum, and Tom Goldstein. "DP-InstaHide: Data Augmentations Provably Enhance Guarantees Against Dataset Manipulations." ML Safety workshop, NeurIPS 2022.

W18 Sun, Yanchao, Ruijie Zheng, Parisa Hassanzadeh, Yongyuan Liang, Soheil Feizi, Sumitra Ganesh, and Furong Huang. "Certifiably Robust Multi-Agent Reinforcement Learning against Adversarial Communication." Workshop on Responsible Decision Making in Dynamic Environments, ICML 2022.

W17 Liang, Yongyuan, Yanchao Sun, Ruijie Zheng, and Furong Huang. "Efficient Adversarial Training without Attacking: Worst-Case-Aware Robust Reinforcement Learning." Workshop on Responsible Decision Making in Dynamic Environments, ICML 2022.

W16 Xu, Yuancheng, Yanchao Sun, and Furong Huang. "Everyone Matters: Customizing the Dynamics of Decision Boundary for Adversarial Robustness." Workshop on Continuous Time Methods for Machine Learning, ICML 2022.

W15 Wang, Xiyao, Wichayaporn Wongkamjan, and Furong Huang. "Live in the Moment: Learning Dynamics Model Adapted to Evolving Policy." Workshop on Decision Awareness in Reinforcement Learning (DARL), Spotlight, ICML 2022.

W14 An, Bang, Zora Che, Mucong Ding, and Furong Huang. "Transfer Fairness under Distribution Shifts." Workshop on Socially Responsible Machine Learning, ICLR 2022.

W13 Rabbani, Tahseen, Brandon Feng, Yifan Yang, Arjun Rajkumar, Amitabh Varshney, and Furong Huang. "Comfetch: Federated Learning of Large Networks on Memory-Constrained Clients via Sketching." Workshop on Trustable, Verifiable and Auditable Federated Learning, AAAI 2022.

W12 Sun, Yanchao, Ruijie Zheng, Yongyuan Liang, and Furong Huang. "Who Is the Strongest Enemy? Towards Optimal and Efficient Evasion Attacks in Deep RL." Workshop on Safe and Robust Control of Uncertain Systems (SafeRL), **Best Paper Award**, Oral, NeurIPS 2021.

W11 Liang, Yongyuan, Yanchao Sun, Ruijie Zheng, and Furong Huang. "Efficiently Improving the Robustness of RL Agents against Strongest Adversaries." Workshop on Safe and Robust Control of Uncertain Systems (SafeRL), Oral, NeurIPS 2021.

W10 Ding, Mucong, Kezhi Kong, Jiuhai Chen, John Kirchenbauer, Micah Goldblum, David Wipf, Furong Huang, and Tom Goldstein. "A Closer Look at Distribution Shifts and Out-of-Distribution Generalization on Graphs." Workshop on Distribution Shifts: Connecting Methods and Applications (DistShift), Spotlight, NeurIPS 2021.

W9 Yousefzadeh, Roozbeh, and Furong Huang. "Interpretable Insights about Medical Image Datasets: Using Wavelets and Spectral Methods." Workshop on Human Interpretability in Machine Learning (WHI), ICML 2020.

W8 Huang, Furong, Yu-Xiang Wang, and Xuchen You. "A Spectral Method for Off-Policy Evaluation in Contextual Bandits under Distribution Shift." 13th Annual Machine Learning Symposium, 2019.

W7 Su, Jiahao, Milan Cvitkovic, and Furong Huang. "Sampling-Free Learning of Bayesian Quantized Neural Networks." 13th Annual Machine Learning Symposium, 2019.

W6 Su, Jiahao, Jingling Li, Bobby Bhattacharjee, and Furong Huang. "Tensorial Neural Networks: Generalization of Neural Networks and Application to Model Compression." 13th Annual Machine Learning Symposium, 2019.

W5 Li, Jingling, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. "Understanding of Generalization in Deep Learning via Tensor Methods." Workshop on Generalization, ICML 2019.

W4 Su, Jiahao, Jingling Li, Bobby Bhattacharjee, and Furong Huang. "Tensorial Neural Networks: Generalization of Neural Networks and Application to Model Compression." Women in Machine Learning Workshop, NeurIPS 2018.

W3 Xu, Zheng, Furong Huang, Louiqa Raschid, and Tom Goldstein. "Non-negative Factorization of the Occurrence Tensor from Financial Contracts." Tensor-Learn Workshop, NIPS 2016.

W2 Huang, Furong, Niranjan U.N., Ioakeim Perros, Robert Chen, Jimeng Sun, and Animashree Anandkumar. "Scalable Latent Tree Model and its Application to Health Analytics." Machine Learning in Healthcare Workshop, NIPS 2015.

W1 Huang, Furong, Sergiy Matusevych, Anima Anandkumar, Nikos Karampatziakis, and Paul Mineiro. "Distributed Latent Dirichlet Allocation via Tensor Factorization." Optimization for Machine Learning Workshop, NIPS 2014.

Research Software, Datasets, and Benchmarks

8. **Agentic Critical Training (ACT).** Training framework for self-improving foundation models. Project page.
7. **TSRBench.** Multi-task, multimodal time-series reasoning benchmark for generalist models. Benchmark and project page.
6. **TraceGen.** World modeling in 3D trace space for learning from cross-embodiment videos. Project page; code.
5. **MomaGraph.** State-aware unified scene graphs for embodied task planning. Project page; code.
4. **ROVER.** Benchmark for reciprocal cross-modal reasoning in omnimodal generation. Project page; code.
3. **PropensityBench.** Agentic evaluation suite for latent safety risks in large language models. Code and benchmark.
2. **Zebra-CoT.** Dataset for interleaved vision-language reasoning. Dataset.
1. **Imagine, Verify, Execute.** Agentic exploration framework for vision-language models. Project page; code.

Awards and Honors

28. NVIDIA Academic Grant, 2026.
27. Best Paper Award at the New Frontiers in Adversarial Machine Learning (AdvML Frontier) Workshop, NeurIPS 2024.
26. National Science Foundation (NSF) National Artificial Intelligence Research Resource (NAIRR) Pilot project NAIRR240045 “Guardians of Integrity in AI: Establishing Trust, Originality, and Ethical Standards (GATES),” 2024.
25. Adobe Research Gift Grant, 2024.
24. Microsoft Accelerate Foundation Models Research award, 2023.
23. CloudBank Computing Award for Research on Fair Machine Learning in the Long Run, 2023.
22. Outstanding Paper Award at the Trustworthy and Socially Responsible Machine Learning Workshop, NeurIPS 2022.
21. Winner of MIT Technology Review Innovators Under 35 (TR35) Asia Pacific, 2022.
20. Finalist of AI in Research - AI researcher of the year, 2022 Women in AI Awards North America, 2022.
19. Special Jury Recognition of Women in AI Awards North America, 2022.
18. JP Morgan Faculty Research Award, “Repelling Security Vulnerabilities in AI-augmented Financial Decision-Making Systems,” 2022.
17. Best Paper Award at Safe and Robust Control of Uncertain Systems, NeurIPS 2021.
16. IEEE Senior Member, Jan. 2021.
15. JP Morgan Faculty Research Award, “Robust, Private and Fair ML for Financial Models,” 2020.
14. JP Morgan Faculty Research Award, “Methods to Identify Communities and Trading Behavior over Financial Data Streams,” 2019.
13. Capital One Faculty Research Award, “Unsupervised Learning via Probabilistic Models,” 2018.
12. Adobe Faculty Research Award, “Understanding User Features with Text, Social and Behavior Data through Deep Tensor Neural Network Learning,” 2017.
11. Travel Grant Award: Heidelberg Laureate Forum, 2017.
10. Travel Grant Award: CRA-W Early Career Mentoring Workshop, 2016.
9. MLconf Industry Impact Student Research winner sponsored by Google, San Francisco, 2015.
8. Travel Grant Award: NIPS, 2015.
7. Travel Grant Award: Junior Scientist Workshop on Machine Learning and Computer Vision, 2015.
6. Electrical Engineering and Computer Science Department Graduate Fellowship, University of California Irvine, 2010-2014.
5. Travel Grant Award: WiML, 2014.
4. Travel Grant Award: CRA-W Grad. Cohort Workshop, 2012.
3. Travel Grant Award: WiML, 2012.
2. Department Fellowship, 2010.
1. Best Undergraduate Thesis Award, Zhejiang University, 2010.

Grants

Total: $13,278,032, My Share: $9,229,399.

20. Feizi, S. (Lead Investigator), Huang, F. (Co-Investigator), Wilson, A. (Co-Investigator), “Consistency, Robustness and Generalization Guarantees for Understanding Language Models,” Sponsored by DARPA-Defense Advanced Research Projects Agency, $1,800,000. (July 1, 2025 - June 30, 2028).
19. Huang, F. (Lead Investigator), Goldstein, T. (Lead Investigator), Li, B. (Co-Investigator), Martin-Martin, R. (Co-Investigator), Zhu, Y. (Co-Investigator), “RAMSES: A Domain-Agnostic System with Robust Features and Symbolic Reasoning,” Sponsored by DARPA-Defense Advanced Research Projects Agency, $2,800,000. (July 1, 2024 - June 30, 2027).
18. Huang, F. (Single Lead Investigator), “Securing Retrieval-Augmented Generation (RAG) Systems,” Sponsored by Peraton Inc., $100,037. (01/01/2025 - 12/31/2025).
17. Huang, F. (Single Lead Investigator), “Encoded Reasoning in Agentic Systems, FAKE-LLM: Knowledgeable Exploitation, and Reasoning-Trace Augmented Training,” Sponsored by Good Ventures (recommended by Open Philanthropy), $1,118,700. (October 1, 2025–September 30, 2026).
16. Huang, F. (Single Lead Investigator), “Robust and Efficient Machine Learning Project 2,” Sponsored by Booz Allen and Hamilton, Inc. (Prime: Air Force Material Command), $236,178. (Mar 15, 2024 - Mar 14, 2025).
15. Huang, F. (Single Lead Investigator), “E-VERIFY: Robust and Efficient Ultra-large Sized Graph Learning - Project 13,” Sponsored by Booz Allen and Hamilton, Inc. (Prime: Air Force Material Command), $150,000. (Feb 13, 2023 - Feb 12,

2024).

14. Huang, F. (Single Lead Investigator), “Autonomous Decision-Making Systems under (Adversarial) Distribution Shifts: Learning to Adapt, Generalize and Robustify,” Sponsored by the Air Force Office of Scientific Research (AFOSR), $517,881. (December 15, 2022–December 14, 2025).
13. Huang, F. (Single Lead Investigator), “Repelling Security Vulnerabilities in AI-Augmented Financial Decision-Making Systems,” Sponsored by J.P. Morgan, $80,000. (Fall 2022–Fall 2023).
12. Huang, F. (Single Lead Investigator), “Trustworthy Real-time Communicative Multi-Agent Reinforcement Learning for Decision-making in Unstable & Adversarial Systems”, Sponsored by Office of Naval Research, $863,877. (June 1, 2022 - May 30, 2025).
11. Huang, F. (Lead Investigator), Wu, M. (Co-Investigator), Dachman-Soled D. (Co-Investigator), “FAI: Toward Fair Decision Making and Resource Allocation with Application to AI-Assisted Graduate Admission and Degree Completion”, Sponsored by National Science Foundation and Amazon, $1,000,000. (February 15, 2022 - January 31, 2025).
10. Huang, F. (Single Lead Investigator), “E-VERIFY: Task 11: Efficient and Expressive State-of-the-art Deep Learning Model Design through Tensor Methods,” Sponsored by Booz Allen and Hamilton, Inc. (Prime: Air Force Material Command), $180,000. (June 1, 2022 - May 30, 2023).
9. Huang, F. (Single Lead Investigator), “CRII: RI: Principled Methods for Learning and Understanding of Neural Networks,” Sponsored by National Science Foundation, $175,000. (May 1, 2019 - April 30, 2022).
8. Huang, F. (Single Lead Investigator), “E-VERIFY: Task 10: TNN Based Machine Learning 2 (FY21),” Sponsored by Booz Allen and Hamilton, Inc. (Prime: Air Force Material Command), $ 180,000. (August 23, 2021 - August 22, 2022).
7. Goldstein, T. (Lead Investigator), Huang, F. (Co-Investigator), “Robust, Private and Fair ML for Financial Models,” Sponsored by J. P. Morgan, $150,000 (my share: $75,000). (May 1, 2020 - August 31, 2021).
6. Raschid, L. (Lead Investigator), Huang, F. (Co-Investigator), Rossi, A. (Co-Investigator), “Methods to Identify Communities and Trading Behavior over Financial Data Streams,” Sponsored by J. P. Morgan, $150,000 (my share: $75,000). (March 4, 2019 - March 3, 2021).
5. Huang, F. (Single Lead Investigator), “E-VERIFY: Task 6: High-Performance Machine Learning with Tensorial Neural Networks on Heterogeneous Systems,” Sponsored by Booz Allen and Hamilton, Inc. (Prime: Air Force Material Command), $180,000. (January 1, 2020 - January 31, 2021).
4. Goldstein, T. (Lead Investigator), Katz, J. (Co-Investigator), Huang, F. (Co-Investigator), Shrivastava, A. (Co-Investigator), Jacobs, D. (Co-Investigator), Dickerson, J. (Co-Investigator), “Repelling Evasion and Poisoning Attacks: A Principled Way Forward,” Sponsored by DOD-DARPA-Defense Advanced Research Projects Agency, $3,226,359 (my share: $537,726). (December 12, 2019 - November 30, 2023).
3. Raschid, L. (Lead Investigator), Huang, F. (Co-Investigator), “Customizing Machine Learning Approaches to Interpret Patterns in Financial Data Streams,” Sponsored by AI in Business and Society Seed Grant, $20,000 (my share: $10,000). (2018).
2. Huang, F. (Lead Investigator), “Unsupervised Learning via Probabilistic Models,” Sponsored by Capital One, $300,000. (2018–present).
1. Huang, F. (Single Lead Investigator), “Understanding User Features with Text, Social and Behavior Data through Deep Tensor Neural Network Learning,” Sponsored by Adobe, $50,000. (2017).

Academic Talks

90. Panelist, “Vision, Multimodal AI & Agents,” KAIST@ICML 2026, Seoul, South Korea, Jul. 2026.
89. Invited speaker at the ITGenNexus Workshop, IEEE International Symposium on Information Theory (ISIT), “Reasoning as Control: Adaptive Test-Time Compute for Planning Agents,” Guangzhou, China, Jul. 2026.
88. Panelist, “From Information to Intelligence: What Can These Communities Learn from Each Other?,” ITGenNexus Workshop at IEEE ISIT, Guangzhou, China, Jul. 2026.
87. Lightning talk at the Workshop on Reasoning and Planning, hosted by Apple, “From More Thinking to Better Thinking: Adaptive Test-Time Control for Reasoning and Planning Agents,” Jun. 2026.
86. Invited speaker at DatologyAI’s Summer of Data Seminar, “A Few Early Steps Away: Building Self-Improving Foundation Models—Auditors, Actuators, and Amplifiers for Trustworthy AI,” Jun. 2026.
85. Invited speaker at Scale AI’s ML Research Seminar, “A Few Early Steps Away: Building Self-Improving Foundation Models—Auditors, Actuators, and Amplifiers for Trustworthy AI,” Jun. 2026.
84. Invited speaker at the ICLR Workshop on Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA), “Rethinking Test-Time Thinking: From Token-Level Rewards to Robust Generative Agents,” Apr. 2026.
83. Panelist at the ICLR Workshop on Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA), Apr. 2026.
82. Colloquium at Johns Hopkins University, “Rethinking Test-Time Compute,” 2026.

81. Keynote speaker at RL Workshop at IISc, “From Perception to Action: World Model Learning for Generalist Agents”, Jan., 2026.
80. Keynote speaker at NeurIPS Workshop on Generative AI in Finance, “LLMs Are Not Accountants: Data Without Disclosure via Rule-Grounded World Models for Financial Ledgers”, NeurIPS, Dec., 2025.
79. Keynote speaker at the 4th workshop on Computer Vision in the Wild, “From Perception to Action: World Model Learning for Generalist Agents”, CVPR, Jun., 2025.
78. Keynote speaker at the 2nd Texas Colloquium on Distributed Learning (TL;DR), “Rethinking Test-Time Thinking: From Token-Level Rewards to Robust Generative Agents”, May., 2025.
77. Keynote speaker at the GenAI Watermarking Workshop, “Erasing the Invisible: Riding the WAVES of Watermark Robustness”, at ICLR, Apr., 2025.
76. Keynote speaker at the World Models Workshop, “Foundation Model for Sequential Decision-Making: A World-Model Learning Perspective”, at ICLR, Apr., 2025.
75. Spotlight speaker at the Alignment Workshop, “Aligning AI on the Fly,” Singapore, Apr. 2025.
74. Invited Talk at the 2024 IEEE International Conference on Image Processing (ICIP) Workshop SPVis ’24 – Security and Privacy of Machine Learning-Based Vision Processing in Autonomous Systems, “Crafting and Cracking AI in the Shadows of Language- Poison Data and Jailbreak Prompts for LLMs,” Oct., 2024.
73. Keynote speaker at the New York Academy of Sciences, 15th Annual Machine Learning Symposium, “Towards Generative AI Security: An Interplay of Stress-Testing and Alignment,” Oct., 2024.
72. Invited talk at UT Austin AI Institute Seminar (IFML), “Foundation Models for Sequential Decision-Making,” Oct., 2024.
71. Invited talk at Vanderbilt University ML Seminar, “Integrity in AI: Multi-Modality Approaches to Combat Misinformation for Content Authenticity,” Sep., 2024.
70. Invited talk at Capital One, “Foundation Models for Sequential Decision-Making,” Sep., 2024.
69. Invited talk at MSRNYC Seminar, “Foundation Models for Sequential Decision-Making,” Aug., 2024.
68. Keynote speaker at the 1st CVPR Workshop on Dataset Distillation (DDCV) at CVPR, “Advancing AI with Data-Centric Strategies—Boosting Efficiency, Generalization, and Trust,” Jun., 2024.
67. Invited talk at AI Community of Practice Seminar at the U.S. Securities and Exchange Commission (SEC), “Integrity in AI: Multi-Modality Approaches to Combat Misinformation for Content Authenticity,” May, 2024.
66. Invited talk at JP Morgan, “Invisible Foes: Crafting and Cracking AI in the Shadows of Language,” May, 2024.
65. Invited talk at Maryland Robotics Center (MRC) Symposium, “Foundation Models for Sequential Decision-Making,” May, 2024.
64. Invited talk at Qualcomm AI Security Lecture Series, “Invisible Foes: Crafting and Cracking AI in the Shadows of Language — Poison Finetuning Data and Jailbreak Prompts for LLMs,” Mar., 2024.
63. Seminar talk at Values-Centered Artificial Intelligence (VCAI) seminar series, “Algorithmic Fairness in an Ever-Changing World,” College Park, MD, Feb., 2024.
62. Lightning talk at NSF-Amazon FAI PI meeting, “Algorithmic Fairness in an Ever-Changing World,” Amazon HQ2, Arlington, VA, Jan., 2024.
61. Department Colloquium at University of Maryland, College Park, “Trustworthy Machine Learning in an Ever-Changing World,” Sep., 2023.
60. Panelist at Interactive Learning with Implicit Human Feedback workshop, ICML, Jul., 2023.
59. Keynote speaker at ROADS to Mega-AI Models Workshop, MLSys, “Efficient Machine Learning at the Edge,” Jun., 2023.
58. Invited talk at the 3rd Workshop of Adversarial Machine Learning on Computer Vision: Art of Robustness, “Robust Reinforcement Learning in an Ever-Changing World,” CVPR, Jun., 2023.
57. Invited talk at Reincarnating RL workshop, “Adaptable Reinforcement Learning in An Ever-Changing World,” ICLR, May., 2023.
56. Panelist at Reincarnating RL workshop, ICLR, May., 2023.
55. Spotlight talk at NSF-IEEE workshop: Toward Explainable, reliable, and sustainable machine learning in signal & data science, “Trustworthy Machine Learning in Complex Environments,” Mar., 2023.
54. Invited talk at 57th Annual Conference on Information Science and Systems, CISS, “Efficient Machine Learning at the Edge in Parallel,” Mar., 2023.
53. Invited talk at the 2023 Information Theory and Applications Workshop (ITA), “Trustworthy Machine Learning in Complex Environments,” Feb. 2023.
52. Invited talk at UTSA Matrix AI Seminar, “Trustworthy Machine Learning in Complex Environments,” Jan., 2023.
51. Invited talk at 2022 MSR Asia Theory Workshop on Data-driven Optimization, “Efficient Machine Learning at the Edge in Parallel,” Nov., 2022.

50. Invited talk at UCSD, “Trustworthy Machine Learning in Complex Environments,” Oct., 2022.
49. Invited talk at Caltech, “Trustworthy Machine Learning in Complex Environments,” Oct., 2022.
48. Invited talk at UCLA, “Trustworthy Machine Learning in Complex Environments,” Oct., 2022.
47. Invited talk at USC, “Trustworthy Machine Learning in Complex Environments,” Oct., 2022.
46. Invited talk at UCI, “Trustworthy Machine Learning in Complex Environments,” Oct., 2022.
45. Invited talk at SIAM Conference on Mathematics of Data Science (MDS22), “Data-driven Design of Transformers for Improved Generalization or Efficiency,” Oct., 2022.
44. Contributed talk at ICML workshop Decision Awareness in Reinforcement Learning, “Live in the Moment: Learning Dynamics Model Adapted to Evolving Policy,” Jul., 2022.
43. Invited talk at the Applied Statistics Symposium at the ICSA (International Chinese Statistical Association) International Conference, “Spectral Methods in Deep Learning,” Sep., 2021.
42. Invited talk at Code Girls Club at Montgomery Blair High School, “Learning in High-dimensional Data,” Jun., 2021.
41. Invited talk at Institute for Pure & Applied Mathematics (IPAM), Efficient Tensor Representations for Learning and Computational Complexity Workshop, “ Understanding, Interpreting and Design Neural Network Models Through Tensor Representation,” May, 2021.
40. Invited talk at the annual East Coast Optimization Meeting (ECOM), “Escaping from saddle points using asynchronous coordinate gradient descent,” Apr., 2021.
39. Invited talk at the SIAM CSE21 minisymposium “High-performance tensor computations in scientific computing and machine learning,” “Tensors in Modern Machine Learning,” Mar., 2021.
38. Invited talk at ARO Sponsored Workshop on Assured Autonomy, “Poisoning Attacks in Reinforcement Learning,” Jun., 2020.
37. Invited talk at High Performance Computing and Data Analytics Workshop, “Efficient Machine Learning via Tensorial Neural Networks,” Sep., 2019.
36. IEEE CIS/SPS lecture, “Understanding of Generalization in Deep Learning via Tensor Methods,” May 1, 2019.
35. Invited talk at Virginia Tech AI-AI Workshop, AI Workshop, “Tensorial Neural Networks: Generalization of Neural Networks and Application to Model Compression,” May, 2019.
34. Invited talk at ICERM Workshop, Theory and Practice in Machine Learning and Computer Vision, “Tensorial Neural Networks: Generalization of Neural Networks and Application to Model Compression,” Feb., 2019.
33. Seminar talk at University of Montreal, “Theory in Deep Learning,” Dec., 2018.
32. Invited talk at UMD-NCI Data Science Workshop, “Discovery of Latent Factors in High-dimensional Data via Spectral Methods,” Nov., 2018.
31. Invited talk at the 55th Annual Allerton Conference on Communication, Control, and Computing, session on Large-Scale Optimization, “Principled Methods for Learning and Understanding of Neural Networks,” Monticello, Illinois, Oct., 2018.
30. Invited talk at Workshop on Quantum Machine Learning, “Discovery of Latent Factors in High-dimensional Data via Spectral Methods,” Sep., 2018.
29. Seminar talk at Virginia Tech, “Discovery of Latent Factors in High-dimensional Data via Spectral Methods,” Oct., 2018.
28. Talk at Deep Learning RIT, “Principled Methods for Learning and Understanding of Neural Networks,” University of Maryland, Oct., 2018.
27. Seminar talk at Laboratory for Telecommunication Science (LTS), “Guaranteed Learning of Latent Variable Models through Tensor Methods,” Jul., 2018.
26. Invited tutorial at ACM Sigmetrics, Irvine, California, Jun., 2018.
25. Seminar talk at Laboratory for Telecommunication Science (LTS), “Fast Detection of Overlapping Communities via Online Tensor Methods on GPUs,” Nov., 2017.
24. Undergraduate Honors Guest Lecture, “Matrices and Tensors in Machine Learning,” Nov., 2017.
23. Seminar talk at the Department of Computer Science, “Guaranteed Learning of Latent Variable Models through Tensor Methods,” Virginia Tech, Nov., 2017.
22. Department Colloquium at Computer Science Department, “Guaranteed Learning of Latent Variable Models through Tensor Methods,” University of Maryland College Park, Oct., 2017.
21. Invited talk at the Matrix and Tensor Factorization Methods in Computer Vision, ICCV conference, Venice, Italy, Oct., 2017.
20. Invited talk at the 55th Annual Allerton Conference on Communication, Control, and Computing, session on Large-Scale Optimization, “Nonconvex Optimization in Tensor Decomposition,” Monticello, Illinois, Oct., 2017.
19. Seminar talk at New York University, “Guaranteed Learning of Latent Variable Models through Tensor Methods,” New York City, Apr., 2017.
18. Spotlight talk at 11th Annual Machine Learning Symposium, “Convolutional Dictionary Learning through Tensor

Factorization," New York City, Mar., 2017.

17. Welcome and opening remark at the live streaming event for "Women in Data Science" conference at Stanford University, located in New York City, Feb., 2017.
16. Tea talk at Microsoft Research New York City, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Nov., 2016.
15. Invited talk at Machine Learning conference, "Guaranteed Learning of Latent Variable Models through Tensor Methods," New York City, Apr., 2016.
14. Invited talk at University of Maryland, College Park, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Mar., 2016.
13. Invited talk at IBM T. J. Watson Research Center in Yorktown Heights, "Guaranteed Learning of Latent Variable Models through Tensor Methods," New York, Feb., 2016.
12. Invited talk at Microsoft Research New York City, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Feb., 2016.
11. Invited talk at Simons Foundation New York City, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Feb., 2016.
10. Invited talk at Boston University, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Mar., 2016.
9. Invited talk at Tufts University, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Mar., 2016.
8. Invited talk at University of California Davis, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Mar., 2016.
7. Invited talk at Pennsylvania State University, "Guaranteed Learning of Latent Variable Models through Tensor Methods," Mar., 2016.
6. Invited research talk and tutorial "Method of Moments and Tensor Decomposition" at Janelia Research Campus, Ashburn, Virginia, Oct., 2015.
5. Oral presentation in COLT, "Escaping from Saddle Points — Online Stochastic Gradient for Tensor Decomposition," Paris, France, Jul., 2015.
4. Spotlight talk at Non-convex optimization NIPS Workshop, "Escaping from Saddle Points — Online Stochastic Gradient for Tensor Decomposition," Dec., 2015.
3. Spotlight talk "Method of Moments and Tensor Decomposition" at BigNeuro NIPS Workshop, Dec., 2015.
2. Contributed Talk "Fast Detection of Overlapping Communities Via Online Tensor Methods" at Topic Model NIPS Workshop, Dec., 2013.
1. Invited Talk in AI/ML seminar, Donald Bren Hall ICS department at UC Irvine, Mar., 2013.

Editorial and Peer-Review Service

7. **Action Editor:** Transactions on Machine Learning Research (TMLR), Jan. 2024–present.
6. **Guest Editor:** Frontiers Research Topic "Tensor Methods for Deep Learning," 2021.
5. **Senior Area Chair:** ICLR (2025, 2026; appointed for 2027); ICML (2026).
4. **Area Chair:** AAAI (appointed for 2027); CoRL (2026); NeurIPS (2021–2025); ICML (2019, 2025); ICLR (2021–2024).
3. **Program Committee:** IJCAI-ECAI (2018); SysML (2018, 2019).
2. **Selected journal reviewing:** *Annals of Statistics*; *Journal of Machine Learning Research*; *IEEE/ACM Transactions on Networking*.
1. **Selected conference reviewing:** AAAI, AISTATS, CVPR, GLOBECOM, ICLR, ICML, IJCAI, ISIT, MobiHoc, MILCOM, and NeurIPS.

Media and Public Engagement

12. UMD's PropensityBench framework was featured in CMNS News after Meta used it to evaluate the safety of a new multimodal AI model, Apr. 2026.
11. Research on latent safety risks in agentic AI was featured in CMNS News, Dec. 2025.
10. Research on security threats to web AI agents was featured in CMNS News, May 2025.
9. Commercialization of AI-safety research was featured in CMNS News, Mar. 2025.
8. Interviewed by the Institute for Trustworthy AI in Law & Society (TRAILS) for "TRAILS Researchers Discuss How AI Will Change the Way We Live," May 2024.
7. Hosted a public Reddit Ask Me Anything on artificial intelligence and machine learning, May 2024.
6. Participation in the federal National Artificial Intelligence Research Resource pilot was featured by UMIACS, May 2024.

5. Research on autonomous jailbreak attacks and large-language-model safety was featured by Science News Explores, Apr. 2024, and Science News, Feb. 2024.
4. Research on the limits of AI-generated text detection was named one of the CMNS Top Stories of 2023, Dec. 2023, and covered by TechXplore, May 2023.
3. Research on fairness in AI and the NSF–Amazon Fairness in AI program was featured in Maryland Today, Apr. 2022.
2. Best Paper recognition at the NeurIPS 2021 SafeRL Workshop was featured by the UMD Department of Computer Science, Jan. 2022.
1. UMD's partnership with Capital One was covered by the Washington Business Journal, Nov. 2018.

TEACHING, MENTORING, AND ADVISING ACTIVITIES

Courses Taught

15. Summer 2026, Generative AI Agent, DeepLearn 2026, 13th International School on Deep Learning.
14. Spring 2026, CMSC 422, *Introduction to Machine Learning* (course website)
13. Fall 2025, CMSC 848N, Generative AI Agents(course website)
12. Spring 2024, CMSC 422, *Introduction to Machine Learning* (course website)
11. Fall 2023, CMSC 742, *Algorithms in Machine Learning: Guarantees and Analyses* (course website)
10. Spring 2023, CMSC 742, *Algorithms in Machine Learning: Guarantees and Analyses* (course website)
9. Fall 2022, CMSC 422, *Introduction to Machine Learning* (course website)
8. Spring 2022, CMSC 422, *Introduction to Machine Learning* (course website)
7. Fall 2021, CMSC 742, *Algorithms in Machine Learning: Guarantees and Analyses* (course website)
6. Spring 2021, CMSC 422, *Introduction to Machine Learning* (course website)
5. Fall 2020, CMSC 828U, *Algorithms in Machine Learning: Guarantees and Analyses* (course website)
4. Spring 2019, CMSC 828U, *Algorithms in Machine Learning: Guarantees and Analyses* (course website)
3. Fall 2018, CMSC 498V, *Advanced Topics in Machine Learning* (course website)
2. Spring 2018, CMSC 422, *Introduction to Machine Learning* (course website)
1. Fall 2017, CMSC 828R, *Topics in Machine Learning* (course website)

Course and Curriculum Development

3. Developed CMSC848N, a graduate course on Generative AI Agents, including lectures and projects focused on LLM-based reasoning, planning, tool use, memory, and evaluation. Designed new assignments and hands-on projects on building robust agentic systems for real-world applications.
2. Developed new Machine Learning graduate course, including lectures and projects, focusing on ML methods that provide theoretical guarantees on performance. Methods taught are applicable to real-world problems, including social networks, biological systems, and natural language processing.
1. Developed new topics for undergraduate course, including Learning Theory, Latent Variable Models, and principled methods in Deep Learning. Developed new assignments and projects for these topics.

Research Advising

Doctoral Students and Research Collaborators

Formal Ph.D. advisees and research collaborators are identified explicitly below.

42. Jiawei Xu, Ph.D. advisee, 2025–present
41. Andrew Mendez, Ph.D. advisee, 2025–present
40. Minghui Liu, Ph.D. advisee, 2025–present
39. Sy-Tuyen Ho, Ph.D. advisee, 2025–present
38. Weize Liu, Ph.D. advisee, 2025–present
37. Lingzhi Yuan, Ph.D. advisee, 2025–present
36. Yoonkyo Jung, Ph.D. advisee, 2025–present
35. Shayan Shabihi, Ph.D. advisee, 2025–present

34. Seungjae (Jay) Lee, Ph.D. advisee, 2024–present
33. Michael-Andrei Panaitescu-Liess, Ph.D. advisee, 2024–present
32. Cheryl Yongyuan Liang, Ph.D. advisee, 2024–present
31. Aakriti Agrawal, Ph.D. advisee, 2023–present
30. Souradip Chakraborty, Ph.D. advisee, 2022–present
29. Pankayaraj Pathmanathan, Ph.D. advisee, 2022–present
28. Chenghao Deng, Ph.D. advisee, 2021–present
27. Xiyao Wang, Ph.D. advisee, 2022–2026 (Tencent Hunyuan, North America)
26. Mucong Ding, Ph.D. advisee, 2021–present
25. Frank Ruijie Zheng, Ph.D. advisee, 2022–present
24. Sicheng Zhu, Ph.D. advisee, 2022–2025
23. Bang An, Ph.D. advisee, 2021–2025
22. Xiaoyu Liu, Ph.D. advisee, 2020–2024
21. Marco Bornstein, Ph.D. advisee, 2020–2025
20. Yuancheng Xu, Ph.D. advisee, 2020–2025
19. Yifan Yang, doctoral research mentee, 2020–2025
18. Xiangyu Liu, Ph.D. advisee, 2022–2024
17. Tahseen Rabbani, Ph.D. advisee, 2019–2024
16. Zora Che, Ph.D. advisee, 2023–2024
15. Jiaxin Yuan, doctoral research mentee, 2022–2026
14. Wichayaporn Wongkamjan, Ph.D. advisee, 2021–2024
13. Yanchao Sun, Ph.D. advisee, 2018–2023
12. Jiahao Su, Ph.D. advisee, 2017–2022
11. Zhi Zhang, Ph.D. advisee, 2020–2021
10. Chen Chen, Ph.D. advisee, 2019–2021
9. David Haoran Chan, Ph.D. advisee, 2021
8. MG Hirsch, Ph.D. advisee, 2021
7. Jin-peng Liu, research collaborator, 2018–2020
6. Tongyang Li, research collaborator, 2017–2020
5. Xuchen You, Ph.D. advisee, 2017–2019
4. Jingling Li, Ph.D. advisee, 2017–2019
3. Yogesh Balaji, research collaborator, 2017–2019
2. Seyed Abdulaziz Esmaeili, Ph.D. advisee, 2017–2018
1. Jialin Li, Ph.D. advisee, 2017–2018

Master's Research Mentoring

All entries below are M.S. research advisees.

10. Dat Dao, 2025–present
9. Xinchen Yu, 2025–present
8. Rohit Suresh, 2025–present
7. Andrew Zheng, 2025–present
6. Mohneet Kaur, 2026–present
5. Kimia Samieinejad, 2025–present
4. Ignacio Hernández Bas, 2025–present
3. Quan Nguyen, 2025–present
2. Vinay Samuel, 2025–present
1. Sudhamshu Dandu, 2025–present

Undergraduate Research Mentoring

All entries below are undergraduate research mentees.

47. Raghav Thind, 2026 - present
46. Sarvesh Baskar, 2025 - present
45. Muhammad Islam, 2026 - present
44. Kaushal Janga, 2025 - present

43. Andrew Wang, 2025 - present
42. Ankit Nakhawa, 2025 - present
41. Nandhu Pillai, 2026 - present
40. Amar Dhillon, 2026 - present
39. Jude Lwin, 2025 - present
38. Arjun Rajaram, 2025 - present
37. Aayush Talreja, 2025 - present
36. Hyunwoo Jae, 2025 - 2026
35. Keenan Powell, 2025 - 2026
34. Neel Jay, 2025 - 2026
33. Sungbin Oh, 2025 - 2026
32. Aadi Palnitkar, 2025 - 2026
31. Rohit Karthickeyan, 2025 - 2026
30. Satya Shah, 2025 - 2026
29. Anirudh Pappu, 2025 - 2026
28. Jess Choe, 2025 - 2026
27. Mingzhi Chen, 2025 - 2026
26. Shile Li, 2025 - 2026
25. Jiaming Meng, 2025 - 2026
24. Ignacio Hernandez Bas, 2025 - 2026
23. Baozan Yan, 2025 - 2026
22. Andrew Xu, 2025 - 2026
21. Anirudh Satheesh, 2023 - present
20. Justin Huang, 2022-2023
19. Benjamin Kreiswirth, 2022-2023
18. Evan Wang, 2021-2023
17. Zora Che, 2021-2023
16. Cheryl Yongyuan Liang, 2021-2023
15. Frank Ruijie Zheng, 2020-2022
14. Kate Hu, 2020-2021
13. Arjun Rajkumar, 2019-2021
12. Huimin Zeng, 2019-2021
11. Elly Do, 2021
10. George Li, 2021
9. Son Phan, 2020
8. Shiqi Wang, 2019-2020
7. Xiangyu Ying, 2019-2020
6. Zhixuan Zhou, 2019-2020
5. Amy Huddell, 2019
4. Ruochun Wang, 2019
3. Jeremy Hu, 2018- 2020
2. Christopher DeCarolis, 2017- 2019
1. Yu Wang, 2017- 2018

Junior Faculty Mentoring

2. Fumeng Yang, 2024–present.
1. Sanghamitra Dutta, 2022–present.

Ph.D. and M.S. Committee and Examination Service

Unless another institution is named, service was at the University of Maryland. UMD units: CS (Computer Science), ECE (Electrical and Computer Engineering), AMSC (Applied Mathematics, Statistics, and Scientific Computation), INFO (Information Studies), and CMNS (College of Computer, Mathematical, and Natural Sciences).

79. Jiuhai Chen (Ph.D.) — Committee member, CMNS (2026).
78. Dylan Sam (Ph.D.) — Ph.D. committee member, Carnegie Mellon University (2025–2026). Advisor: Zico Kolter.
77. Yingchen Xu (Ph.D.) — External Ph.D. examiner, UCL (2026). Advisors: Tim Rocktäschel and Edward Grefenstette.

76. Tom Sander (Ph.D.) — External thesis reviewer (rapporteur) and defense jury member, École Polytechnique / Meta FAIR (2026). Advisors: Alain Durmus and Chuan Guo.
75. Sanae Lotfi (Ph.D.) — Dissertation committee member, NYU (2025). Advisor: Andrew Gordon Wilson.
74. Hanyu Wang (Ph.D.) — Committee member, CMNS (2025).
73. Yuxin Wen (Ph.D.) — Committee member, CS (2025).
72. Faisal Adamu Hamman (Ph.D.) — Committee member, ECE (2025).
71. Shuaiyi Huang (Ph.D.) — Committee member, CS (2025).
70. Wenxiao Wang (Ph.D.) — Committee member, CS (2025).
69. Ryan Peter Sullivan (Ph.D.) — Committee member, CS (2025).
68. Samyadeep Basu (Ph.D.) — Committee member, CS (2025).
67. Fuxiao Liu (Ph.D.) — Committee member, CS (2025).
66. Dehao Yuan (Ph.D.) — Committee member, CS (2025).
65. Yuhang Zhou (Ph.D.) — Dean's representative, INFO (2025).
64. Arpit Bansal (Ph.D.) — Committee member, CS (2025).
63. Zachary McBride Lazri (Ph.D.) — Committee member, ECE (2025).
62. Gowthami Somepalli (Ph.D.) — Committee member, CS (2024).
61. Sakila S. Jayaweera Samaranayake (Ph.D.) — Committee member, ECE (2024).
60. Wei-Hsiang Wang (Ph.D.) — Committee member, ECE (2024).
59. Alireza Ganjdanesh (Ph.D.) — Committee member, CS (2024).
58. Neha Mukund Kalibhat (Ph.D.) — Committee member, CS (2024).
57. Wenhan Xian (Ph.D.) — Committee member, CS (2024).
56. Zhengmian Hu (Ph.D.) — Committee member, CS (2024).
55. Priyatham Kattakinda (Ph.D.) — Committee member, ECE (2024).
54. Aquia Richburg (Ph.D.) — Committee member, AMSC (2024).
53. Kezhi Kong (Ph.D.) — Committee member, CS (2024).
52. Ruichen Wang (Ph.D.) — Committee member, ECE (2024).
51. David Li (Ph.D.) — Committee member, CS (2024).
50. Thomas David Wrona (M.S.) — Committee member and Dean's representative, CS (2023).
49. Henok Gebeyehu (M.S.) — Committee member, ECE (2023).
48. Neha Kalibhat (Ph.D.) — Preliminary examination committee member, CS (2023).
47. Renkun Ni (Ph.D.) — Committee member, CS (2023).
46. Bo He (Ph.D.) — Committee member, CS (2023).
45. Wenhan Xian (Ph.D.) — Preliminary examination committee member, CS (2023).
44. Manli Shu (Ph.D.) — Committee member, CS (2023).
43. Shlok Kumar Mishra (Ph.D.) — Committee member, CS (2023).
42. Biao Jia (Ph.D.) — Committee member, CS (2023).
41. Alexander Jacob Levine (Ph.D.) — Committee member, CS (2023).
40. Yu Shen (Ph.D.) — Committee member, CS (2023).
39. Yigitcan Kaya (Ph.D.) — Committee member, CS (2023).
38. Samuel William Atwell Dooley (Ph.D.) — Committee member, CS (2023).
37. Valeriia Cherepanova (Ph.D.) — Committee member, AMSC (2023).
36. Brandon Feng (Ph.D.) — Committee member, CS (2023).
35. Hao Chen (Ph.D.) — Committee member, CS (2023).
34. Yushan Feng (Ph.D.) — Committee member, CS (2023).
33. Avi Koplon Schwarzschild (Ph.D.) — Committee member, AMSC (2023).
32. Jordan Kirby Terry (Ph.D.) — Committee member, CS (2023).
31. Jun Wang (Ph.D.) — Committee member, ECE (2022–2023).
30. Justin Terry (Ph.D.) — Committee member, CS (2023).
29. Mohammad Amin Ghiasi (Ph.D.) — Committee member, CS (2022).
28. Varun Suryan (Ph.D.) — Committee member, CS (2022).
27. Xin Tian (Ph.D.) — Committee member, ECE (2022).
26. Chen Zhu (Ph.D.) — Committee member, CS (2022).
25. Shouvanik Chakrabarti (Ph.D.) — Committee member, CS (2022).
24. Zeyad Ali Sami Emam (Ph.D.) — Committee member, AMSC (2022).

23. Chihuang Liu (Ph.D.) — Committee member, ECE (2021–2022).
22. Gaurav Shrivastava (M.S.) — Committee member, CS (2021).
21. Parsa Saadatpanah (Ph.D.) — Committee member, CS (2021).
20. Yuqian Hu (Ph.D.) — Committee member, ECE (2021).
19. Wells Robinson (Ph.D.) — Committee member, CS (2021).
18. Xitong Yang (Ph.D.) — Committee member, CS (2021).
17. Mingliang Chen (Ph.D.) — Committee member, ECE (2021).
16. Fengyu Wang (Ph.D.) — Committee member, ECE (2021).
15. Sushant Patkar (Ph.D.) — Committee member, CS (2021).
14. Sai Deepika Regani (Ph.D.) — Committee member and Dean's representative, ECE (2020–2021).
13. Xiaoxu Meng (Ph.D.) — Committee member, CS (2020).
12. Ali Shafahi (Ph.D.) — Committee member, CS (2020).
11. Tongyang Li (Ph.D.) — Committee member, CS (2020).
10. Yuan Su (Ph.D.) — Committee member, CS (2020).
9. Yiqi Li (M.S.) — Committee member and Dean's representative, ECE (2019).
8. Xing Niu (Ph.D.) — Committee member, CS (2019).
7. Roozbeh Yousefzadeh (Ph.D.) — Committee member, CS (2019).
6. Thang Dai Nguyen (Ph.D.) — Committee member, CS (2019).
5. Karthik Abinav Sankararaman (Ph.D.) — Committee member, CS (2019).
4. Pan Xu (Ph.D.) — Committee member, CS (2019).
3. Zheng Xu (Ph.D.) — Committee member, CS (2019).
2. Ruofei Du (Ph.D.) — Committee member, CS (2018).
1. Mohammed E. Fathy Salem (Ph.D.) — Committee member, CS (2018).

SERVICE ACTIVITIES

Professional Service

7. Organizer of NeurIPS 2024 Competition: Erasing the Invisible: A Stress-Test Challenge for Image Watermarks, Sep. - Dec. 2024.
6. Chair of NSF-Amazon FAI (Fairness in AI) PI Meeting, Jan. 9-10, 2024, Arlington, VA.
5. Co-chair of NSF-IEEE workshop: Toward Explainable, Reliable, and Sustainable Machine Learning in Signal & Data Science, Mar. 20-21, 2023, College Park, MD.
4. Chair, IEEE Signal Processing Society (SPS) Washington Chapter, 2022–present.
3. Organizer: "Dagstuhl Seminar on Tensor Computations: Applications and Optimization" 2020-2021 and 2021-2022.
2. Organizer: "Matrix Factorization" Workshop at 5th Heidelberg Laureate Forum, Heidelberg, Germany, Sep. 2017.
1. Program Steering Committee, SysML 2018.

Non-University Panels

9. NSF panelist at Electrical, Communications, and Cyber Systems (ECCS) EPCN Unsolicited Panel 0515, May. 2023.
8. NSF panelist at CCF-CIF Small ML+Optim P in the Division of Computing and Communications Foundations, Apr. 2020.
7. Jury of the ACM India Doctoral Dissertation Award, Best thesis, 2020.
6. NSF panelists at RI CRII RH Panel in the Information & Intelligent Systems Division, Dec., 2019.
5. Jury of the ACM India Doctoral Dissertation Award, Best thesis, 2019.
4. NSF panelists at CISE/IIS ML/AI Medium Projects Panel, Feb., 2019.
3. Jury of the ACM India Doctoral Dissertation Award, Best thesis, 2018.
2. SysML Conference Panel, 2018.
1. US-Czech AI Workshop Panel, 2018.

University Service

29. Chair, Diversity and Inclusion Committee, 2025–2026.
28. AIM Tenure-Track Faculty Search Committee, 2025–2026.
27. UMIACS Director Search Committee, 2025–2026.
26. Merit Review Committee, 2024.
25. UMIACS APT Committee, 2023 - 2024.
24. Faculty Hiring Committee, 2023-2024.
23. University of Maryland AI Working Group Committee, University of Maryland, 2023.
22. Rising Star in ML Committee, 2023.
21. Department council, 2022-2023.
20. Graduate student fellowship committee, 2022-2023.
19. I4C Director Search Committee, 2022-2023.
18. Field Committee Chair - Artificial Intelligence, 2022-2023.
17. Rising Star in ML Committee, 2022.
16. Admission Committee, 2021-2022.
15. Field Committee Chair - Artificial Intelligence, 2021-2022.
14. Faculty Hiring Committee, 2020-2021.
13. Rising Star in ML Committee, 2021.
12. Field Committee Chair - Artificial Intelligence, 2020-2021.
11. STAT 400 CS Department Representative, 2020.
10. Rising Star in ML Committee, 2020.
9. Field Committee Chair - Artificial Intelligence, 2019-2020.
8. CS-ECE joint capstone project course design, 2019.
7. Rising Star in ML Committee, 2019.
6. AMSC SC track reform committee, 2019.
5. Academic Job Search Panel, 2019.
4. Field Committee Chair - Artificial Intelligence, 2018-2019.
3. Middlestates Assessment Committee, 2018-2019.
2. Faculty Hiring Committee, 2017-2018.
1. Committee for Google PhD Fellowship Department Nomination, Nov. 2017.

SYNERGISTIC ACTIVITIES

- **Trustworthy and agentic AI with real-world impact.** Research connects foundations of machine learning with AI safety, fairness, robustness, multimodal reasoning, and sequential decision-making. Partnerships with government, nonprofit, and industry sponsors translate these methods into deployable evaluations and systems.
- **Research communities and convening.** Organizer of the Matrix Factorization Workshop at the fifth Heidelberg Laureate Forum; Dagstuhl Seminars on Tensor Computations (2020, 2022); Rising Star in Machine Learning; and the University of Maryland Machine Learning Reading Group.
- **Scientific leadership.** Senior Area Chair for ICLR (2025, 2026; appointed for 2027) and ICML (2026); Area Chair for ICLR, ICML, NeurIPS, CoRL, and AAAI; Action Editor for TMLR; and Program Committee member for IJCAI-ECAI and SysML.
- **Broad research dissemination.** More than 80 invited, keynote, panel, and research presentations across academia, government, and industry, including the tutorial “The Role of Tensors in Deep Learning” at ACM SIGMETRICS 2018.
- **Open-source research.** Research code, datasets, benchmarks, and project pages are released through the Huang Group GitHub repository and the artifact links accompanying the publication record.