# Yiwei Yang

**University of Washington**

819 Northeast 67th st                              yanyiwei.github.io
Seattle, WA 98115                                  yanyiwei@uw.edu

## Education

| | |
|---|---|
| *10/20 — Present*<br>*Seattle, WA* | **University of Washington**<br>Ph.D. in Information Science<br>Research Interests: Large Multi-modal Models, Reliability,<br>    Spurious Correlations, Data-Centric ML<br>Advisor: Bill Howe |
| *09/15 – 05/19*<br>*Ann Arbor, MI* | **University of Michigan**<br>B.S. in Computer Science and Engineering |

## Professional Experience

| | |
|---|---|
| *10/20 — Present*<br>*Seattle, WA* | **University of Washington**<br>Graduate Student and Researcher |
| *06/22 – 09/22*<br>*Remote* | **SONY AI**<br>Research Intern (Mentors: William Thong, Alice Xiang) |
| *05/18 – 08/18*<br>*San Jose, CA* | **IBM Research, Almaden**<br>Research Intern (Mentors: Eser Kandogan, Prithviraj Sen, Yunyao Li) |

## Research Projects

*UW* **Benchmarking Spurious Correlations in Multi-modal LLMs**
Synthesizing a multimodal benchmark showing that spurious correlation is hard to even the very large proprietary models (e.g., GPT-4o). Proposed prompting and finetuning methods to improve model robustness to out-of-distribution spurious correlations.

*UW* **Improving Robustness to Spurious Correlations with Concepts**
Existing works on spurious correlations require group labels. Introduced a framework that first uses out-of-distribution examples (which can be created with one line of prompt to Stable Diffusion) to infer group labels, then applies robust training methods with the inferred group labels 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*

*SONY AI* **Bias Propagation in Knowledge Distillation for Vision Models**
Demonstrated that fairness properties transfer from teacher to student model via knowledge distillation.

## Publications

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*.