Efficient Fine-Tuning of Behavior Cloned Policies with Reinforcement Learning from Limited Demonstrations

Samyeul Noh, Seonghyun Kim, Ingook Jang
ETRI
Daejeon 34129, Republic of Korea
samuel@etri.re.kr

Abstract

Behavior cloning (BC) is a supervised learning technique in which an agent mimics expert behavior based on demonstration data. While BC is widely applied in robotics due to its simplicity, it is constrained by challenges such as dataset bias, limited adaptability, and an inability to outperform the expert. In this study, we introduce an efficient fine-tuning approach that combines BC with reinforcement learning (RL), leveraging limited demonstrations to mitigate these limitations. Our approach refines the pre-trained BC policy by incorporating a world model to facilitate synthetic rollouts for planning and policy optimization, and by balancing data sampling between expert-provided demonstrations and agent-driven online interactions. We perform experiments in environments with high-dimensional image observations and employ sparse reward signals in place of human-engineered dense reward functions. The experimental results demonstrate that our method significantly improves sample efficiency, enabling the successful learning of complex robotic manipulation tasks within a restricted budget of 100K environment steps.

1 Introduction

Behavior cloning (BC) is a supervised learning technique where an agent learns to replicate expert behavior by mimicking actions demonstrated in specific situations (Atkeson & Schaal, 1997). BC has been widely applied in robotics and autonomous systems, particularly for tasks such as autonomous driving (Codevilla et al., 2019; Ly & Akhloufi, 2020; Li et al., 2022; Wang et al., 2024) and robotic manipulation (Ahn et al., 2022; Brohan et al., 2022, 2023a). Its primary advantages lie in its simplicity and ease of implementation, as it directly learns from expert demonstrations without requiring complex reward functions. However, BC has significant limitations, including dataset bias, poor generalization to unseen scenarios, and the inability to exceed the performance of the expert. These challenges make BC less suitable for tasks that require adaptability to novel environments or improvement beyond the expert's performance.

In contrast, reinforcement learning (RL) is a machine learning paradigm in which an agent learns optimal actions through direct interaction with its environment, receiving feedback in the form of rewards or penalties (Sutton, 2018). RL excels in solving complex, sequential decision-making tasks by maximizing long-term rewards through a trial-and-error process (Mnih et al., 2015; Levine et al., 2016; Silver et al., 2017; Vinyals et al., 2019). However, RL is often hindered by sample inefficiency, as it typically requires a large amount of interaction data and computational resources to converge to an optimal policy. This inefficiency is especially problematic in real-world robotics applications, where data collection is costly, time-consuming, or unsafe.

To address the limitations inherent in both BC and RL, we propose an efficient fine-tuning approach that integrates these two methods. Our approach leverages a limited set of expert demonstrations to initialize a policy through BC, followed by RL-based fine-tuning to enhance adaptability and

38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024).

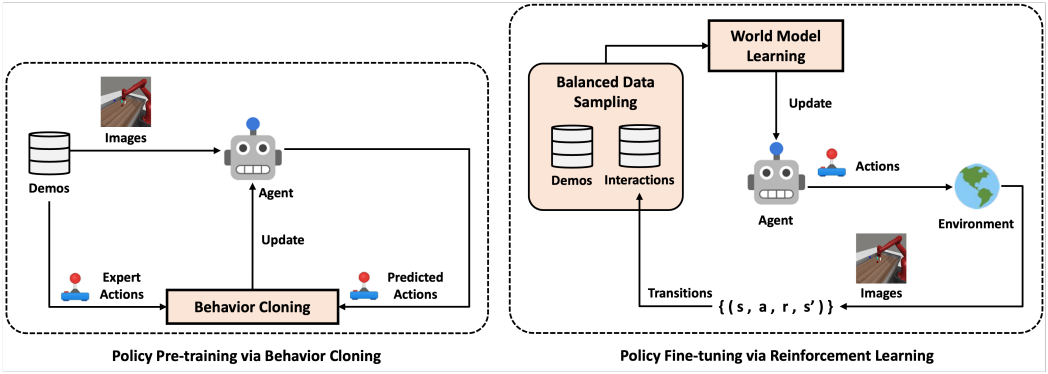


Figure 1: Efficient RL-based fine-tuning approach utilizing limited demonstrations. Our method employs a limited set of expert demonstrations to initialize a policy using BC and subsequently fine-tunes it using RL to improve adaptability and generalization. Specifically, our method fine-tunes the pre-trained BC policy by learning a world model that facilitates long-term planning, and by balancing data sampling between expert-provided demonstrations and agent-driven online interactions.

generalization. Specifically, our method fine-tunes the pre-trained BC policy by learning a world model that facilitates long-term planning, and by balancing data sampling between expert-provided demonstrations and agent-driven online interactions.

Our method further addresses the challenge of reward function design by employing sparse reward signals during RL-based fine-tuning in environments with high-dimensional image observations. This eliminates the need for human-engineered dense reward functions, which are not only challenging to design but may also introduce unintended biases. Through a series of experiments on various robotic manipulation tasks, we demonstrate that our approach significantly improves sample efficiency, enabling robust policy learning within a constrained budget of interaction steps while outperforming baseline methods in both efficiency and performance.

2 Related Work

2.1 Behavior Cloning

BC is a supervised learning approach that aims to mimic expert demonstrations to learn optimal policies. This technique has been widely used as a pre-training method for RL tasks to accelerate policy learning (Atkeson & Schaal, 1997). In the context of robotics and control, BC has proven to be effective in initializing policies that can then be fine-tuned using RL to achieve better performance in complex environments (Brohan et al., 2023b, 2022). Recent advancements in offline RL (Chebotar et al., 2023) have further highlighted the synergy between BC and offline datasets, allowing models to learn effectively from static data before transitioning to an online fine-tuning phase. In this study, in contrast to previous studies, we utilize only five expert demonstrations to pre-train policies, and these same five demonstrations are reused during RL-based fine-tuning.

2.2 Reinforcement Learning

RL has been extensively studied for its ability to learn optimal policies through trial and error interactions with the environment. Fine-tuning policies that have been pre-trained via BC has been a common approach to accelerate learning in challenging environments. Some RL methods such as proximal policy optimization (PPO) (Schulman et al., 2017) and deep Q-networks (DQN) (Mnih, 2013) have benefited from BC as a starting point, enabling faster convergence and improved performance. Additionally, meta-learning approaches, such as model-agnostic meta-learning (MAML) (Finn et al., 2017) framework, have been employed to enhance the adaptability of RL agents, making them more efficient in fine-tuning for new tasks. The combination of RL with pre-training strategies has been a key focus in recent literature to improve sample efficiency and reduce the computational burden of training RL agents from scratch. In this study, we fine-tune the pre-trained BC policy by incorporating world model learning to facilitate synthetic rollouts for planning and policy opti-

mization, and by balancing data sampling between expert-provided demonstrations and agent-driven online interactions, further improving sample efficiency and ensuring effective policy learning in environments with sparse rewards.

3 Methodology

In this section, we present our efficient fine-tuning approach via RL, which incorporates two key components: (a) learning a world model to enable long-term planning and improve sample efficiency, and (b) balancing data sampling between expert-provided demonstrations and agent-driven online interactions, as illustrated in Fig. 1. These components work in synergy to optimize policy learning in complex, sequential decision-making tasks.

3.1 Policy Pre-training via Behavior Cloning

We initially pre-train a policy using expert demonstrations through BC. BC is a straightforward yet effective technique that leverages expert behavior to train a policy, π_θ, which predicts expert actions based on corresponding observations. This method has demonstrated significant success across various robotic tasks, particularly when large datasets of expert demonstrations are available. However, in this study, we specifically focus on a setting with limited demonstrations — only five expert demonstrations are utilized for pre-training. Our hypothesis is that, while a BC policy trained with limited demonstrations may struggle to solve complex tasks, it can establish a useful inductive prior for subsequent fine-tuning through RL.

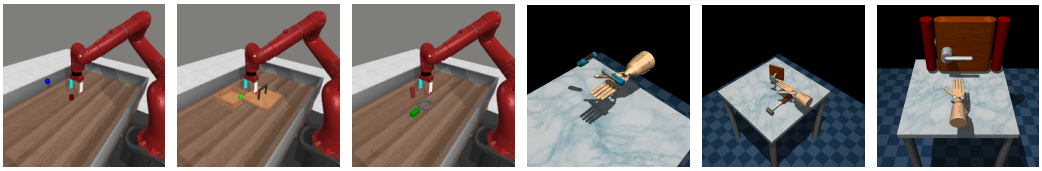
Behavior Cloning BC trains a parameterized policy, $\pi_\theta : \mathcal{S} \to \mathcal{A}$, to predict expert actions $a_t \in \mathcal{A}$ given an observation $s_t \in \mathcal{S}$. The objective is to minimize the discrepancy between the expert's action and the agent's predicted action. Despite its simplicity, BC suffers from inherent limitations: it is limited by the quality and quantity of the provided demonstrations and cannot surpass the expert's performance since it lacks any feedback mechanism to assess task success beyond imitation. This limitation underscores the necessity of integrating RL to refine the BC policy, allowing the agent to exceed expert-level performance by exploring the environment and optimizing based on a reward signal.

3.2 Policy Fine-tuning via Reinforcement Learning

After initializing the policy with BC, we fine-tune it using RL to enhance adaptability and generalization in novel environments, such as those where object positions vary. Our method incorporates two key components: (a) learning a world model to facilitate sample efficiency and long-term planning, and (b) balancing the data sampling between expert-provided demonstrations and online interactions driven by the agent.

World Model Learning World models provide the agent with an explicit understanding of its environment by allowing it to predict future states and rewards. This internal model serves as a mechanism for model-based RL, where the agent performs rollouts in the latent space to optimize long-term outcomes. In environments with high-dimensional image observations, we employ latent dynamics models, which abstract high-dimensional inputs into compact latent representations. These latent representations enable forward predictions that guide planning and policy improvement.

Two notable approaches for world model learning are Dreamer (Hafner et al., 2020) and TD-MPC (Hansen et al., 2022), with the latter being particularly relevant for this study due to its decoder-free world model architecture. TD-MPC operates by performing local trajectory optimization in the latent space, minimizing the reliance on raw image reconstruction and focusing on latent dynamics. This approach allows for more efficient rollouts, optimizing five core components: the encoder, latent



(a) (b) (c) (d) (e) (f)

Figure 2: Two challenging robotic manipulation benchmarks, including three tasks from the Meta-World benchmark: (a) *pick-place*, (b) *box-close*, and (c) *assembly*, as well as three dexterous object manipulation tasks from the Adroit benchmark: (d) *adroit-pen*, (e) *adroit-hammer*, and (f) *adroit-door*.

dynamics, reward predictor, terminal value, and policy prior, as follows:

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

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

The policy π_θ 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:

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

where $l_{\text{hd}} = ||d_\theta(z_t, a_t) - \text{sg}(h_\theta(s'_t))||_2^2$ 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.

By using TD-MPC, our method achieves greater efficiency in environments with high-dimensional image observations, facilitating long-term planning and enabling sample-efficient policy optimization.

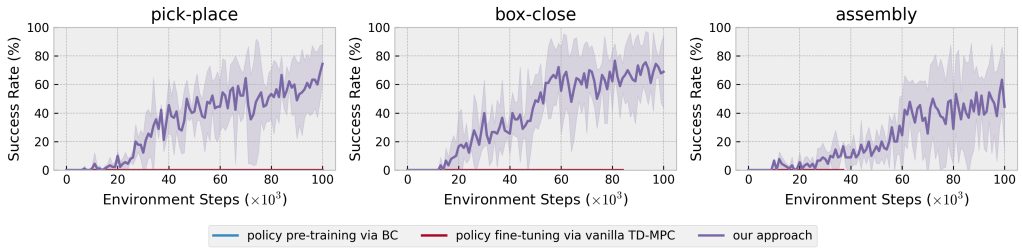
Balanced Data Sampling Balanced data sampling plays a crucial role in improving the efficiency of policy fine-tuning within the RL framework. Our method balances data between expert-provided demonstrations and agent-driven online interactions to ensure a more effective exploration-exploitation tradeoff. By combining the structured knowledge from demonstrations with the agent's exploratory data, we prevent the policy from becoming overly reliant on either strategy. This approach allows the agent to leverage expert knowledge while continuously improving its policy through exploration of novel states and actions, thus enhancing both sample efficiency and generalization to unseen scenarios. The balanced sampling method also mitigates the issue of sparse rewards in RL by providing structured data points from demonstrations, which helps the agent focus on high-value actions early in training. In this study, we use a 50:50 ratio for expert demonstrations and agent-driven online interactions to optimize learning performance across diverse environments.

4 Experiments

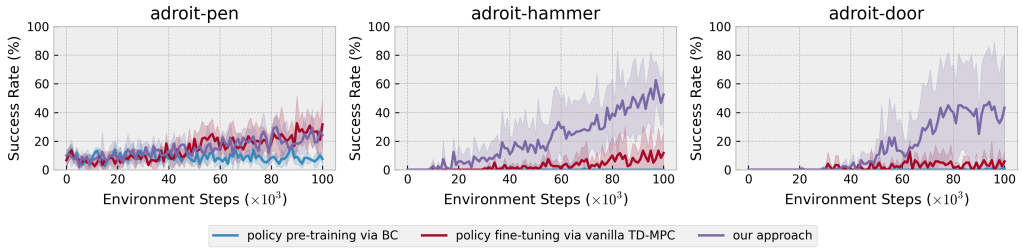
In this section, we present experimental results on various robotic manipulation tasks to evaluate the effectiveness of our RL-based fine-tuning approach, with a focus on sample efficiency and asymptotic performance.

4.1 Setup

We evaluate our method on a subset of robotic manipulation tasks from the Meta-World benchmark (Yu et al., 2020) and dexterous object manipulation tasks from the Adroit benchmark (Rajeswaran et al., 2018). In these benchmarks, we consider learning from high-dimensional image observations



(a)



(b)

Figure 3: 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.

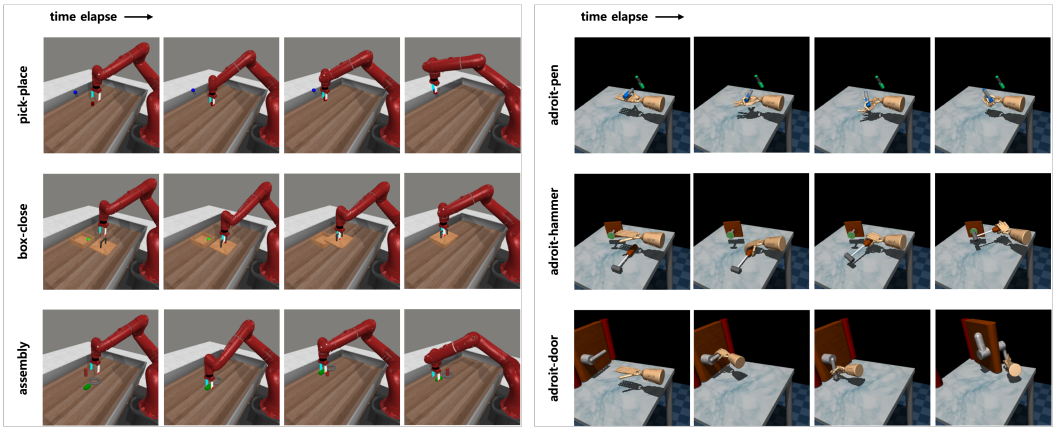
and sparse task-completion rewards, rather than human-engineered dense rewards. In this context, we limit the number of online interactions to a budget of 100K environment steps. The primary focus is on assessing how effectively our method can learn under sparse reward conditions and with limited data availability.

4.2 Baselines

We consider two baselines: (a) pre-training a policy using BC with limited demonstrations, without any further fine-tuning, and (b) fine-tuning the pre-trained BC policy using vanilla TD-MPC.

4.3 Results

The experimental results, illustrated in Fig. 3, compare three different configurations: (a) policy pre-training via BC, specifically pre-training a policy using BC with limited demonstrations, without further fine-tuning); (b) policy fine-tuning via vanilla TD-MPC, specifically fine-tuning the pre-trained BC policy using vanilla TD-MPC); and (c) our approach, specifically fine-tuning the pre-training BC policy by learning a world model to facilitate long-term planning and by balancing data sampling between expert-provided demonstrations and agent-driven online interactions. Our findings indicate that policy pre-training via BC, particularly when using a limited number of expert demonstrations, is insufficient for solving the robotic manipulation tasks. This result indicates that BC requires a larger dataset to succeed in these tasks. Although TD-MPC has demonstrated state-of-the-art performance in sample efficiency in environments with human-engineered dense rewards, it struggles to perform well under sparse reward conditions, both in the Meta-World and Adroit benchmarks. Notably, incorporating balanced data sampling with world model learning significantly accelerates the RL-based fine-tuning process across all evaluated tasks. This result suggests that balancing data sampling between expert-provided demonstrations and agent-driven online interactions is a crucial factor for improving sample efficiency and achieving successful task completion, even within a constrained interaction budget and under sparse reward conditions. We provide visualizations of the successful trajectories generated by policies fine-tuned with our approach in Fig. 4.



(a) (b)

Figure 4: Visualization of successful trajectories generated by our RL-based fine-tuning approach on two challenging robotic manipulation benchmarks. (a) Visualization of three tasks from the Meta-World benchmark: *pick-place*, *box-close*, and *assembly*. (b) Visualization of three dexterous object manipulation tasks from the Adroit benchmark: *adroit-pen*, *adroit-hammer*, and *adroit-door*.

5 Conclusion

In this study, we introduced an efficient fine-tuning method that integrates BC with RL to address their respective limitations, particularly when working with limited demonstrations. Our approach uses BC to initialize the policy and then refines it with RL, improving adaptability and generalization by mitigating dataset bias and performance constraints inherent in BC. By incorporating world model learning and balancing data sampling between expert demonstrations and agent-driven interactions, our method enhances both sample efficiency and asymptotic performance, especially under sparse reward conditions.

Experimental results demonstrated the effectiveness of our approach in terms of sample efficiency and overall performance compared to baseline methods. The balanced data sampling strategy proved to be key in accelerating the RL-based fine-tuning process. These finding suggests that our method provides a practical solution for improving policy learning in environments with high-dimensional image observations and sparse rewards.

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].

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *International Conference on Machine Learning*, volume 97, pp. 12–20. Citeseer, 1997.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023a.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pp. 287–318. PMLR, 2023b.

Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pp. 3909–3928. PMLR, 2023.

Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, pp. 8387–8406. PMLR, 2022.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Guofa Li, Zefeng Ji, Shen Li, Xiao Luo, and Xingda Qu. Driver behavioral cloning for route following in autonomous vehicles using task knowledge distillation. *IEEE Transactions on Intelligent Vehicles*, 8(2):1025–1033, 2022.

Abdoulaye O Ly and Moulay Akhloufi. Learning to drive by imitation: An overview of deep behavior cloning methods. *IEEE Transactions on Intelligent Vehicles*, 6(2):195–209, 2020.

Volodymyr Mnih. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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.

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 *Robotics: Science and Systems*, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

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.

Xudong Wang, Chao Wei, Hanqing Tian, Weida Wang, and Jibin Hu. Implicit predictive behavior cloning for autonomous driving decision-making in urban traffic. *IEEE Transactions on Intelligent Vehicles*, 2024.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pp. 1094–1100. PMLR, 2020.