Yiwei Yang

Seattle, WA | yanyiwei.github.io | yanyiwei@uw.edu

Research Interests

Large Language Model Agents, Tool Use & Reasoning, Post-Training, Alignment, Robustness to Spurious Correlations.

Education

University of Washington Seattle, WA
Ph.D. in Information Science (Advisor: Bill Howe) 2020 – Present
Research focus: Evaluating and improving the robustness of LLM agents.

University of Michigan Ann Arbor, MI
B.S. in Computer Science and Engineering 2015 – 2019

Professional Experience

Adobe San Jose, CA
Machine Learning Engineer Intern Summer 2025

Sony AI Remote
Research Intern Summer 2022

Selected Research Projects

Robust Tool-Using LLM Agents under Spurious Correlations

- **Problem:** Investigated whether RL post-training (e.g., **GRPO**) causes agents to learn ”shortcuts,” such as triggering tool calls based on prompt formatting cues rather than logical necessity.
- **Method:** Developed a specialized RL framework incorporating a **Tool Necessity Reward** to penalize illogical tool calls and encourage principled reasoning trajectories. Analyzed token-level entropy to identify entropy collapse during tool-token generation.
- **Impact:** Improved end-to-end agent accuracy and correct tool call rates on out-of-distribution (OOD) tasks compared to standard RLVR baselines; currently preparing for submission to **COLM 2026**.

SpuriVerse: Mitigating Spurious Correlations in VLMs

- **Problem:** Investigated the systemic reliance of Vision-Language Models (VLMs) on spurious features—such as background cues and object co-occurrences—which leads to reasoning failures under distribution shift.
- **Method:** Developed a synthetic data pipeline to generate 2k+ counterfactual image pairs and showed that **post-training** can decorrelate target labels from non-causal visual features.
- **Impact:** Demonstrated that training on diverse spurious samples improves out-of-distribution (OOD) generalization by 43.2% on unseen data; published as first-author at **NeurIPS 2025**.

Label-Efficient Group Robustness via OOD Concepts

- **Problem:** Addressed the reliance on expensive, manual group annotations required for **Group Robustness** training (e.g., DRO), which typically limits the scalability of robust models.
- **Method:** Proposed a framework that infers latent group labels from **out-of-distribution (OOD) examples**, enabling robust optimization without requiring human-labeled group metadata.
- **Impact:** Achieved a 33.1% increase in worst-group accuracy on Waterbirds and CelebA benchmarks; published as first-author at **CVPR 2024**.

Selected Publications

- **Y. Yang**, C. Lee, S. Feng, et al. *Escaping the SpuriVerse: Can Large Vision-Language Models Generalize Beyond Seen Spurious Correlations?* **NeurIPS 2025**.
- **Y. Yang**, A. Liu, R. Wolfe, A. Caliskan, B. Howe. *Label-Efficient Group Robustness via Out-of-Distribution Concept Curation.* **CVPR 2024**.
- **Y. Yang**, B. Howe. *Does a Fair Model Produce Fair Explanations? Relating Distributive and Procedural Fairness.* **HICSS 2024**.
- R. Wolfe, ..., **Y. Yang**, et al. *Laboratory-scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings.* **FAccT 2024**.
- B. Han, **Y. Yang**, A. Caspi, B. Howe. *Towards Zero-shot Annotation of the Built Environment with Vision-Language Models.* **SIGSPATIAL 2024**.
- **Y. Yang**, A. Liu, R. Wolfe, A. Caliskan, B. Howe. *Regularizing Model Gradients with Concepts to Improve Robustness to Spurious Correlations.* **ICML SCIS 2023**.

Technical Skills

Frameworks: PyTorch, HuggingFace (Transformers, Accelerate), DeepSpeed, vLLM.
Algorithms: Reinforcement Learning (GRPO), RLVR, SFT, Reward Modeling.