Integrating Advanced Perception Models with Reinforcement Learning for Enhanced Object Interaction

Ahmad Farooq and Kamran Iqbal

afarooq@ualr.edu, kxiqbal@ualr.edu

Department of Electrical and Computer Engineering , University of Arkansas at Little Rock

(a) FloorPlan1 (b) FloorPlan2 (c) FloorPlan3 (d) FloorPlan4

Fig. 1: Top-down views of the four kitchen environments in AI2-THOR used in our experiments [6].

Abstract

This paper presents a novel approach integrating vision foundation models with reinforcement learning to enhance object interaction capabilities in simulated environments. We combine the Segment Anything Model (SAM) and YOLOv5 with a Proximal Policy Optimization (PPO) agent operating in the AI2-THOR simulation environment. Our experiments, conducted across four indoor kitchen settings, demonstrate significant improvements in object interaction and navigation efficiency compared to a baseline agent without advanced perception. The results show a 68% increase in average reward, a 52.5% improvement in object interaction success, and a 33% increase in navigation efficiency, highlighting the potential of combining foundation models with reinforcement learning for complex robotic tasks.

Index Terms: Reinforcement Learning, Object Interaction, Vision Foundation Models, Segment Anything Model, AI2-THOR Simulation

Introduction

Autonomous robots require sophisticated perception and decision-making capabilities to operate effectively in complex environments. Advanced perception allows robots to interpret their surroundings, recognize objects, and understand spatial relationships [1], [2].

Reinforcement learning (RL) offers a framework for training agents to learn optimal policies through interactions with the environment [3]. However, integrating advanced perception models with RL agents presents challenges, including computational complexity and the need for efficient real-time processing.

In this work, we propose an approach that combines vision foundation models with reinforcement learning to enhance object interaction in simulated environments. Specifically, we integrate the Segment Anything Model (SAM) [4] and YOLOv5 [5] into the perception pipeline of an RL agent operating within the AI2-THOR simulation environment [6].

Our contributions include:
(1) developing an RL agent that integrates SAM and YOLOv5 for improved perception and object interaction,
(2) addressing challenges in designing an effective reward function,
(3) demonstrating significant performance improvements over a baseline agent, and
(4) providing insights into the integration process of foundation models with RL for robotic applications.

TABLE I: Performance Comparison

Metric	Perception-Enhanced	Baseline
Avg. Reward	136.41 ± 70.56	81.15 ± 80.37
Obj. Interaction Success	73.5%	48.2%
Navigation Efficiency	82.1%	61.7%

Methodology

Environment and RL Framework:

- AI2-THOR simulation: Four indoor kitchen scenes
- Discrete action space: Navigation and object interaction
- Proximal Policy Optimization (PPO) for training
- Reward function: $r_t = \alpha \cdot \Delta dt + \beta \cdot st - \gamma \cdot ct$ Where Δdt: change in distance to target, st: success indicator, ct: penalty term

Perception Pipeline:

- SAM: Scene segmentation for object boundaries and spatial relationships
- YOLOv5: Real-time object detection and classification
- Combination: Overlay SAM masks on YOLOv5 bounding boxes
- Rich environmental representation for informed decision-making

Training and Evaluation:

- 440,000 timesteps across four environments
- Asynchronous data loading and GPU acceleration
- Checkpointing and periodic evaluation
- Comparison with baseline agent using standard RGB observations

Results

Key Improvements:

- 68% increase in average reward
- 52.5% improvement in object interaction success
- 33% increase in navigation efficiency

Qualitative Observations:

1. Superior object recognition in cluttered scenes
2. Improved spatial awareness and efficient navigation
3. More informed and coherent task planning
4. Quick adaptation to novel object arrangements
5. Efficient path planning and obstacle avoidance

Discussion

- Vision foundation models significantly enhance RL performance
- Trade-off: Improved capabilities vs. computational overhead
- Enhanced perception leads to better environment understanding and decision-making
- Challenges remain in computational efficiency and deployment
- Critical importance of reward function design
- Complementary benefits of SAM (detailed segmentation) and YOLOv5 (efficient detection)

Conclusions

- Successfully integrated vision foundation models with RL for enhanced object interaction
- Demonstrated significant improvements in reward, interaction success, and navigation efficiency
- Promising direction for more capable and adaptable robotic systems

Future Work

- Efficient integration techniques
- Transfer learning to real-world scenarios
- Multi-task learning capabilities
- Performance in dynamic environments
- Incorporating human-robot interaction

References

[1] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision,
2017, pp. 2961–2969.
[2] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in
Proceedings of the IEEE conference on computer vision and pattern
recognition, 2017, pp. 7263–7271.
[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Prox-
imal policy optimization algorithms,” arXiv preprint arXiv:1707.06347,
2017.
[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson,
T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., “Segment anything,”
in Proceedings of the IEEE/CVF International Conference on Computer
Vision, 2023, pp. 4015–4026.
[5] G. Jocher, “YOLOv5 by Ultralytics,” May 2020. [Online]. Available:
https://github.com/ultralytics/yolov5
[6] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti,
M. Deitke, K. Ehsani, D. Gordon, Y. Zhu et al., “Ai2-thor: An interactive
3d environment for visual ai,” arXiv preprint arXiv:1712.05474, 2017.

UA LITTLE ROCK

IROS '24 ABU DHABI