Integrating Vision Foundation Models with Reinforcement Learning for Enhanced Object Interaction

Ahmad Farooq*[1] and Kamran Iqbal[2]

***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**

I. 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.

II. Methodology

A. Environment and RL Framework

We use the AI2-THOR simulation environment [6], focusing on four indoor kitchen scenes (FloorPlan1, FloorPlan2, FloorPlan3, and FloorPlan4). The agent operates with a discrete action space comprising navigation and interaction actions, including moving forward, rotating left and right, looking up and down, picking up objects, and dropping them. The observation space consists of RGB image frames captured from the agent's point of view.

We employ the Proximal Policy Optimization (PPO) algorithm for training due to its stability and efficiency in high-dimensional action spaces. The design of the reward function is critical; we aim to encourage exploration, object interaction, and goal-oriented behaviors. The reward at time t is defined as:

$$r_t = \alpha \cdot \Delta d_t + \beta \cdot s_t - \gamma \cdot c_t, \tag{1}$$

where Δd_t is the change in distance to the target object, s_t is a success indicator for interactions, c_t is a penalty term, and α, β, γ are weighting factors.

Figure 1 shows top-down views of the four kitchen environments (FloorPlan1-4) used in our experiments. These diverse layouts provide a range of object configurations and spatial relationships for the agent to learn from, enhancing the robustness of our approach.

B. Perception Pipeline

We integrate SAM for scene segmentation and YOLOv5 for object detection. SAM generates accurate masks for objects in an image without task-specific training, providing the agent with detailed information about object boundaries and spatial relationships. YOLOv5 provides bounding boxes and class labels for objects in real-time, chosen for its stability and lower computational requirements.

The outputs from SAM and YOLOv5 are combined to create a rich representation of the environment. Specifically, we overlay the segmentation masks from SAM onto the detected bounding boxes from YOLOv5 to refine object localization. This information is processed and encoded into a feature representation fed into the agent's policy network, allowing it to make informed decisions based on the perceived scene.

C. Training and Evaluation

The agent is trained for 440,000 timesteps across the four environments. To manage computational demands, we employ techniques such as asynchronous data loading and GPU acceleration. We use checkpointing and evaluation mechanisms to monitor the agent's progress and save models periodically, enabling resumption of training and analysis of intermediate results.

To evaluate the effectiveness of our approach, we compare the performance of the agent with the advanced perception pipeline against a baseline agent that uses standard RGB observations without the integration of SAM and YOLOv5.

This work was not supported by any organization.

*Corresponding Author: Ahmad Farooq (afarooq@ualr.edu)

(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].

Both agents are trained under the same conditions, and their performance is assessed based on success rates in object interaction tasks and cumulative rewards.

III. Results and Discussion

Our results demonstrate significant improvements in the perception-enhanced agent's performance. Table I presents the quantitative results:

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%

The perception-enhanced agent shows a 68% higher average reward, a 52.5% improvement in object interaction success rate, and a 33% increase in navigation efficiency. Qualitatively, we observed several key differences:

1) Object Recognition: The perception-enhanced agent demonstrated superior ability in identifying and distinguishing between various kitchen objects, even in cluttered scenes.
2) Spatial Awareness: Leveraging the detailed segmentation from SAM, the perception-enhanced agent showed improved understanding of spatial relationships between objects, leading to more efficient navigation.
3) Task Planning: The enhanced perception allowed the agent to make more informed decisions about which objects to interact with and in what order, resulting in more coherent and goal-directed behavior.
4) Adaptability: When faced with novel object arrangements across the four different kitchen environments, the perception-enhanced agent adapted more quickly, leveraging its improved scene understanding to generalize its learned behaviors.
5) Efficient Path Planning: The perception-enhanced agent demonstrated more efficient path planning and navigation, often taking shorter routes to target objects and avoiding obstacles more effectively.

While the integration of vision foundation models significantly improves performance, it introduces computational overhead. The perception-enhanced agent requires more GPU memory and has longer inference times, presenting a trade-off between performance and computational cost.

The enhanced perception allows the agent to better understand the environment, leading to more efficient decision-making and demonstrating the value of integrating advanced perception models in reinforcement learning for robotic tasks.

However, challenges remain in computational efficiency and reward function design. The training time for the perception-enhanced agent was longer than for the baseline agent, which may limit deployment in resource-constrained environments or real-time applications with strict latency requirements.

Fine-tuning the reward function proved critical; our iterative approach, balancing immediate rewards for successful interactions with longer-term rewards for task completion, was crucial in achieving the reported performance gains.

The integration of SAM and YOLOv5 provided complementary benefits. While YOLOv5 offered efficient object detection and classification, SAM's detailed segmentation masks enabled finer-grained spatial reasoning. This combination allowed the agent to not only identify objects but also understand their shape, boundaries, and spatial relationships more accurately across the various kitchen layouts.

IV. Conclusion and Future Work

We presented an approach integrating vision foundation models with reinforcement learning, demonstrating significant performance improvements in object interaction tasks. Future work will focus on efficient integration techniques, transfer learning to real-world scenarios, multi-task learning, performance in dynamic environments, and incorporating human-robot interaction. Despite challenges in computational efficiency, our work highlights a promising direction for developing more capable and adaptable robotic systems.

References

[1] K. He, G. Gkioxari, P. Dollár, 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, "Proximal 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.