# Jialiang Fan

jfan5@nd.edu • jialiangfan.com • Scholar | GitHub

I am a second-year doctoral student in Computer Science and Engineering at the University of Notre Dame, advised by Professor Fanxin Kong. My research focuses on safe and trustworthy AI, with expertise in LLM post-training (SFT, RLHF, GRPO), agentic AI, Vision-Language-Action (VLA) models, world models and world-action models, reinforcement learning, and task planning with formal methods.

## Education

- **University of Notre Dame** — **South Bend, IN, USA**
  *Ph.D. in Computer Science and Engineering, GPA: 3.917/4.0* — *Jun 2024 – Present*
- **Lanzhou University** — **Lanzhou, P.R. China**
  *Master of Electronic Information, Computer Technology* — *2020.9–2023.6*
- **Shandong University** — **Jinan, P.R. China**
  *Bachelor of Engineering, Software Engineering* — *2015.9–2019.6*

## PhD Publications

- Z. Wang, **J. Fan**, R. Zuo, Q. Qiu, and F. Kong, "SafeNet: A Neural-Symbolic Network for Safe Planning in Robotic Systems using Formal Method-Guided LLM Fine-Tuning," in ***Proc. IEEE Int. Conf. Robot. Autom. (ICRA)***, 2026.
- **J. Fan**, M. Liu, S. Jiang, and F. Kong, "Vulnerability Exploration of Safe Reinforcement Learning on Cyber-Physical Systems via STL Mining," in ***Proc. ACM/IEEE Int. Conf. Cyber-Phys. Syst. (ICCPS)***, 2026. *(Top venue in CPS)*
- **J. Fan** and F. Kong, "A Survey of Signal Temporal Logic Specification Mining: Techniques, Applications, and Future Directions," in ***Proceedings of SPIE Defense + Commercial Sensing 2025***, Orlando, FL, 2025.

### Preprints

- **J. Fan**, W. Xu, M. Liu, O. Sokolsky, I. Lee, and F. Kong, "SafeGen-LLM: Enhancing Safety Generalization in Task Planning for Robotic Systems," ***arXiv preprint arXiv:2602.24235***, 2026. *(Submitted to IROS 2026)*
- **J. Fan**, S. Jiang, M. Liu, and F. Kong, "Vulnerability Analysis of Safe Reinforcement Learning via Inverse Constrained Reinforcement Learning," ***arXiv preprint arXiv:2602.16543***, 2026. *(Submitted to IROS 2026)*

## Skills

- **Programming:** Python, C/C++, MATLAB
- **ML / LLM Frameworks:** PyTorch, HuggingFace Transformers, TRL, PEFT/LoRA, vLLM, DeepSpeed, Accelerate
- **LLM Post-Training:** SFT, RLHF, DPO, GRPO, RLVR, reward modeling
- **Reinforcement Learning:** PPO, SAC, constrained RL, inverse RL, safe RL
- **Robotics & VLA:** OpenVLA, LIBERO, ROS, MuJoCo, Safety-Gymnasium, manipulation policy learning
- **Formal Methods:** Signal Temporal Logic (STL), reward machines, specification mining
- **Infra & Tools:** CUDA, Linux, Git, distributed training (FSDP / DeepSpeed), SLURM

## Research/Project Experience

- **Research at University of Notre Dame**
  ***SafeVLA: Formal Safety Alignment for Vision-Language-Action Models*** *2025 – Present*

  VLA models maximize task success but ignore *how* tasks are executed, learning policies that collide, exceed workspace bounds, or move unsafely near humans.
  - Propose the first formal specification-guided VLA post-training framework, using Signal Temporal Logic (STL) robustness as a continuous, step-level safety reward inside GRPO/RLVR fine-tuning of OpenVLA on LIBERO. *Targeting CoRL/NeurIPS 2026.*

- **Collaboration between University of Notre Dame and University of Pennsylvania**
  ***SafeGen-LLM: Safety Generalization for Robot Task Planning*** *2025 – Present*
  - Design LLM-based robot task planners with formal verification rewards to improve safety generalization.
  - Fine-tune LLMs via supervised learning and GRPO with reward machines for safety-constrained robot planning.

- **Collaboration between University of Notre Dame and Syracuse University**
  ***SafeNet: Neural-Symbolic Network for Safe Planning in Robotic Systems*** *2024 – 2025*
  - Build neural-symbolic architectures integrating formal logic into LLM post-training for safe trajectory generation.
  - Combine Signal Temporal Logic specifications with LLM fine-tuning for safety-guaranteed robot planning. *Accepted by IEEE ICRA 2026.*

- **Research at University of Notre Dame** **South Bend, Indiana, USA**
  ***Safe and Robust Learning for Robotic and Cyber-Physical Systems*** *May 2024 – Present*
  - Develop adversarial attack methods to analyze vulnerabilities of safe reinforcement learning controllers in CPS.
  - Mine Signal Temporal Logic (STL) specifications from trajectories to guide adversarial perturbation generation.

## Work Experience

- **Tencent Technology (Shenzhen) Co., Ltd.** **Shenzhen, P.R. China**
  ***Research Intern, Tencent Robotics X*** *Dec 2021 – Aug 2022*
  - Developed learning-based policies for dexterous manipulation tasks on a robotic arm platform.
  - Built a ROS-based whole-body control stack integrating a custom mobile base with a Kinova Jaco manipulator.