Addressing Sparse Rewards in Visual Reinforcement Learning via Balanced Online–Offline Sampling Under Scarce Demonstrations

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

Seonghyun Kim
ETRI
Daejeon, South Korea
kim-sh@etri.re.kr

Ingook Jang
ETRI
Daejeon, South Korea
ingook@etri.re.kr

***Abstract*—Sparse reward environments pose a significant challenge in reinforcement learning (RL) due to the difficulty of acquiring sufficient informative experiences through exploration alone. This challenge is particularly severe in visual RL, where agents must learn from high-dimensional pixel observations, making exploration less efficient and reward propagation more difficult. Incorporating expert demonstrations can alleviate this issue, but in many real-world scenarios, only a limited number of high-quality demonstrations are available. In this paper, we propose Balanced Online–Offline Sampling (BOOS), a novel online RL training strategy that adaptively combines scarce offline demonstrations with online interaction data to improve sample efficiency in sparse reward settings. BOOS dynamically adjusts the sampling ratio between offline and online data using an exponential decay schedule, prioritizing demonstrations during early training while progressively increasing reliance on online exploration. We evaluate BOOS on two widely used visual robotic manipulation benchmarks, Meta-World and Adroit, where sparse rewards and high-dimensional observations present significant learning challenges. Experimental results show that BOOS significantly outperforms state-of-the-art pure online RL algorithms in both sample efficiency and final task performance. These findings highlight the potential of BOOS as a practical solution for visual RL under the dual constraints of sparse rewards and scarce demonstrations.**

***Index Terms*—demonstrations, robotic manipulation, sparse rewards, visual reinforcement learning.**

I. Introduction

Reinforcement Learning (RL) has achieved remarkable progress in a variety of domains, including robotic manipulation, game playing, and autonomous navigation [1]–[6]. These successes, however, have been predominantly demonstrated in environments with dense and informative reward signals, where agents can obtain frequent feedback to guide policy improvement. In contrast, sparse reward environments—where rewards are provided only upon task completion or after reaching specific milestones—pose a severe challenge for RL agents [7], [8]. The lack of frequent feedback often results in inefficient exploration, slow convergence, and unstable performance.

The challenge becomes even more pronounced in visual RL, where agents learn directly from high-dimensional pixel observations rather than low-dimensional state representations [9]–[16]. High-dimensional inputs exacerbate the credit assignment problem and further reduce the efficiency of exploration, as the agent must not only discover rewarding behaviors but also extract meaningful features from raw images. As a result, training an effective policy in sparse reward visual RL often requires an impractically large number of environment interactions.

A promising approach to mitigate sparse reward difficulties is to incorporate expert demonstrations, which can provide informative examples of successful behavior and bootstrap policy learning [17]–[19]. While this approach has proven effective in various domains, it often relies on access to a large number of high-quality demonstrations. In practice, however, such data can be expensive or infeasible to obtain—especially in robotic manipulation tasks, where expert time is limited and physical data collection is costly. This leads to the scarce demonstration setting, where the available demonstration data is both limited in quantity and critical to the agent's learning success.

Existing hybrid RL methods that combine offline demonstration data with online environment interactions offer a promising way to address this challenge [20]–[22]. However, most existing methods rely on fixed or heuristically chosen sampling ratios between offline and online data, which can be suboptimal. In the early stages of training, demonstration data is most valuable for establishing an initial policy, while in later stages, the agent should rely more heavily on online exploration to discover novel strategies and improve generalization. Without an adaptive sampling mechanism, scarce demonstrations risk being underutilized or overexploited, leading to poor sample efficiency or overfitting.

To address these limitations, we propose Balanced Online–Offline Sampling (BOOS), a novel online RL training strategy that adaptively integrates scarce offline demonstrations with online interaction data. BOOS employs an exponential decay schedule to prioritize demonstration data in the early stages and gradually shift toward online experience as learning progresses. This approach ensures that scarce demonstrations provide maximum benefit during policy initialization while still allowing for sustained exploration in later phases. We

evaluate BOOS on two widely used visual robotic manipulation benchmarks, Meta-World and Adroit, which are both characterized by sparse rewards and high-dimensional observations. Experimental results show that BOOS consistently outperforms state-of-the-art online RL methods in terms of both sample efficiency and final task performance.

Our contributions can be summarized as follows:

- An adaptive sampling framework for integrating scarce demonstrations with online data in sparse reward visual RL tasks.
- A dynamic sampling schedule that prioritizes demonstrations in early training and gradually transitions to exploration-driven online learning.
- Extensive empirical evaluation on Meta-World and Adroit, showing substantial improvements in both sample efficiency and final task performance over state-of-the-art pure online RL baselines.

Through this work, we aim to provide a practical and effective solution for real-world RL applications where agents must learn under the combined constraints of sparse rewards, high-dimensional visual observations, and limited demonstration availability.

II. Related Work

A. Sparse Reward Reinforcement Learning

Sparse reward environments have long been recognized as a fundamental challenge in RL. In such settings, agents receive meaningful feedback only after completing a sequence of successful actions, making random exploration highly inefficient. A variety of methods have been proposed to address this issue, including reward shaping [7], intrinsic motivation [8], and goal-conditioned RL [23], [24]. While these approaches can accelerate learning, they often rely on prior domain knowledge or additional engineered signals, which may not be available in real-world scenarios. Moreover, in visual RL, sparse reward problems become even more severe, as the agent must concurrently learn both the task policy and a high-dimensional visual representation, significantly increasing the sample complexity.

B. Visual Reinforcement Learning

Visual RL focuses on learning policies from raw pixel observations rather than low-dimensional state features. Recent works have explored representation learning techniques such as contrastive learning [9], [10], autoencoders [11], [12], and data augmentation [13]–[16] to improve sample efficiency. However, in sparse reward visual RL, these methods alone often fail to overcome the exploration bottleneck, as effective task learning still requires discovering rare rewarding states. This difficulty motivates the use of expert demonstrations to guide exploration and representation learning simultaneously.

C. Learning from Demonstrations (LfD)

Learning from Demonstrations has been widely studied as a means to bootstrap RL agents in challenging environments. Imitation Learning (IL) methods, such as behavioral cloning (BC) [17] and inverse RL [25], directly learn policies from demonstration data. While effective with abundant high-quality demonstrations, IL suffers when demonstration coverage is insufficient. Offline RL methods [18], [19] extend this paradigm by enabling policy optimization without further environment interaction, but typically require large, diverse datasets. In scarce demonstration settings, neither pure IL nor offline RL is sufficient, as limited data coverage can lead to poor generalization and overfitting.

D. Hybrid Online–Offline Reinforcement Learning

Hybrid approaches that combine offline demonstrations with online interaction have shown promise in addressing both exploration inefficiency and data scarcity. Notable examples include DAPG [20], which fine-tunes policies initialized with demonstrations, and methods that interleave offline and online updates [21], [22]. However, most existing hybrid strategies use fixed or heuristically chosen sampling ratios between offline and online data, which can be suboptimal—especially in scarce demonstration regimes. Without a mechanism to adaptively adjust the sampling balance over time, these methods risk either overfitting to the small demonstration set or underutilizing its guidance.

Our work builds on this line of research by introducing an adaptive sampling strategy that dynamically balances offline and online data during training. By prioritizing demonstrations in the early phase and gradually shifting toward online exploration, our method maximizes the utility of scarce demonstrations while maintaining exploration efficiency, particularly in visual RL tasks with sparse rewards.

III. Methodology

In this section, we present BOOS, a novel online RL training strategy designed to address the combined challenges of sparse rewards, high-dimensional sensory inputs (specifically, image pixels), and scarce demonstrations. BOOS dynamically balances the use of scarce offline demonstrations and online environment interactions during training. The key idea is to leverage demonstrations heavily at the start of learning—when they provide maximum guidance—and gradually shift toward online exploration as the agent's policy improves.

A. Problem Formulation

We consider an RL agent interacting with a Markov Decision Process (MDP) defined by $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state (or observation) space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ is the transition dynamics, $r(s,a)$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. In *visual RL*, the agent observes $o_t \in \mathbb{R}^{H\times W\times C}$, a high-dimensional image rather than a low-dimensional state s_t. We assume access to a scarce offline demonstration dataset $D_{\text{offline}} = \{(o,a)\}_{i=1}^{N}$ collected by an expert policy π_E, where N is very small (e.g., 1–5 trajectories). The objective is to learn a policy $\pi_\theta(a|o)$ that maximizes the expected discounted return while making efficient use of D_{offline} and maintaining strong online exploration.

B. Balanced Online–Offline Sampling Schedule

A central component of BOOS is the adaptive sampling ratio $p_{\text{offline}}(t)$, which determines the fraction of training samples drawn from the offline demonstration dataset at training step t. This ratio is designed to start high to accelerate policy bootstrapping from expert data and decay gradually to encourage online exploration while preventing overfitting. Accordingly, we define:

$$p_{\text{offline}}(t) = \max\left(\alpha \cdot e^{-\beta t},\, p_{\min}\right), \tag{1}$$

where $\alpha \in (0, 1]$ is the initial offline sampling proportion, $\beta > 0$ controls the decay speed, and $p_{\min} > 0$ ensures demonstrations are still occasionally used in later stages.

At each training iteration, we sample a batch of size B:

$$B_{\text{offline}} = \lfloor p_{\text{offline}}(t) \cdot B \rfloor, \quad B_{\text{online}} = B - B_{\text{offline}}. \tag{2}$$

Here, $p_{\text{offline}}(t) \in [0, 1]$ denotes the fraction of offline samples used at iteration t, and the floor operator $\lfloor \cdot \rfloor$ ensures integer batch sizes.

C. Policy Optimization

BOOS is agnostic to the choice of the underlying RL algorithm. In this work, we implement BOOS on top of TD-MPC [26] for continuous control. The combined batch $\mathcal{B} = \mathcal{B}_{\text{offline}} \cup \mathcal{B}_{\text{online}}$ is used to update a latent world model via the following objectives:

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

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

The policy π_θ is optimized to maximize long-term returns by guiding the agent towards high-value trajectories. The overall objective of the world model is to jointly minimize the latent state prediction error, reward prediction error, and temporal difference (TD)-error, as formalized in the following loss function:

$$\mathcal{L}(\theta) \doteq \mathbb{E}_{(s,a,r,s')_{0:H} \sim \mathcal{D}} \left[\sum_{t=0}^{H} \lambda^t \Big(l_{\text{P}} + l_{\text{R}} + l_{\text{Q}} \Big) \right], \tag{3}$$

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

IV. Experiments

We evaluate BOOS on two challenging visual robotic manipulation benchmarks, Meta-World and Adroit, focusing on sample efficiency and final task performance under sparse reward conditions and scarce demonstrations.

A. Benchmarks

We conduct experiments on:

- **Meta-World** [27]: A suite of MuJoCo-based multi-task manipulation environments with high-dimensional image observations. We use sparse reward variants of three representative tasks, including `pick-place`, `box-close`, and `assembly`, where rewards are provided only upon successful completion.
- **Adroit** [20]: A set of dexterous manipulation tasks (`pen`, `hammer`, and `door`) featuring complex dynamics, high-dimensional observations, and sparse success-based rewards. We use the image-based task versions to match the visual RL setting.

B. Baselines

We compare BOOS against:

- **TD-MPC**: A model-based RL algorithm achieving state-of-the-art sample efficiency in dense reward settings.
- **BC Pre-training**: Behavioral cloning from demonstrations followed by no further RL fine-tuning.

C. Demonstration Setup

For each task, we collect only five demonstration trajectories using an expert policy. To emulate real-world constraints, no additional demonstration data is available during training. Demonstrations are stored in D_{offline}, while D_{online} is populated during interaction.

D. Implementation Details

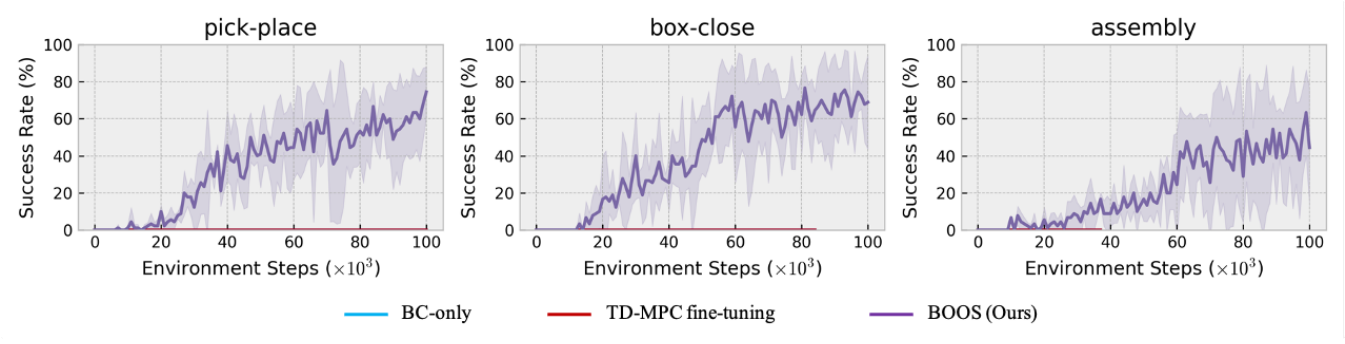
We implement BOOS and baselines in PyTorch. The offline sampling ratio parameters are set as $\alpha = 0.75$, $\beta = 1 \times 10^{-6}$, $p_{\min} = 0.25$.

E. Results

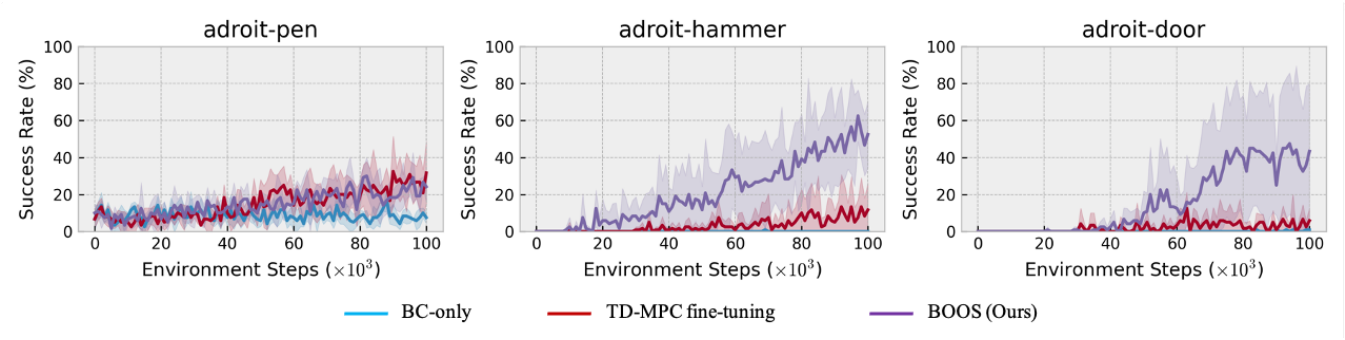
Figure 1 compares three configurations: (a) **BC-only**: pre-training with BC using limited demonstrations without RL fine-tuning; (b) **TD-MPC fine-tuning**: initializing with BC and training via vanilla TD-MPC; (c) **BOOS** (ours): initializing with BC and fine-tuning using our adaptive online–offline sampling with a world model.

Our experimental results show that **BC-only** fails to solve most tasks, reflecting the limitations of imitation learning when demonstration coverage is small. **TD-MPC fine-tuning** performs well in dense reward settings but struggles under sparse rewards, exhibiting failure on most tasks in Meta-World and slow convergence and low final success rates in Adroit. **BOOS** achieves substantially faster learning and higher asymptotic performance across all tasks, particularly under severe demonstration scarcity. Our adaptive sampling consistently accelerates fine-tuning by leveraging demonstration data early and enabling robust exploration later.

These results confirm that balancing expert and agent-generated data is critical for improving sample efficiency and achieving high performance under sparse reward constraints.



(a)



(b)

Fig. 1: Experimental results averaged over four random seeds for two challenging robotic manipulation benchmarks. Shaded regions represent 95% confidence intervals. (a) Performance results for three robotic manipulation tasks from the Meta-World benchmark. (b) Performance results for three dexterous object manipulation tasks from the Adroit benchmark.

V. Conclusion

In this study, we proposed BOOS, an effective online RL training strategy to address both sparse reward and scarce demonstration challenges. Our method leverages demonstrations heavily in early training and gradually increases online exploration, ensuring efficient use of limited expert data.

Our experimental results on Meta-World and Adroit demonstrate that our method significantly improves sample efficiency and final task performance compared with state-of-the-art pure online RL algorithms and BC. Our method is algorithm-agnostic and well-suited for real-world scenarios where collecting demonstrations is costly.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [25ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling, and evolving ways]. It was also supported by internal fund/grant of ETRI [25YR1500, A study on generalizable action intelligence and its application to physical robots].

References

[1] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," Journal of Machine Learning Research (JMLR), vol. 17, no. 1, pp. 1334–1373, 2016.

[2] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "MT-Opt: Continuous multi-task robotic reinforcement learning at scale," arXiv preprint, vol. arXiv:2104.08212, 2021.

[3] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: Lessons we have learned," International Journal of Robotics Research (IJRR), vol. 40, no. 4-5, pp. 698–721, 2021.

[4] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," Journal of Machine Learning Research (JMLR), vol. 22, no. 1, pp. 1395–1476, 2021.

[5] S. Seo and H. Jung, "A robust collision prediction and detection method based on neural network for autonomous delivery robots," ETRI Journal, vol. 45, no. 2, pp. 329–337, 2023.

[6] H. Park and S. H. Yoon, "Deep reinforcement learning for base station switching scheme with federated lstm-based traffic predictions," ETRI Journal, vol. 46, no. 3, pp. 379–391, 2024.

[7] A. Y. Ng et al., "Policy invariance under reward transformations: Theory and application to reward shaping," in Proc. International Conference on Machine Learning (ICML), 1999.

[8] D. Pathak, "Curiosity-driven exploration by self-supervised prediction," in Proc. International Conference on Machine Learning (ICML), 2017.

[9] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus, "Improving sample efficiency in model-free reinforcement learning from images," AAAI Conference on Artificial Intelligence (AAAI), vol. 35, no. 12, pp. 10674–10681, 2021.

[10] J. Auh, C. Cho, and S.-t. Kim, "Improved contrastive learning model via identification of false-negatives in self-supervised learning," ETRI Journal, vol. 46, no. 6, pp. 1020–1029, 2024.

[11] S. B. Alex and L. Mary, "Variational autoencoder for prosody-based speaker recognition," ETRI Journal, vol. 45, no. 4, pp. 678–689, 2023.

[12] Y. Ko, S. Ko, and Y. Kim, "Generative autoencoder to prevent overregularization of variational autoencoder," ETRI Journal, vol. 47, no. 1, pp. 80–89, 2025.

[13] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," in Proc. International Conference on Learning Representations (ICLR), 2022.
[14] R. Zheng, X. Wang, Y. Sun, S. Ma, J. Zhao, H. Xu, H. Daum´ e III, and F. Huang, "TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning," Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[15] G. Xu, R. Zheng, Y. Liang, X. Wang, Z. Yuan, T. Ji, Y. Luo, X. Liu, J. Yuan, P. Hua, et al., "DrM: Mastering visual reinforcement learning through dormant ratio minimization," in Proc. International Conference on Learning Representations (ICLR), 2024.
[16] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in Proc. International Conference on Learning Representations (ICLR), 2020.
[17] C. Atkeson and S. Schaal, "Robot learning from demonstration," in Proc. International Conference on Machine Learning (ICML), vol. 97, pp. 12–20, 1997.
[18] A. Kumar, et al., "Conservative q-learning for offline reinforcement learning," in Proc. Neural Information Processing Systems (NeurIPS), pp. 1179–1191, 2020.
[19] Y. Chebotar et al., "Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions," in Proc. Conference on Robot Learning (CoRL), pp. 3909–3928, 2023.
[20] A. Rajeswaran et al., "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in Proc. Robotics: Science and Systems (RSS), 2018.
[21] A. Nair et al., "AWAC: Accelerating online reinforcement learning with offline datasets," arXiv preprint, vol. arXiv:2006.09359, 2021.
[22] Q. Yang and Y. Wang, "ATraDiff: Accelerating online reinforcement learning with imaginary trajectories," in Proc. International Conference on Machine Learning (ICML), pp. 56485–56500, 2024.
[23] M. Liu, M. Zhu, and W. Zhang, "Goal-conditioned reinforcement learning: Problems and solutions," in Proc. International Joint Conference on Artificial Intelligence (IJCAI), pp. 5502–5511, 2022.
[24] M. Bortkiewicz et al., "Accelerating goal-conditioned reinforcement learning algorithms and research," in Proc. International Conference on Learning Representations (ICLR), 2025.
[25] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in Proc. International Conference on Machine Learning (ICML), 2004.
[26] N. Hansen, H. Su, and X. Wang, "Temporal difference learning for model predictive control," in Proc. International Conference on Machine Learning (ICML), pp. 8387–8406, 2022.
[27] T. Yu et al., "Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning," in Proc. Conference on Robot Learning (CoRL), pp. 1094–1100, 2020.