Learning Dexterous Object Manipulation with a Robotic Hand via Goal-Conditioned Visual Reinforcement Learning Using Limited Demonstrations

Samyeul Noh
ETRI, KAIST
Daejeon 34129, South Korea
samuel@etri.re.kr

Hyun Myung*
School of Electrical Engineering, KAIST
Daejeon 34141, South Korea
hmyung@kaist.ac.kr

Abstract

Deep reinforcement learning (deep RL) has recently achieved significant advancements, enabling agents to tackle complex tasks such as video games, locomotion, and manipulation directly from high-dimensional image pixels. Despite these successes, deep RL typically relies on domain-specific reward functions, which require expert knowledge. In this paper, we propose a goal-conditioned visual RL method that effectively learns dexterous object manipulation with a robotic hand using a goal image and limited demonstrations, without relying on domain-specific dense reward functions. Our approach leverages limited demonstrations to pre-train a policy, which is then optimized through balanced sampling between the demonstrated and online interaction data. During online interaction, it replaces human-specified dense reward functions with goal-conditioned rewards generated by a goal image and the VIP model. Experimental results demonstrate that our method achieves superior sample efficiency in dexterous object manipulation tasks with a robotic hand, even in environments with sparse or no rewards.

1. Introduction

Deep reinforcement learning (RL) has emerged as a powerful technique, capable of learning optimal policies from interaction data without the need for an explicit, hand-coded dynamics model. This approach is versatile, handling both discrete and continuous actions while utilizing either low-dimensional state vectors or high-dimensional sensor readings [6, 8, 10, 12, 14–16, 20]. Despite its promise, applying deep RL to real-world scenarios, such as robotic learning, presents significant challenges [2], including the difficulty of reliably tracking the complete system state and the complexity of crafting informative reward functions. Although recent advancements in data augmentation and self-supervised learning have improved the sample efficiency of policy learning from image observations [9, 17–19, 21], reward engineering remains a bottleneck. This process often requires domain-specific knowledge and lacks scalability to physical systems. Consequently, there is a pressing need to enhance the exploration capabilities of RL agents in environments with sparse or no rewards [4, 7].

In this paper, we introduce a goal-conditioned visual RL method that effectively learns dexterous object manipulation with a robotic hand from image observations, even in environments with sparse or no rewards. Our method leverages a goal image and the value-implicit pre-training (VIP) model [11] to generate goal-conditioned reward signals directly from image observations, eliminating the need for human-specified reward functions. The VIP model provides a self-supervised pre-trained visual representation capable of generating dense and smooth reward functions for unseen robotic tasks. Our approach begins with limited demonstrations to pre-train a policy, which is then optimized by goal-conditioned reward signals through balanced sampling between online interactions and demonstrated data. This method demonstrates superior sample efficiency on goal-image-specified robotic hand manipulation tasks from Adroit [13], achieving significant results within an extremely limited budget of 100K environment steps. Our findings reveal the potential of goal-conditioned visual RL in environments with sparse or no rewards.

2. Methodology

Our method leverages a goal image and a limited number of demonstrations to effectively learn dexterous object manipulation with a robotic hand in environments with sparse or no rewards. The process begins by using these limited demonstrations to pre-train a policy. This pre-trained policy is then optimized through goal-conditioned rewards generated by combining the goal image and the VIP model, while employing balanced sampling between demonstrated and online interaction data. Notably, in this study, we employed only five demonstrations per task. Fig. 1 provides an illus-

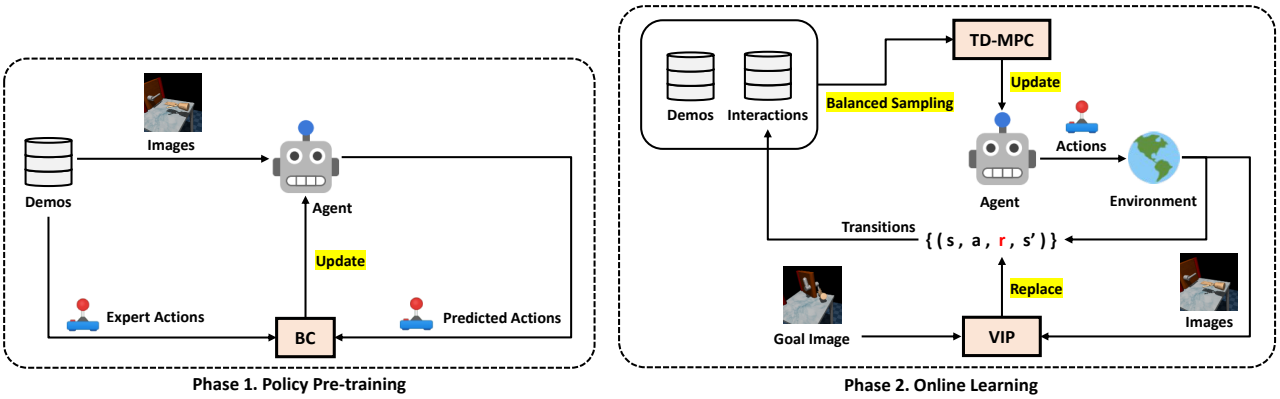


Figure 1. Goal-conditioned visual reinforcement learning with limited demonstrations. Our method first pre-trains a policy using limited demonstrations and then optimizes it with goal-conditioned rewards generated from a goal image and the VIP model. The optimization process incorporates balanced sampling between demonstrated and online interaction data, enabling effective learning in environments with sparse or no rewards.

tration of our method, and the full training procedure for goal-conditioned visual RL is summarized in Algorithm 1.

Algorithm 1 Goal-conditioned visual RL using demos

// Phase 1: Policy pre-training
for step $t = 0...T$ **do**
  Sample state-action pairs: $\{s_t, a_t\} \sim \mathcal{D}^{\text{demo}}$
  Train π_θ from $\mathcal{D}^{\text{demo}}$ using BC.
// Phase 2: Online learning
while *not converged* **do**
  Choose action: $a_t \sim \pi_\theta(s_t)$
  Step environment: $s_{t+1} \leftarrow \text{env}(a_t)$
  Fill in reward: $r_t \leftarrow r_t^{\text{VIP}}$ $(+\, r_t^{\text{sparse}})$
  Add transition (s_t, a_t, r_t, s_{t+1}) to $\mathcal{D}^{\text{online}}$.
  Train π_θ from $\mathcal{D}^{\text{demo}}$ and $\mathcal{D}^{\text{interaction}}$ using TD-MPC.

2.1. Policy Pre-training

We utilize a limited number of demonstrations to pre-train a policy using behavior cloning (BC) [3], thereby enhancing sample efficiency by establishing an inductive prior through this policy pre-training phase. BC is a straightforward yet effective approach for training a policy with expert demonstrations, often yielding impressive performance.

BC We employ BC, a widely used technique in imitation learning, to establish an inductive prior. BC trains a parameterized policy, $\pi_\theta : \mathcal{S} \to \mathcal{A}$, with the objective of predicting expert actions based on corresponding observations. However, BC has an inherent limitation: it cannot surpass the performance of the expert, as it lacks an intrinsic measure of task success. This shortcoming highlights the need to combine demonstrations with a sample-efficient RL approach to achieve superior performance.

2.2. Online Learning

We utilize the model-based RL method TD-MPC [5] to optimize the pre-trained policy through online interactions, chosen for its state-of-the-art performance in sample efficiency among online RL methods. Instead of relying on human-specified dense reward functions, we generate goal-conditioned rewards using a goal image and the VIP model during online interactions. To further enhance sample efficiency, we implement balanced sampling between demonstrated and online interaction data.

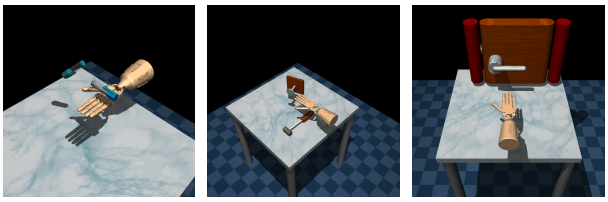
TD-MPC We employ TD-MPC as a model-based RL algorithm. TD-MPC performs local trajectory optimization in the latent space of a learned implicit (decoder-free) world model. Specifically, it optimizes the following five components:

$$\begin{aligned} &\text{Encoder} && z = h_\theta(s) \\ &\text{Latent dynamics} && z' = d_\theta(z, a) \\ &\text{Reward predictor} && \hat{r} = R_\theta(z, a) \\ &\text{Terminal value} && \hat{q} = Q_\theta(z, a) \\ &\text{Policy prior} && \hat{a} = \pi_\theta(z) \end{aligned} \tag{1}$$

where s represents a state, a represents an action, and z represents a latent representation.

The policy π_θ designed to guide planning towards high-return trajectories and is optimized to maximize temporally weighted Q-values. The remaining components are jointly optimized to minimize latent state prediction errors, reward prediction errors, and TD-errors. The overall objective is formulated as:

$$\mathcal{L}(\theta) \doteq \mathbb{E}_{(s,a,r,s')_{0:H} \sim \mathcal{D}} \left[\sum_{t=0}^{H} \lambda^t \Big(e_{\text{hd}} + e_{\text{R}} + e_{\text{Q}} \Big) \right], \tag{2}$$



(a) (b) (c)

Figure 2. Object manipulation tasks using the Shadow dexterous hand in Adroit. (a) adroit-pen, (b) adroit-hammer, and (c) adroit-door.

where $e_{\text{hd}} = ||d_\theta(z_t, a_t) - \text{sg}(h_\theta(s'_t))||_2^2$ represents latent state prediction errors, $e_{\text{R}} = ||R_\theta(z_t, a_t) - r_t||_2^2$ represents reward prediction errors, and $e_{\text{Q}} = ||Q_\theta(z_t, a_t) - (r_t + \gamma Q_{\bar{\theta}}(z'_t, \pi_\theta(z'_t))||_2^2$ represents TD-errors. Here, $\bar{\theta}$ denotes an exponential moving average of θ.

VIP We utilize the VIP model to generate smooth and dense reward signals from image observations based on a given goal image. The VIP model learns the optimal goal-conditioned value function through a dual offline goal-conditioned RL formulation, defined as:

$$\mathcal{L}(\phi) \doteq \mathbb{E}_{p(g)}\Big[(1-\gamma)\mathbb{E}_{\mu_0(s;g)}\big[-V_\phi(s;g)\big] + \log\mathbb{E}_{(s,s';g)\sim\mathcal{D}}\big[\exp\big(-V_\phi(s;g) - (1-\gamma V_\phi(s';g))\big)\big]\Big], \tag{3}$$

where $V_\phi(s;g) = -||\phi(s) - \phi(g)||_2$ denotes the goal-conditioned value function, parameterized implicitly as the L_2 distance in the embedding space by ϕ; g represents a goal image; $p(g)$ indicates the goal distribution; and $\mu_o(s;g)$ refers to the goal-conditioned distribution of the initial image. The implicit value function effectively provides a dense reward signal conditioned on the specified goal.

3. Experiments

We present experimental results on three dexterous object manipulation tasks using a robotic hand from Adroit. The specific tasks include 'adroit-pen,' 'adroit-hammer,' and 'adroit-door.'

Setup The tasks involve object manipulation with the Shadow dexterous hand [1], as illustrated in Fig. 2. In these experiments, we consider learning from image observations and focus on goal-image-specified tasks, where each task is specified via a goal image. During online interaction, VIP-generated reward signals are used to replace the human-engineered (simulation-provided) dense rewards. We evaluate our method under a limited budget of 100K online interactions.

Baseline As a baseline, we consider training TD-MPC with sparse rewards, substituting the human-specified (simulation-provided) dense rewards.

Results The experimental results, presented in Fig. 3, evaluate five different configurations: (a) baseline, (b) baseline with policy pre-training, (c) baseline with policy pre-training and VIP-generated rewards, (d) baseline with policy pre-training, VIP-generated rewards, and balanced sampling, and (e) baseline with policy pre-training, VIP-generated rewards, and balanced sampling, but eliminating sparse rewards. Our findings reveal that TD-MPC struggles to succeed in dexterous object manipulation tasks under sparse reward conditions, despite its state-of-the-art sample efficiency in environments with human-specified dense rewards. Although pre-training the policy with limited demonstrations offers some advantages, its overall effect remains marginal. Notably, VIP-generated rewards significantly accelerate the RL training process in the 'adroit-hammer' task, though they are less effective in other tasks. Balanced sampling emerges as a critical factor for enhancing sample efficiency. Finally, training TD-MPC exclusively with VIP-generated rewards achieves satisfactory success rates in the 'adroit-door' task, even in reward-free environments, although this approach generally falls short in other scenarios.

4. Conclusion

In this study, we introduced a goal-conditioned visual RL method utilizing limited demonstrations. Our approach begins by pre-training a policy with these limited demonstrations to establish an inductive prior, and then optimizes the pre-trained policy through goal-conditioned rewards generated by a goal image and the VIP model, eliminating the need for human-specified reward functions. Additionally, balanced sampling between demonstrated and online interaction data is employed to enhance sample efficiency. Our results show that our method achieves satisfactory success rates using only a goal image, even in environments without explicit rewards. Furthermore, the success rates improve when sparse rewards are available. Future work will explore unsupervised RL approaches to further enhance sample efficiency by pre-training a world model in an unsupervised manner prior to policy optimization.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [24ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling, and evolving ways].

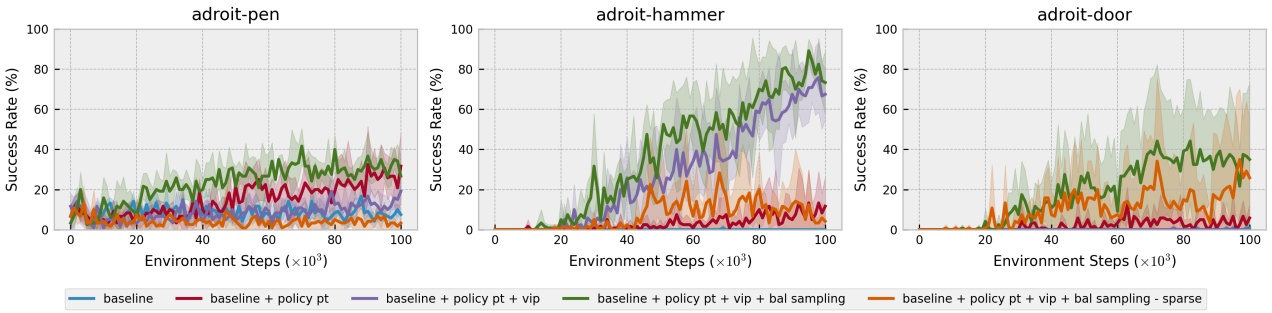


Figure 3. Experimental results for three dexterous object manipulation tasks, averaged over four seeds. Shaded areas indicate 95% confidence intervals.

References

[1] Shadow dexterous hand, [Online]. https://www.shadowrobot.com/dexterous-hand-series/. Accessed: (Aug. 2, 2024). 3

[2] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *IJRR*, 39 (1):3–20, 2020. 1

[3] Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *ICML*, pages 12–20, 1997. 2

[4] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *ICLR*, 2019. 1

[5] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *ICML*, pages 8387–8406, 2022. 2

[6] Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: Lessons we have learned. *Int. J. Robot. Res.*, 40(4-5):698–721, 2021. 1

[7] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *CoRL*, pages 651–673. PMLR, 2018. 1

[8] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *J. Mach. Learn. Res.*, 22(1): 1395–1476, 2021. 1

[9] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *ICML*, pages 5639–5650, 2020. 1

[10] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17(1):1334–1373, 2016. 1

[11] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023. 1

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. 1

[13] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *RSS*, 2018. 1

[14] William B Shen, Danfei Xu, Yuke Zhu, Leonidas J Guibas, Li Fei-Fei, and Silvio Savarese. Situational fusion of visual representation for visual navigation. In *ICCV*, pages 2881–2890, 2019. 1

[15] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

[16] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. 1

[17] Guowei Xu, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, et al. DrM: Mastering visual reinforcement learning through dormant ratio minimization. In *ICLR*, 2024. 1

[18] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *ICLR*, 2020.

[19] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *ICLR*, 2022. 1

[20] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *ICCV*, pages 15222–15232, 2021. 1

[21] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé III, and Furong Huang. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. *NeurIPS*, 36, 2023. 1