Zizhao **Wang**

zizhao.wang@utexas.edu · https://wangzizhao.github.io/ · google scholar: https://tinyurl.com/zizhaowangscholar

Research focus

- LLM post-training
- Reinforcement Learning (RL)
- Causal Reasoning

Skills

- LLM
  - LLM Agents
  - RL post-training
  - Reasoning
  - Safety
- decision making
  - model-based RL
  - RLHF
  - offline RL
  - hierarchical RL
  - imitation learning
  - planning
- representation learning
- generalization and robustness
- interpretability and explainability
- ML frameworks (PyTorch, TensorFlow, vLLM, Transformers)
- distributed training (deepspeed, PEFT)
- Python
- data structure, algorithms

Education

2020 - 25	**PhD**, Electrical and Computer Engineering expected graduation: Dec, advisor: **Peter Stone**	**University of Texas at Austin**
2018 - 19	**MS**, Computer Science	**Columbia University**
2016 - 18	**BS**, Computer Engineering (dual degree program)	**University of Michigan**
2014 - 18	**BS**, Electrical and Computer Engineering	**Shanghai Jiao Tong University**

Work Experience

2025/03 — Student Researcher — Google

- Designed an adversarial **RL post-training** framework that enhance **LLM agent** security again prompt injections, by co-training two LLMs: an attacker that learns to create diverse prompt injections and an agent that learns to defend against them.
- Implemented the data collection pipeline with vLLM and parallel simulation environments for fast LLM agent rollout inference.
- Fine-tuned the LLM model with the GRPO algorithm, implemented with Transformer, deepspeed, and LoRA for fast and memory-efficient training.
- Reduced the attack success rate by 21% and improve task success rate by 18% compared to the untrained model.

2024/06 — Research Intern — Microsoft Research

- Designed an **generative world model** that can synthesize images of novel scenarios, by using object-centric representations and disentangled representations.
- Enhanced the generalization of **reinforcement learning** policies by 30%, when learning with generated out-of-distribution data.

2024/01 — Research Intern — Honda Research Institute

- Developed a **motion prediction** algorithm for **autonomous driving** that, reduced prediction error by 48%, by applying **causal reasoning** to vehicle interactions.
- Sped up model training with **distributed training** and **efficient CUDA implementations** for sparse attention.

Research Experience

2021-22 — Causal World Model (ICML oral, AAAI oral) — University of Texas at Austin

- Developed a **world model** that can analyzes **causal relationships** between state factors (e.g., whether an object moves because of itself or other objects).
- Increased the model's **out-of-distribution generalization** by 46%, by leveraging the identified relationships and conditioning predictions only on relevant inputs.
- Derived a theoretically-grounded state abstraction for model-based RL, which improved sample efficiency and generalization in **planning** for **robotics** tasks.

2022-23 — Unsupervised Skill Learning (NeurIPS) — University of Texas at Austin

- Proposed a skill discovery method for **structured decision-making** tasks, where reusable skills are learned to induce interactions between state factors.
- Implemented a novel **hierarchical RL** algorithm for skill learning in PyTorch – the high-level policy selects the interaction to induce and the low-level policy learns to induce it using primitive actions.
- Enhanced skill diversity and downstream task performance on long-horizon robotics tasks and structured decision-making tasks by 40%.

Selected Publications

See google scholar (https://tinyurl.com/zizhaowangscholar) for a complete list of publications.

1. Adversarial Reinforcement Learning for LLM Agent Safety, *In submission*
   Z Wang, D Li, V Keshava, P Wallis, A Balashankar, P Stone, L Rutishauser.
2. SkiLD: Unsupervised Skill Discovery Guided by Local Dependencies, *NeurIPS 2024*
   Z Wang*, J Hu*, C Chuck*, S Chen, R Martín-Martín, A Zhang, S Niekum, P Stone.
3. Building Minimal and Reusable Causal State Abstractions for RL, *AAAI 2024* (**oral**)
   Z Wang*, C Wang, X Xiao, Y Zhu, and P Stone.
4. ELDEN: Exploration via Local Dependencies, *NeurIPS 2023*
   Z Wang*, J Hu*, R Martín-Martín, and P Stone.
5. Causal Dynamics Learning for Task-Independent State Abstraction, *ICML 2022* (**oral**)
   Z Wang, X Xiao, Z Xu, Y Zhu, and P Stone.