

Research Interests

My research sits at the intersection of reinforcement learning, large language models, and robot learning. Below is a more detailed account of the specific directions I am pursuing and the questions that drive my work.

RL Post-Training for Language Models

A central theme of my research is understanding how reinforcement learning can be used *after* supervised pre-training to unlock new capabilities in large language models. Recent work has shown that RL-based post-training—such as RLHF [1] and RLVR [2]—can substantially improve reasoning, instruction following, and alignment. I am particularly interested in:

- **Self-distillation and self-play:** can models improve by training on their own outputs, bootstrapping capability without new human data? This connects to ideas from expert iteration [3] and self-play in games [4].
- **Reward modeling and reward hacking:** how do we design reward signals that remain faithful proxies for human intent as models become more capable? The problem of reward hacking [5] is a key bottleneck.
- **Continual RL post-training:** rather than a single RL fine-tuning phase, can we develop curricula that iteratively refine models over time, akin to continual learning [6]?

Robot Learning and Visuomotor Policies

The second pillar of my work concerns learning sensorimotor policies for robotic manipulation. Foundation models pre-trained on internet-scale data provide rich representations, but bridging the gap to physical action remains an open challenge. I am exploring:

- **Vision-language-action models (VLAs):** architectures that map visual observations and language instructions directly to robot actions [7], and how to fine-tune them efficiently for new tasks.

- **On-policy RL for robotics:** methods like PPO [8] and GRPO [2] applied in simulation and on real hardware, with a focus on sample efficiency and sim-to-real transfer.
- **World models and planning:** learning predictive models of environment dynamics [9] that enable model-based planning and reduce the need for costly real-world interaction.

Unifying Theme

RL is the connective tissue across these interests. Whether the “backbone” is a language model or a visuomotor policy, the core algorithmic questions are shared: how to define good reward signals, how to explore efficiently, and how to leverage pre-trained representations for downstream adaptation. I believe progress on these fronts will be mutually reinforcing—advances in RL post-training for LLMs will inform robot learning, and vice versa.

References

- [1] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.
- [2] Z. Shao *et al.*, “DeepSeekMath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.
- [3] T. Anthony, Z. Tian, and D. Barber, “Thinking fast and slow with deep learning and tree search,” in *Advances in neural information processing systems*, 2017.
- [4] D. Silver *et al.*, “Mastering the game of Go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.
- [5] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, “Defining and characterizing reward hacking,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 20460–20471, 2022.
- [6] J. Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [7] A. Brohan *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv preprint arXiv:2307.15818*, 2023.
- [8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [9] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse domains through world models,” *arXiv preprint arXiv:2301.04104*, 2023.