Efficient Device-Cloud Collaborative Offline-to-Online Reinforcement Learning

Extended Abstract

Honglan Huang
Rocket Force University of Engineering
Xi'an, China
huanghonglan17@alumni.nudt.edu.cn

Wei Shi
Rocket Force University of Engineering
Xi'an, China
shiwei15@nudt.edu.cn

Chaoyue Niu
Shanghai Jiao Tong University
Shanghai, China
Rvince@sjtu.edu.cn

ABSTRACT

Federated reinforcement learning in device-cloud architectures suffers from high interaction costs, low sample efficiency, and slow convergence due to device constraints. We propose a data-centric device-cloud collaborative training method that pre-trains a global model on the cloud to provide a high-quality initial policy and selects high-value samples to guide safe and efficient local policy fine-tuning on devices. Experiments on public benchmarks show our method achieves significantly faster convergence, higher training efficiency, and improved policy stability, outperforming baselines.

KEYWORDS

Device-cloud collaborative; Federated reinforcement learning; Offline-to-online reinforcement learning; Prioritized experience replay

ACM Reference Format:
Honglan Huang, Wei Shi, and Chaoyue Niu. 2026. Efficient Device-Cloud Collaborative Offline-to-Online Reinforcement Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/STCC8399

1 INTRODUCTION

The proliferation of intelligent devices underscores the importance of device-cloud collaboration, which integrates cloud computing with edge sensing to support edge intelligence. Federated reinforcement learning (FRL) is particularly promising in this paradigm due to its distributed and privacy-preserving nature. However, FRL often suffers from low sample efficiency and slow convergence caused by cold starts in practice.

The offline-to-online reinforcement learning (O2ORL) paradigm offers a potential solution by pre-training a policy on offline datasets before fine-tuning it online. In device-cloud collaboration, this corresponds to cloud-based pre-training of a global model. However, a naïve combination often leads to policy collapse due to distribution shift between offline data and the online environment, negating the benefits of pre-training.

Corresponding author: Wei Shi.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/STCC8399

To address this, we propose an O2ORL framework for device-cloud collaboration. The cloud first trains an initial model using offline RL. Then, a designed data selection strategy transmits high-value offline experiences along with the model to clients, effectively guiding and accelerating their online training while mitigating the risk of policy collapse. Our contributions are:

(1) An O2ORL framework for efficient, privacy-preserving agent evolution in device-cloud systems.
(2) A novel offline experience selection strategy to guide client-side online learning and prevent policy collapse.
(3) Experimental validation on public datasets, showing improved convergence speed and performance over baselines.

2 RELATED WORK

FRL. Recent research focuses on improving efficiency, privacy, and handling heterogeneity. Methods include reward-based client selection [1, 2, 11], convergence-guaranteed algorithms for heterogeneous policies [14], and techniques to enhance sample efficiency [5, 6, 8]. Privacy concerns are addressed via differential privacy [16] or secure frameworks [7, 12]. However, leveraging cloud-stored historical data for offline pre-training to accelerate federated online training remains underexplored.

O2ORL. This paradigm aims to bridge the offline-online distribution shift. Key approaches include adaptive update strategies [17], conservative value learning [3, 9], stabilized fine-tuning [15], and improved exploration [4]. Theoretical analysis [10] and techniques for smooth transition [13] have also been investigated. Yet, these methods are not designed for federated constraints concerning heterogeneity and communication efficiency.

3 METHODOLOGY

This section presents the Efficient Device-Cloud Collaborative Offline-to-Online RL (DCC-O2ORL) framework. Its core idea is to leverage cloud-based offline pre-training to obtain a high-quality initial policy, combined with a novel offline data selection and transmission mechanism that mitigates policy collapse during subsequent online fine-tuning. The overall process is illustrated in Fig. 1.

The DCC-O2ORL procedure consists of two main phases:

(1) **Cloud-based Offline Pre-training & Data Selection:** The server trains an initial global policy using an offline RL algorithm on its historical dataset. It then computes the advantage values for offline samples and selects a subset $\mathcal{D}_{\text{off}}$ with high advantage values to guide client-side online training.

(2) **Federated Online Fine-tuning:** The server broadcasts the initial global model and $\mathcal{D}_{\text{off}}$ to clients. Each client integrates $\mathcal{D}_{\text{off}}$ with its locally collected online experiences in a replay buffer for local policy updates. Clients periodically upload encrypted model updates to the server for aggregation, after which the updated global model is redistributed. Steps ⑤-⑧ in the federated phase repeat until convergence.

$\mathcal{D}_{\text{off}}$ provides guided exploration for clients, accelerating online learning and stabilizing the transition from offline pre-training. If necessary, the offline data selection (Steps ②-④) can be repeated during early training rounds to further refine the guidance.

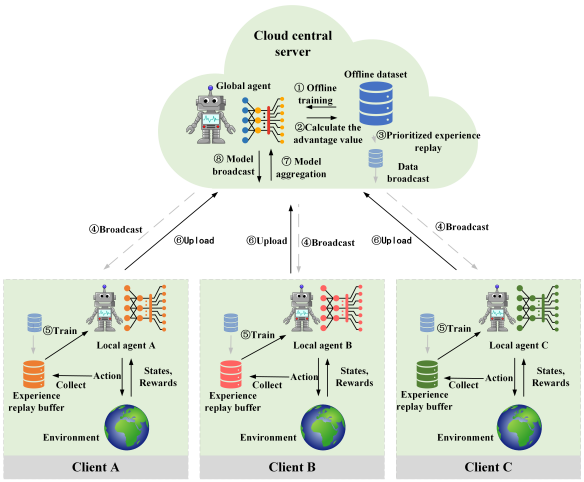

Figure 1: Efficient DCC-O2ORL algorithm flow chart

4 EXPERIMENTS AND ANALYSIS

4.1 Experimental Setup

We evaluate our framework on MuJoCo continuous control tasks (HalfCheetah, Hopper, Walker2d) using offline datasets from D4RL of varying quality (random, medium, expert, etc.). We compare DCC-O2ORL against several baselines: 1) TD3-single (single-agent online RL); 2) TD3 (federated online RL from scratch); 3) TD3-all (federated online RL with full offline data); 4) TD3+BC (federated offline pre-training with random data selection). Our method enhances TD3+BC with prioritized offline data selection (PER).

4.2 Main Results

As summarized in Table 1, DCC-O2ORL consistently achieves the highest or competitive final performance across most dataset qualities. Crucially, it significantly mitigates the policy collapse phenomenon commonly observed during the offline-to-online transition. For example, in the HalfCheetah task (Fig. 2), while baselines suffer severe performance drops initially, DCC-O2ORL maintains stable and rapidly improving performance. This demonstrates the framework's robustness to varying offline data quality and its effectiveness in accelerating online convergence.

Table 1: Mean ± standard deviation of cumulative rewards during online training

Task	Dataset	TD3-all	TD3+BC	DCC-O2ORL
HalfCheetah	random	4474.91±95.00	2677.14±14.66	**5902.79±63.73**
	expert	4629.03±275.13	5384.03±70.28	**5588.86±83.99**
	medium	4272.13±82.26	4046.84±47.98	**5535.34±92.19**
	medium-expert	4501.23±275.66	5464.11±81.76	**5816.99±91.64**
	medium-replay	4653.19±68.65	3665.95±47.08	**5154.96±68.19**
	full-replay	4835.53±87.04	4864.69±60.12	**5793.28±80.39**
Hopper	random	968.18±102.70	360.81±1.54	**1544.46±91.88**
	expert	1617.65±60.64	**1721.57±36.59**	1714.54±36.44
	medium	1543.52±47.72	1554.80±11.37	**1697.32±26.65**
	medium-expert	1617.71±62.91	**1719.04±24.34**	1703.69±60.97
	medium-replay	1595.03±58.78	1596.98±16.11	**1667.50±78.86**
	full-replay	1634.11±66.32	1630.10±52.28	**1697.68±90.01**
Walker2D	random	**392.84±76.22**	208.04±94.71	42.11±36.85
	expert	2332.18±46.97	2458.81±10.63	**2499.13±45.82**
	medium	1946.79±66.61	1954.42±11.70	**2407.86±30.31**
	medium-expert	2284.52±38.88	2501.71±10.58	**2543.17±42.75**
	medium-replay	2130.88±27.34	1972.75±16.63	**2370.78±33.59**
	full-replay	2153.04±33.61	2118.79±12.15	**2485.17±42.84**

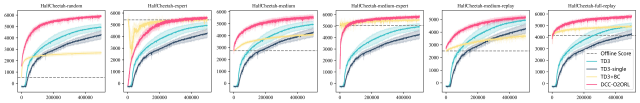

Figure 2: Test performance of different algorithms during online training

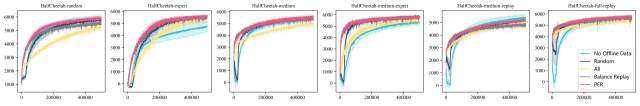

Figure 3: Test performance of client local models under different offline data selection strategies

4.3 Ablation Study on Data Selection

We ablate the offline data selection strategy within our framework. Compared to alternatives (No offline data, Random selection, Full data, Balanced replay), our PER-based strategy proves most effective. As shown in Fig. 3, PER avoids significant policy collapse in the initial online phase, while other strategies exhibit pronounced performance drops. This validates that selectively transmitting high-value offline experiences is key to guiding early online exploration and stabilizing training.

5 CONCLUSION

We propose DCC-O2ORL, an offline-to-online reinforcement learning framework for device-cloud collaboration. It mitigates low sample efficiency and policy collapse via prioritized experience replay that selects high-value offline data from the cloud to clients. Experiments show accelerated convergence, improved performance, and robust policy collapse mitigation across datasets.

REFERENCES

[1] Honglan Huang, Wei Shi, Yanghe Feng, Chaoyue Niu, Guangquan Cheng, Jincai Huang, and Zhong Liu. 2024. Active Client Selection for Clustered Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems* 35, 11 (2024), 16424–16438. https://doi.org/10.1109/TNNLS.2023.3294295

[2] Woonghee Lee. 2023. Reward-based participant selection for improving federated reinforcement learning. *ICT Express* 9, 5 (2023), 803–808. https://doi.org/10.1016/j.icte.2022.08.008

[3] Yixiu Mao, Yun Qu, Qi Wang, and Xiangyang Ji. 2025. Adaptive Neighborhood-Constrained Q Learning for Offline Reinforcement Learning. *arXiv preprint arXiv:2511.02567* (2025).

[4] Trevor A. McInroe, Stefano V. Albrecht, and Amos J. Storkey. 2023. Planning to Go Out-of-Distribution in Offline-to-Online Reinforcement Learning. *ArXiv preprint* abs/2310.05723 (2023). https://arxiv.org/abs/2310.05723

[5] Yun Qu, Qi Wang, Yixiu Mao, Vincent Tao Hu, Björn Ommer, and Xiangyang Ji. 2025. Can prompt difficulty be online predicted for accelerating rl finetuning of reasoning models? *arXiv preprint arXiv:2507.04632* (2025).

[6] Yun Qu, Qi Cheems Wang, Yixiu Mao, Yiqin Lv, and Xiangyang Ji. 2025. Fast and Robust: Task Sampling with Posterior and Diversity Synergies for Adaptive Decision-Makers in Randomized Environments. *arXiv preprint arXiv:2504.19139* (2025).

[7] F. Sattler, K. R. Muller, and W. Samek. 2021. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems* 32, 8 (2021), 3710–3722.

[8] Jin Wang, Jia Hu, Jed Mills, Geyong Min, Ming Xia, and Nektarios Georgalas. 2023. Federated Ensemble Model-Based Reinforcement Learning in Edge Computing. *IEEE Transactions on Parallel and Distributed Systems* 34, 6 (2023), 1848–1859. https://doi.org/10.1109/TPDS.2023.3264480

[9] Zhe Wang, Jia Hu, Geyong Min, Zhiwei Zhao, and Zi Wang. 2024. Agile Cache Replacement in Edge Computing via Offline-Online Deep Reinforcement Learning. *IEEE Transactions on Parallel and Distributed Systems* 35, 4 (2024), 663–674. https://doi.org/10.1109/TPDS.2024.3368763

[10] Teng Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. 2021. Policy Finetuning: Bridging Sample-Efficient Offline and Online Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27395–27407. https://proceedings.neurips.cc/paper/2021/hash/e61eaa38aed621dd776d0e67cfeee366-Abstract.html

[11] Zhijie Xie and Shenghui Song. 2023. FedKL: Tackling Data Heterogeneity in Federated Reinforcement Learning by Penalizing KL Divergence. *IEEE Journal on Selected Areas in Communications* 41, 4 (2023), 1227–1242. https://doi.org/10.1109/JSAC.2023.3242734

[12] Jiangchao Yao, Feng Wang, Kunyang Jia, Bo Han, Jingren Zhou, and Hongxia Yang. 2021. Device-Cloud Collaborative Learning for Recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 3865–3874. https://doi.org/10.1145/3447548.3467097

[13] Zishun Yu and Xinhua Zhang. 2023. Actor-Critic Alignment for Offline-to-Online Reinforcement Learning. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:260957123

[14] Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. 2024. Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning. arXiv:2401.15273 [cs.LG] https://arxiv.org/abs/2401.15273

[15] Yinmin Zhang, Jie Liu, Chuming Li, Yazhe Niu, Yaodong Yang, Yu Liu, and Wanli Ouyang. 2023. A Perspective of Q-value Estimation on Offline-to-Online Reinforcement Learning. *ArXiv preprint* abs/2312.07685 (2023). https://arxiv.org/abs/2312.07685

[16] Fangyuan Zhao, Xuebin Ren, Shusen Yang, Peng Zhao, Rui Zhang, and Xinxin Xu. 2023. Federated multi-objective reinforcement learning. *Information Sciences: an International Journal* 624 (2023), 811–832. https://doi.org/10.1016/j.ins.2022.12.083

[17] Han Zheng, Xufang Luo, Pengfei Wei, Xuan Song, Dongsheng Li, and Jing Jiang. 2023. Adaptive Policy Learning for Offline-to-Online Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:257505480