Yi Ding

✉ ding432@purdue.edu
☎ (+1) 765-694-9620

Education

- **Purdue University** — **IN, USA**
  Ph.D. in Computer Science — *Aug. 2025 – 2030 (Expected)*
  - Advisor: **Dr. Ruqi Zhang**
- **Tianjin University** — **Tianjin, China**
  B.S. in Data Science and Big Data Technology, School of Mathematics — *Aug. 2021 – Jun. 2025*
  - Advisor: **Prof. Bing Cao** & **Prof. Qinghua Hu**

Research Interest

My primary research goal is to develop reliable and efficient machine learning models/algorithms to address real-world challenges. With this vision, my work focuses on steering Foundation Models (FMs), including LLMs, VLMs, and diffusion models toward human preference and improved reasoning. Currently, my research interests include:

- **LLM Post-Training**
  LLM/VLMs Alignment, RLHF, Reasoning, Self-Improving LLM/VLMs
- **Trustworthy AI**
  AI Safety, Fairness, Uncertainty, etc.
- **Multimodal Learning**
  Multimodal Fusion, Imbalanced Multimodal Learning

Publications

(* denotes equal contribution)

[P1] **SafeWork-R1: Coevolving Safety and Intelligence under the AI-45° Law**
Shanghai AI Lab, · · · , **Yi Ding**, [100+ Authors]
Technical Report

[C1] **Rethinking Bottlenecks in Safety Fine-Tuning of Vision Language Models**
Yi Ding*, Lijun Li*, Bing Cao, Jing Shao
International Conference on Learning Representations (**ICLR 2026**)

[C2] **Sherlock: Self-Correcting Reasoning in Vision-Language Models**
Yi Ding, Ruqi Zhang
Neural Information Processing Systems (**NeurIPS 2025**)

[C3] **Visual Contextual Attack: Jailbreaking MLLMs with Image-Driven Context Injection**
Ziqi Miao*, **Yi Ding***, Lijun Li, Jing Shao
Empirical Methods in Natural Language Processing Main Conference (**EMNLP 2025**)

[C4] **ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time**
Yi Ding, Bolian li, Ruqi Zhang
International Conference on Learning Representations (**ICLR 2025**)

[C5] **Test-Time Dynamic Image Fusion**
Bing Cao (Advisor), Yinan Xia*, **Yi Ding***, Changqing Zhang, Qinghua Hu
Neural Information Processing Systems (**NeurIPS 2024**)

[C6] **Predictive Dynamic Fusion**
Bing Cao (Advisor), Yinan Xia*, **Yi Ding***, Changqing Zhang, Qinghua Hu
International Conference on Machine Learning (**ICML 2024**)

Awards

- **NSF ACCESS Discover Project Award** **2025-2026**
  LLM Post-Training

Research Experience

- **RZ-Lab, Purdue University** **IN, USA**
  Research Intern, Advised by **Dr. Ruqi Zhang** *May 2024–May 2025*
- **Open Trust Lab, Shanghai AI Laboratory** **Beijing, China**
  Research Intern, Advised by **Dr. Lijun Li** *Dec. 2024–Mar. 2025*
- **MLDM Lab, Tianjin University** **Tianjin, China**
  Research Intern, Advised by **Prof. Bing Cao** and **Prof. Qinghua Hu** *Sep. 2023–Dec. 2024*

Skill

Languages: Chinese Mandarin (Native), English (TOEFL 102(22))

Research Abilities: Proficient in coding: Python, LaTeX, MATLAB; Enjoys mathematical derivations; Solid foundation in mathematics and statistics.

Service

- **Conference Reviewer**
  ICML 2026; NeurIPS 2025; ICLR 2025,2026; ARR 2025 May