Yi Ding

ding432@purdue.edu
(+86) 135 2267 3805

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 – Jul. 2025* (*Expected*)
  - 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] **Sherlock: Self-Correcting Reasoning in Vision-Language Models**
Yi Ding, Ruqi Zhang
Under Review (**Preprint**)

[P2] **Rethinking Bottlenecks in Safety Fine-Tuning of Vision Language Models**
Yi Ding*, Lijun Li*, Bing Cao, Jing Shao
Under Review (**Preprint**)

[C1] **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**)

[C2] **Test-Time Dynamic Image Fusion**
Bing Cao (Advisor), Yinan Xia*, **Yi Ding***, Changqing Zhang, Qinghua Hu
Neural Information Processing Systems (**NeurIPS 2024**)

[C3] **Predictive Dynamic Fusion**
Bing Cao (Advisor), Yinan Xia*, **Yi Ding***, Changqing Zhang, Qinghua Hu
International Conference on Machine Learning (**ICML 2024**)

Research Experience

- **RZ-Lab, Purdue University** — **IN, USA**
  Research Intern, Advised by **Dr. Ruqi Zhang** — *May 2024–May 2025*
  - Introduced Sherlock, a self-correction and self-improvement training framework tailored to enhance reasoning ability for Vision-Language Models. Sherlock uses only 20k randomly sampled annotated data to unlock reasoning and self-correction capabilities of the base VLM, then leveraging visual perturbation, and the inherent quality gap between responses before and after self-correction to self-construct a preference dataset to conduct trajectory-level self-improvement training without any annotation. (*In submission*)

- Proposed a two-phase plug-and-play alignment framework (ETA) featuring a multimodal evaluator and bi-level alignment, provided a new perspective of safety challenges in vision-language models caused by continuous visual token embeddings. ETA ensures responses are both safe and useful, improving harmlessness and helpfulness without additional training or data while maintaining the VLM's general performance. (*Accepted at ICLR 2025*)

Open Trust Lab, Shanghai AI Laboratory **Beijing, China**
Research Intern, Advised by **Dr. Lijun Li** *Dec. 2024–Mar. 2025*

- Proposed a new safety fine-tuning paradigm by constructing multi-image unsafe scenarios, creating SFT data with visual reasoning, while improving the model's helpfulness and harmlessness. Additionally, I introduced a large-scale multi-image safety benchmark, which demonstrates how current models struggle to capture unsafe intentions formed by associations between images, leading to successful jailbreaks. (*In Submission*)

MLDM Lab, Tianjin University **Tianjin, China**
Research Intern, Advised by **Prof. Bing Cao** and **Prof. Qinghua Hu** *Sep. 2023–Dec. 2024*

- Revealed that the key to dynamic fusion lies in the correlation between the weights and the loss, providing theoretical foundations for multimodal decision-level fusion from the perspective of the generalization error bound. Based on this insight, proposed the *Predictive Dynamic Fusion* (PDF) algorithmic framework offers trustworthy priors for decision-level fusion in multimodal systems or multi-agent settings, thereby achieving better generalization. (*Accepted at ICML 2024*)
- Proved theoretically that dynamic image fusion outperforms static image fusion and introduced Relative Dominability, providing a formal framework to enhance the interpretability of complex network architectures. This theoretical proof supports dynamically fusing advantageous regions and adjusting fusion weights at test time, leading to significant improvements in image quality across various baselines. (*Accepted at NeurIPS 2024*)

UNITES Lab, UNC Chapel Hill **North Carolina, USA**
Research Intern, Advised by **Dr. Tianlong Chen** *Feb. 2024-May 2024*

- Developed a neural network ensemble framework for time series stock prediction, integrating multiple models (*1D-CNN, GRU, etc.*) to improve accuracy. Additionally, proposed an efficient fine-tuning system for time series foundation models (EFT-TSFM) to enhance parameter and memory efficiency. Findings were summarized in the NN in Finance technical report.

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
- NeurIPS 2025, ICLR 2025