

# Xijia “Polina” Zhang

✉ [zhang.xijia@gatech.edu](mailto:zhang.xijia@gatech.edu) 🎓 [Zhang Xi-Jia](#) 🌐 [xijia.me](https://xijia.me) in [Xijia Polina Zhang](#)

## Research Interests

I am deeply passionate about **mechanistic interpretability**, especially structures like transformer circuits and sparse autoencoders. I’m also interested in reverse-engineering how LLMs reason: how they acquired these capabilities, how we can steer them in latent space, and why such interventions work. I believe mechanistic interpretability will be critical to LLM safety and alignment in the near future.

I am currently exploring how to probe latent Chain-of-Thought (CoT) reasoning to understand why it works, whether it can be aligned with world concepts, and whether such alignment could improve the performance of foundational models.

## Education

**Georgia Institute of Technology**

Aug. 2024 – Dec. 2029

*Ph.D. in Robotics*

*\*Willing to drop out for a full-time research position at Anthropic.*

**University of Michigan**

Sep. 2022 – May. 2024

*B.S.E in Computer Science*

**Shanghai Jiao Tong University**

Sep. 2020 – Aug. 2024

*B.E in Electrical and Computer Engineering*

## Publications

[“Model-Agnostic Policy Explanations with Large Language Models”](#)

**Zhang Xi-Jia**, Yue Guo, Shufei Chen, Simon Stepputtis, Matthew Gombolay, Katia Sycara, Joseph Campbell  
COLM 2025 (**top 5% of submissions**)

- Propose a method for generating natural language explanations of agent behavior based only on observed states and actions, without access to the agent’s underlying model.

[“Towards Human-Free Semantic Interpretability in Reinforcement Learning via Vision-Language Models”](#) (in submission)

Zhaoxin Li\*, **Zhang Xi-Jia**\*, Batuhan Altundas, Letian Chen, Rohan Paleja, Matthew Gombolay

- Develop an automated RL framework that leverages foundational models for semantic feature extraction and interpretable tree-based models for policy optimization.

[“Learning Effective Action Advising in the Face of Changing Rewards”](#)

Yue Guo, **Zhang Xi-Jia**, Simon Stepputtis, Joseph Campbell, Katia Sycara  
CoLLAs 2024 (**Oral**)

- Enable the teacher policy to continually learn and adapt its reward function through ongoing observation of the student when providing action advices.

[“Sensor Array Optimization for the Electronic Nose via Different Deep Learning Methods”](#)

**Zhang Xi-Jia**\*, Tao Wang\*, Wangze Ni, Yongwei Zhang, Wen Lv, Min Zeng, Jianhua Yang, Nantao Hu, Rui Zhan, Guang Li, Zhiqiang Hong, Zhi Yang  
Sensors and Actuators: B, 2024

- Compare lightweight machine learning models on an Electronic Nose and investigate how their performance scales with dataset size.

“Teaching the Teacher: Enhancing Human-to-Robot Skill Demonstration with Live Foundation Model and Augmented Reality Feedback”

Nina Moorman, Matthew Luebbers, **Zhang Xi-Jia**, Zulfiqar Zaidi, Marcus Lau, Yixing Yao, Megan Langwasser, Letian Chen, Sanne van Waveren, Matthew Gombolay (in submission)

- Support non-expert users in providing kinesthetic demonstrations through language explanations and policy visualizations.
- Contributions: Language explanations.

“Communication and Verification in LLM Agents towards Collaboration under Information Asymmetry

Run Peng, Ziqiao Ma, Amy Pang, Sikai Li, **Zhang Xi-Jia**, Yingzhuo Yu, Cristian-Paul Bara, Joyce Chai (in submission)

- Investigate the performance of LLM agents in human-robot collaboration tasks under information asymmetry.
- Contributions: TIAGo robot development.

Skills

Toolchain	ROS, ROS2, Linux, Docker, Huggingface, RLlib
Frameworks & Libraries	Gym, NLTK, Transformers, Pytorch, Scikit-Learn
Foundation Model	LLM fine-tuning, prompting; LoRA, PPO, GRPO
Robotics Development	TIAGo, Jaco Arm, Khepera, Fetch; RViz, Gazebo

Honors & Awards

Sep. 2024	Robotics Fellowship	Georgia Institute of Technology
Mar. 2024	James B. Angell Scholar	University of Michigan
Apr. 2024	Elected Member	Tau Beta Pi, Michigan Gamma
Dec. 2023	Dean’s List	University of Michigan
Apr. 2023	Dean’s List	University of Michigan
Dec. 2022	Dean’s List	University of Michigan
May. 2023	Chun-Tsung Scholar	Shanghai Jiao Tong University
Oct. 2021	Rongchang Innovation Scholarship Nomination	Shanghai Jiao Tong University
Nov. 2021	Silver Medal	The University Physics Contest