

Xijia “Polina” Zhang

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

Research Interests

I am deeply passionate about **mechanistic interpretability**. I aim to reverse-engineer foundation models to be steerable, grounded, and transparent, and to bridge their latent representations with human-level semantics through language.

Education

Georgia Institute of Technology

Aug. 2024 – Dec. 2029

Ph.D. in Robotics

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%**)

- 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.

[“Benchmarking Adaptation Efficiency in Diffusion-Based Vision-Language Action Models”](#)

Zhang Xi-Jia, Roman Yakunin, Che Chen, Shivang Chopra, Chengyue Huang, Xiujin Liu, Joseph Campbell, Zsolt Kira
(in submission)

- Propose and empirically evaluate a set of metrics for mechanistic analysis of Vision Language Action models, and use them to compare adaptation behavior across different model architectures.

[“Thinking Deeper Without Growing Larger: Scaling Virtual Depth in Pretrained LLMs for Enhanced](#)

[Reasoning”](#)

(in submission)

Shivang Chopra, **Zhang Xi-Jia**, Chengyue Huang, Karmesh Yadav, Yusuf Ali, Zsolt Kira

- Introduce dynamic layer routing via lightweight adapters that learn to adaptively skip or loop transformer blocks. Analyze activations, entropy, and model responses for mechanistic insights.

[“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 HRI 2026

- 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	Fine-tuning, Prompting, Deploying; LoRA, PPO, GRPO
Robotics Development	TIAGo, Jaco Arm, Khepera, Fetch; RViz, Gazebo

Honors & Awards

Sep. 2024	Robotics Fellowship	<i>Georgia Institute of Technology</i>
Mar. 2024	James B. Angell Scholar	<i>University of Michigan</i>
Apr. 2024	Elected Member	<i>Tau Beta Pi, Michigan Gamma</i>
2022-2023	Dean's List	<i>University of Michigan</i>
May. 2023	Chun-Tsung Scholar 150/13,000	<i>Shanghai Jiao Tong University</i>
Oct. 2021	Rongchang Scholarship Nomination 6/786	<i>Shanghai Jiao Tong University</i>
Nov. 2021	Silver Medal	<i>The University Physics Contest</i>