

Yiwei Yang

University of Washington

819 Northeast 67th st
Seattle, WA 98115

yanyiwei.github.io
yanyiwei@uw.edu

Education

- 10/20 – Present* **University of Washington**
Seattle, WA Ph.D. in Information Science
Research Interests: Reliability, Spurious Correlations, Reward
Hacking, Reasoning
Advisor: Bill Howe
- 09/15 – 05/19* **University of Michigan**
Ann Arbor, MI B.S. in Computer Science and Engineering

Professional Experience

- 10/20 – Present* **University of Washington**
Seattle, WA Graduate Student and Researcher
- 06/25 – 09/25* **Adobe**
San Jose, CA Applied Scientist Intern (Mentors: Yongsung Kim, Joshua Hailpern, Yunyao Li)
- 06/22 – 09/22* **SONY AI**
Remote Research Intern (Mentors: William Thong, Alice Xiang)

Research Projects

- UW Robust Reward Modeling Mitigates Reward Hacking**
Proposed a robust reward modeling approach that leverages counterfactual data augmentation to balance training distributions across spurious attributes.
- UW Benchmarking Spurious Correlations in Multi-modal LLMs**
Built a multi-modal benchmark revealing vulnerabilities to spurious correlations in even state-of-the-art proprietary models (e.g., GPT-4o). Showed that fine-tuning on a diverse set of spurious correlations enhances out-of-distribution robustness.
Accepted to ICML 2025 R2-FM,
In Submission to NeurIPS 2025 (Scores: 5 5 5 4)

UW Improving Robustness to Spurious Correlations with Concepts
Proposed a framework that infers group labels from out-of-distribution examples (generated with prompt-based Stable Diffusion), enabling robust training without manual annotations to improve worst-group accuracy.
Accepted to CVPR 2024

UW Zero-Shot Annotation of Urban Features with LMMs
Proposed a zero-shot method that leverages Large Multi-modal Modals (LMMs) along with visual prompts (i.e., Set-of-Marks) to annotate urban features such as stop lines and raised tables.
Accepted to SIGSPATIAL 2024

Publications

- P.14 **Y. Yang**, C. Lee, S. Feng, D. Zhao, B. Wen, A. Liu, Y. Tsvetkov, B. Howe. Escaping the SpuriVerse: Can Large Vision-Language Models Generalize Beyond Seen Spurious Correlations? *ICML R2-FM 2025*.
- P.13 B. Han, **Y. Yang**, A. Caspi, B. Howe. Towards Zero-shot Annotation of the Built Environment with Vision-Language Models. *SIGSPATIAL 2024*.
- P.12 **Y. Yang**., A. Liu, R. Wolfe, A. Caliskan, B. Howe. Label-Efficient Group Robustness via Out-of-Distribution Concept Curation. *CVPR 2024*.
- P.11 R. Wolfe, S. Issac, B. Han, B. Wen, **Y. Yang**, L. Rosenblatt, B. Herman, E. Brown, Z. Qu, N. Weber, B. Howe. Laboratory-scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. *FAccT 2024*.
- P.10 **Y. Yang**, B. Howe. Does a Fair Model Produce Fair Explanations? Relating Distributive and Procedural Fairness. *HICSS 2024*.
- P.09 **Y. Yang**., A. Liu, R. Wolfe, A. Caliskan, B. Howe. Regularizing Model Gradients with Concepts to Improve Robustness to Spurious Correlations. *ICML SCIS 2023*.
- P.08 R. Wolfe, **Y. Yang**, B. Howe, A. Caliskan. Contrastive Lanugage-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. *FAccT 2023*.
- P.07 **Y. Yang**., E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop. *ACL 2019*.
(**Best Poster** at Michigan AI Symposium, 1/55)
- P.06 **Y. Yang**., E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. A study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. *IUI 2019*.

- P.05 A. Lundgard, **Y. Yang**, M.L. Foster, W.S. Lasecki. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. *CHI 2018*.
- P.04 S.W. Lee, Y. Zhang, I. Wong, **Y. Yang**, S. D. O’Keefe, W.S. Lasecki. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping Of Interactive Interfaces. *UIST 2017*.
- P.03 H. Kaur, M. Gordon, **Y. Yang**, J. Teevan, E. Kamar, J. Bigham, W.S. Lasecki. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. *HCOMP 2017*.
- P.02 Y. Chen, S.W. Lee, Y. Xie, **Y. Yang**, W.S. Lasecki, S. Oney. Codeon: On Demand Software Development Assistance. *CHI 2017*.
- P.01 S. W. Lee, **Y. Yang**, S. Yan, Y. Zhang, I. Wong, Z. Yan, M. McGruder, C. M. Homan, W. S. Lasecki. Creating Interactive Behaviors in Early Sketch by Recording and Remixing Crowd Demonstrations. *HCOMP 2016*.