Training Generalizable Visuomotor Policies with Real-world RL

Yinuo Zhao
Beijing Institute of Technology
Beijing, China
yinuozhao007@gmail.com

Abstract

End-to-end visuomotor policies aim to directly map visual observations to low-level control commands, facilitating autonomous manipulation in complex, unstructured environments. In recent years, imitation learning has demonstrated strong performance by leveraging large-scale human teleoperation data. However, the generalization capability of such models is inherently limited by the diversity of object configurations and manipulation strategies present in the demonstrations. In contrast, reinforcement learning (RL) offers the potential for greater scalability by enabling agents to improve through interaction. Nonetheless, deploying RL in real-world scenarios remains challenging due to issues such as inefficient exploration, limited generalization, and the high cost of scalable online training. This work addresses these challenges by proposing methods that (i) improve exploration in sparse reward settings, (ii) enhance visual generalization for robust policy execution, and (iii) enable scalable online learning via human-in-the-loop interaction. Collectively, these contributions advance the development of adaptive and resilient visuomotor policies for real-world robotic manipulation.

1. Introduction

End-to-end visuomotor policies, which map raw visual observations directly to control commands, have become increasingly important in robotic manipulation [1–3], autonomous driving [4–6], and other embodied AI domains [7]. Imitation learning, often leveraging expressive models such as diffusion [3] and Transformers [8], has dominated this field. However, its effectiveness is bounded by demonstration quality, particularly the proficiency of human operators. Reinforcement learning (RL), in contrast, offers the promise of autonomous improvement through interaction, but remains difficult to scale to real-world applications.

Three key challenges limit the deployment of RL for visuomotor control. First, sparse reward signals in real-world tasks make exploration inefficient and policies prone to suboptimal convergence. Prior work has attempted to address this with dense, handcrafted rewards [9], but such approaches require domain-specific knowledge and significant human effort, limiting scalability. Second, robust visual representation learning is critical: policies trained end-to-end often fail under visual variation because perception and control are learned jointly. Although effective techniques such as affordance learning [10, 11] and preference learning have been proposed, scalable unsupervised techniques are still needed to reduce reliance on human supervision. Third, efficient online learning with necessary human assistance remains challenging. Real-world deployment requires both hardware that facilitates seamless human feedback and algorithms that can incorporate such corrections effectively.

My research addresses these fundamental challenges of real-world RL from three complementary perspectives. First, we focus on automatically generating informative intrinsic rewards to compensate for sparse extrinsic feedback. To this end, we propose a curiosity-driven reward module that leverages shared position embedding features, enabling scalable learning in multi-agent systems. Second, we aim to enhance visuomotor policy learning by extracting control-relevant features from high-dimensional inputs. We develop unsupervised, attention-based observation augmentation and feature extraction methods to improve visual generalization and support robust RL control. Third, we incorporate essential human guidance through a low-cost teleoperation system and a human-in-the-loop RL (HIL-RL) framework, allowing efficient online adaptation in real-world settings. Together, these advances aim to enable scalable, adaptive, and resilient visuomotor policies, thereby broadening the applicability of RL in embodied AI domains.

2. Automatically Generating Informative Rewards

Training RL agents in environments with sparse or misleading rewards remains a fundamental challenge, especially in real-world embodied AI where accurate environmental states are often unobtainable due to missing or noisy sensors. Reward shaping is a widely used engineering approach to

provide more informative feedback, but it is prone to reward hacking [12], where agents maximize returns without achieving task goals, and it requires extensive human effort to refine reward functions. To address these limitations, previous research explores generating rewards from agents' own experience. For example, ICM [13] uses prediction error in the feature space as a curiosity signal to encourage exploration of uncertain states. RND [14] introduces intrinsic rewards based on output discrepancies between a predictor and a fixed randomly initialized target network. While effective for short-term exploration, it struggles with long-horizon tasks requiring global exploration. To scale intrinsic rewards to longer decision horizons and multi-agent settings, we developed DRL-CEWS [15], a spatial curiosity model with a sparse reward mechanism designed for large-scale crowdsensing environments with unevenly distributed data. The core innovation is shared position embeddings, which allow agents to leverage others' experiences to predict future positions conditioned on current actions. In addition, a chief–employee distributed computational architecture enhances sample diversity and improves exploration.

With the rapid development of large language models (LLMs) and vision–language models (VLMs), recent research has sought to leverage their strong visual–language understanding, commonsense reasoning, and coding capabilities for reward generation. One line of work queries LLMs or VLMs to synthesize reward functions from environment context or image–language inputs [16, 17]. Another line queries VLMs at every decision step to estimate task progress [18, 19]. However, these methods typically treat model outputs as ground truth. In practice, we found that VLM predictions are often inaccurate in manipulation scenarios, with errors further amplified under occlusions. To address this issue, we propose T^2-VLM [20], a training-free, temporally consistent framework that generates accurate rewards by tracking status changes in VLM-derived subgoals. Unlike prior methods, T^2-VLM requires only a single VLM query per episode, making it computationally efficient and robust to initial estimation errors of VLMs. Experiments demonstrate that our approach substantially improves failure recovery in RL-based robot manipulation policies.

3. Automatically Learning Control-related Representations

While informative reward generation can improve training efficiency, it contributes little to generalization under visual appearance changes. In such cases, extracting control-related features becomes essential. Data augmentation is a classical strategy to enhance visual generalization by enriching observation spaces. However, prior works either indiscriminately augment the entire observation [21] or focus solely on dynamic foregrounds without accounting for task relevance [22]. To address these limitations, we propose EAGLE [23], an efficient training framework for generalizable visuomotor policies. EAGLE employs a self-supervised reconstruction module to learn control-related masks, which are then used to guide control-aware data augmentation by applying strong perturbations to task-irrelevant regions. Experiments on both simulated locomotion and real-world manipulation tasks show that EAGLE achieves robust performance against unseen backgrounds and distractors.

4. Human-Copilot Reinforcement Learning

Although the long-term vision of artificial general intelligence is to reduce reliance on human intervention and scale capabilities purely with computation, human involvement remains essential for deploying RL-based visuomotor policies in the real world, both to ensure safety and to improve learning efficiency. Teleoperation platforms allow humans to collect demonstrations by directly controlling robots. However, existing teleoperation systems—such as VR setups [24], exoskeletons [25], and motion-capture technologies [26]—are primarily designed for unilateral control. In these systems, operators can issue commands but lack real-time feedback, limiting their effectiveness when robots require timely human intervention during autonomous tasks. To address this limitation, we developed HACTS (Human-As-Copilot Teleoperation System), a low-cost solution that enables bilateral, real-time joint synchronization between a robot arm and teleoperation hardware. HACTS is built using only 3D-printed components and off-the-shelf motors, making it both affordable and scalable. Building on this platform, we further propose RLPD-HACTS [27], an online reinforcement learning algorithm that integrates both offline demonstrations and online corrective feedback collected through HACTS. RLPD-HACTS employs a value-based approach to efficiently leverage these two data sources. We validated RLPD-HACTS on real-world manipulation tasks using a single-arm UR5 robot. With only 45 minutes of online training, task success improved from 50% to 80%, while average episode length decreased from 32 to 19 steps compared to offline imitation alone. These results highlight the effectiveness of HACTS in collecting high-quality data and demonstrate the necessity of online RL finetuning for robust visuomotor policies.

5. Conclusion and Future Research

This research advances the efficient training and generalization of visuomotor policies by developing methods for informative reward generation, control-related representation learning, and human-copilot online learning. We demonstrate that self-supervised approaches to reward generation and representation learning, grounded in agents' own experience, improve both training efficiency and generalization. Furthermore, we show that in complex real-world tasks, human-copilot reinforcement learning is essential for en-

hancing success rates through online finetuning. Looking ahead, we aim to extend real-world RL to distributed data collection across heterogeneous robot platforms and diverse manipulation tasks. We will also pursue more scalable RL algorithms that rely primarily on autonomous trial-and-error, thereby reducing the need for human assistance.

References

[1] A. Hämäläinen, K. Arndt, A. Ghadirzadeh, and V. Kyrki, "Affordance learning for end-to-end visuomotor robot control," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1781–1788. 1

[2] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, pp. 1–40, 2016.

[3] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023. 1

[4] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[5] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on robot learning*, 2020, pp. 66–75.

[6] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7153–7162. 1

[7] Y. Zhang, Y. Hu, Y. Song, D. Zou, and W. Lin, "Learning vision-based agile flight via differentiable physics," *Nature Machine Intelligence*, pp. 1–13, 2025. 1

[8] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023. 1

[9] A. Camacho, J. Varley, A. Zeng, D. Jain, A. Iscen, and D. Kalashnikov, "Reward machines for vision-based robotic manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 14 284–14 290. 1

[10] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "Rlafford: End-to-end affordance learning for robotic manipulation," in *2023 IEEE International conference on robotics and automation (ICRA)*, 2023, pp. 5880–5886. 1

[11] N. Yamanobe, W. Wan, I. G. Ramirez-Alpizar, D. Petit, T. Tsuji, S. Akizuki, M. Hashimoto, K. Nagata, and K. Harada, "A brief review of affordance in robotic manipulation research," *Advanced Robotics*, vol. 31, no. 19-20, pp. 1086–1101, 2017. 1

[12] I. F. Shihab, S. Akter, and A. Sharma, "Detecting and mitigating reward hacking in reinforcement learning systems: A comprehensive empirical study," *arXiv preprint arXiv:2507.05619*, 2025. 2

[13] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International conference on machine learning*, 2017, pp. 2778–2787. 2

[14] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations*, 2019. 2

[15] C. H. Liu, Y. Zhao, Z. Dai, Y. Yuan, G. Wang, D. Wu, and K. K. Leung, "Curiosity-driven energy-efficient worker scheduling in vehicular crowdsourcing: A deep reinforcement learning approach," in *2020 IEEE 36th International conference on data engineering (ICDE)*, 2020, pp. 25–36. 2

[16] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," in *The Twelfth International Conference on Learning Representations*, 2024. 2

[17] D. Venuto, M. S. N. Islam, M. Klissarov, D. Precup, S. Yang, and A. Anand, "Code as reward: Empowering reinforcement learning with vlms," in *International Conference on Machine Learning*, 2024. 2

[18] Y. Du, S. Yang, P. Florence, F. Xia, A. Wahid, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, L. P. Kaelbling *et al.*, "Video language planning," in *International Conference on Learning Representations*, 2024. 2

[19] Y. Wang, Z. Sun, J. Zhang, Z. Xian, E. Biyik, D. Held, and Z. Erickson, "Rl-vlm-f: Reinforcement learning from vision language foundation model feedback," in *International Conference on Machine Learning*, 2024. 2

[20] Y. Zhao, J. Yuan, Z. Xu, X. Hao, X. Zhang, K. Wu, Z. Che, C. H. Liu, and J. Tang, "Training-free generation of temporally consistent rewards from vlms," *arXiv preprint arXiv:2507.04789*, 2025. 2

[21] N. Hansen and X. Wang, "Generalization in reinforcement learning by soft data augmentation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 611–13 617. 2

[22] X. Wang, L. Lian, and S. X. Yu, "Unsupervised visual attention and invariance for reinforcement learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6677–6687. 2

[23] Y. Zhao, K. Wu, T. Yi, Z. Xu, Z. Che, C. H. Liu, and J. Tang, "Efficient training of generalizable visuomotor policies via control-aware augmentation," in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 2025, pp. 2832–2834. 2

[24] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems*, 2024. 2

[25] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 031–15 038. 2

[26] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023. 2

[27] Z. Xu, Y. Zhao, K. Wu, N. Liu, J. Ji, Z. Che, C. H. Liu, and J. Tang, "Hacts: a human-as-copilot teleoperation system for robot learning," *arXiv preprint arXiv:2503.24070*, 2025. 2