Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 339-348

Imitation learning to accelerate training process of multi-agent reinforcement learning in 2v2 pong game

Marvin Yonathan Hadiyanto*[1], Budi Harsono[1], Indra Karnadi[1], Ivan Tanra[1]
Department of Electrical Engineering, Krida Wacana Christian University, Indonesia[1]

Article Info
Keywords:
Imitation Learning, MARL, Reinforcement Learning, Pong Game, Learning from Demonstration

Article history:
Received: November 04, 2025
Accepted: February 26, 2026
Published: May 01, 2026

Cite:
M. Y. Hadiyanto, B. Harsono, I. Karnadi, and I. Tanra, "Imitation Learning to Accelerate Training Process of Multi-Agent Reinforcement Learning in 2v2 Pong Game", *KINETIK*, vol. 11, no. 2, May 2026.
https://doi.org/10.22219/kinetik.v11i2.2564

*Corresponding author.
Marvin Yonathan Hadiyanto
E-mail address:
marvin.yonathan@ukrida.ac.id

Abstract
Training multi-agent reinforcement learning (MARL) systems often requires a significant amount of time due to sample inefficiency, particularly when agents must perform extensive exploration in complex environments and coordinate among multiple entities. This study proposes the use of imitation learning to accelerate the MARL training process in a 2v2 pong game by leveraging demonstrations from a 1v1 pong game to shape the initial policy without undergoing inefficient exploration procedures. We employ a deep Q-network (DQN) framework with centralized training and decentralized execution (CTDE) to compare the performance of pretrained and untrained agents in the 2v2 pong environment. Experimental results show that learning from demonstrations in the 1v1 setting improves reward accumulation and game scores of pretrained agents in the 2v2 pong game. The performance improvement peaks at 700 demonstration learning steps and diminishes at larger learning steps due to excessive memorization of the demonstration gameplay. Furthermore, comparative experiments demonstrate that imitation learning with 700 learning steps achieves learning efficiency improvements of approximately 300% and 571% compared to the zonation method and standard reinforcement learning pretraining, respectively. These results indicate that imitation learning from demonstrations can effectively reduce the prolonged training process in MARL, offering a viable solution, particularly when data collection, computational resources, and training time are severely constrained.

1. Introduction

Reinforcement learning (RL) is a machine learning method to tackle the problem of sequential decision making by training an agent through interactions with the environment aiming to achieve optimal behavior that maximizes the cumulative reward over time [1], [2], [3]. One of the main challenges in training a reinforcement learning agent is a low sample efficiency that requires enormous number of samples from interaction with the environment to achieve a desirable level of agent performance [4], [5], [6]. This difficulty makes many implementations of RL in real world problems are not practical especially when data collecting is a complicated task. Learning from a demonstration can be used to reduce the rigor of the training process due to sample inefficiency by providing an example of expert behavior to be imitated by the agent that enables the learning process acceleration [7], [8], [9], [10], [11], [12]. Several previous studies have shown that learning from a demonstration can reduce training steps by mimicking expert data, Todd Hester et al. have proposed algorithm called deep Q-learning from demonstrations (DQfD) that can significantly accelerate the learning process of RL agent in games [13] and Yang Gao et al. have shown a method to improve the agent from imperfect expert demonstrations [14].

In the real world, the environment is complex and involves the dynamic of multiple entities [15], [16], [17], for example the problem of autonomous driving [18], multi-robot warehouse management [19], and multiplayer games which require MARL to enable multiple agents to learn and adapt through multi-entity environment [20]. The nature of multiple entities led to more severe sample inefficiency and greater complexity of data collecting process since MARL requires learning coordination with other agents as well as the dynamics of the environment. Coordination in MARL can be fully cooperative, fully competitive, or combination of cooperation with teams and competition across opponents [21]. There are two common approaches of learning in MARL, namely decentralized and centralized, which depends on the type of global or independent access of state, action, and reward. In centralized learning, agents are trained using global information of state, action, and reward, while in decentralized learning, agents use local information of state, action, and reward [22]. The combination of those two approaches called CTDE has recently emerged as a popular method for training the agents in MARL which can partially access the global information such as state and action of other agents in training while keeping each agent act independently using only its local observation [23]. Several studies have been conducted by implementing this method, for example, Zhou et al. have implemented this method in game StarCraft

II [24] and Marvin et al. for their double snake game [25], both have shown good results of the method. However, these previous results still suffer from inefficient training due to the sample inefficiency that is exacerbated by the complex nature of the dynamics of MARL environment.

Although prior studies have demonstrated the effectiveness of learning from demonstrations and CTDE in MARL, most existing approaches assume that expert demonstrations are collected within the same task structure and interaction dynamics as the target MARL environment. As a result, these methods still require extensive data collection in complex multi-agent settings and remain vulnerable to severe sample inefficiency as the number of agents increases. Moreover, there is limited empirical investigation into whether demonstrations obtained from simpler, lower-entity environments, such as single-agent or 1v1 scenarios can be effectively transferred to accelerate learning in more complex cooperative MARL problems with substantially different interaction dynamics. To address this methodological and empirical gap, this paper proposes an imitation learning framework that transfers demonstrations from a 1v1 Pong environment to initialize and accelerate MARL training in a 2v2 Pong setting. The main contribution of this work is an empirical demonstration that cross-task imitation learning can improve learning efficiency and early performance in cooperative MARL, while also identifying an optimal demonstration length that balances imitation and reinforcement learning.

In this paper we proposed an imitation learning method to accelerate MARL training process to learn playing 2v2 pong game. We use demonstrations of playing 1v1 pong game then transfer the policy for further adaptation into 2v2 pong game. The proposed 2v2 pong game involved 4 players that is team up 2 by 2 that are controlled by their respective agent to maximize score of their own team. The rules of this game are derived from the classic pong game, however the 2v2 pong game is played by 2 teams that consist of 2 players for each team, a point is collected when a pong successfully score the ball into the opponent zone. In our experiment, firstly, we train agents by using DQN from the state, action, and reward in 1v1 pong game demonstration, the resulting DQN is then used by two agents in the same team as the initial policy to further learn 2v2 pong game against two agents without any prior model as the opponent team. The results in this work show that by using demonstration in 1v1 pong game for the pretrained team, the learning rate of MARL in 2v2 pong game is improved compared to the untrained team. With this result we have shown that learning from demonstration can be used to resolve sample inefficiency under a different and substantially more complex dynamics than the original demonstration is conducted.

2. Research Method

2.1 Pong Game

Pong game is a simplified form of tennis where two paddles attempt to keep a ball in play. It has been widely used as a benchmark environment in RL due to its relatively simple mechanics and extended interaction horizon. The observation space is moderately complex, containing features such as the players' scores and visual information like side walls, paddle positions, and ball positions. Each agent controls a paddle located on either the left or right side of the screen, each paddle can be controlled to move up, move down, or rest. The objective of playing pong game is to hit a moving ball back and forth and prevent it from passing one's paddle, a team score is counted when the opponent fails to return the ball. In this work we propose the 2v2 pong game which derives from the classical pong game.

The number of players in 2v2 pong game is 4 players that are teamed up in 2 by 2 as can be seen in Figure 1. The 2v2 pong game starts with a ball moving in a random direction at an initial speed, as depicted in Figure 1(b). When the ball collides with the upper or lower barrier, the vertical speed of the ball is reversed while the horizontal speed of the ball will remain the same, this will result in a bounce vertically when the ball touches the upper or lower barrier. When the ball collides with a paddle the horizontal direction of the ball is reversed and the total speed is increased by 10%, resulting in a horizontal bounce with a speed 1.1 times before the collision. The score of a team will be increased by one point when the ball successfully surpasses the opponents paddle line and the pong game will be reset to start the initial condition. In 2v2 pong game environment, each agent will be rewarded with small amounts of point (less than 1) if the agent is chasing the ball in vertical direction, while all agents of scoring team will be rewarded with 1 point respectively if a team successfully scoring. The state of 2v2 pong game that is used to train the agents is a vector of 16 elements, this state is:

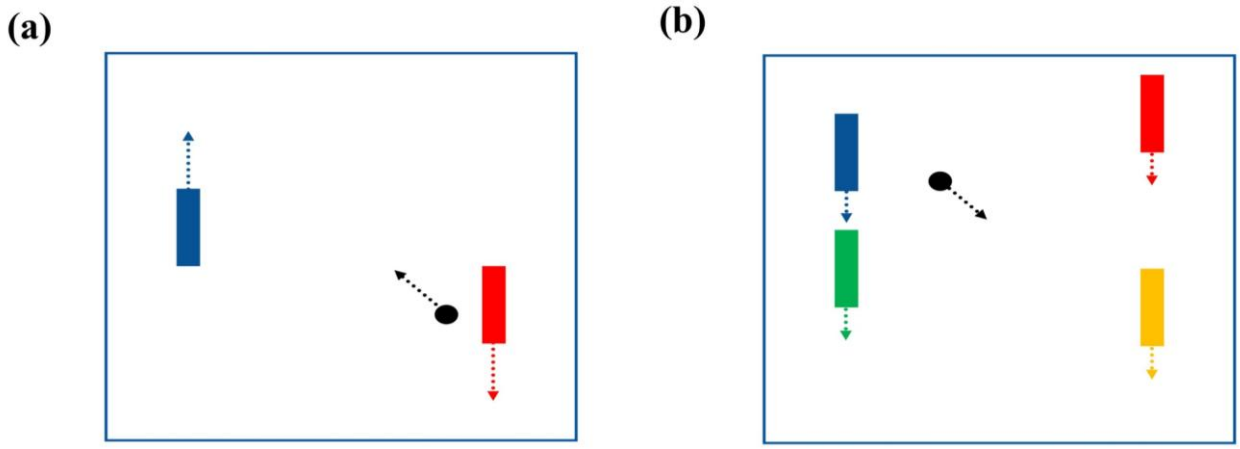

Figure 1. (a) 1v1 Pong Game and (b) 2v2 Pong Game

2.2 Imitation Learning

To accelerate the training process of MARL agents in learning to play 2v2 pong game, as shown in Figure 2, we initialize the agents learning with demonstration that controls the agents' action in 1v1 pong game. In training the initial policy, action is taken from the demonstration while reward and state are obtained from the 1v1 pong game environment. With this approach, we bypass the initial exploration by providing agents with a more reliable state, action, and reward tuple from the demonstration to avoid the sample inefficiency during training due to the unpredictability nature of the exploration.

After the initial training guided by demonstration for a certain number of learning steps in 1v1 pong game, the pretrained agents are transferred into 2v2 pong game. Albeit the same game mechanics between 1v1 and 2v2 pong game, the pretrained agents encounter a different dynamic due to the emergence of two additional players that impose the necessity of coordination among players. The illustration of MARL in 2v2 pong game can be seen in Figure 3, the left side are the pretrained agents while the right side are the untrained agents. Each agent consists of an independent DQN that can predict the value of actions when a particular input state is given. In the training process of an agent in 2v2 pong game, action is controlled by agent to improve their respective reward function through RL. The left side players will benefit from the initial demonstration of 1v1 pong game while the right side players will be started without any prior knowledge about the pong game. By comparing the performance of pretrained and untrained teams, we can observe the effectiveness of imitation learning of 1v1 pong game in the dynamics of 2v2 pong game to reduce the sample inefficiency in MARL training process.

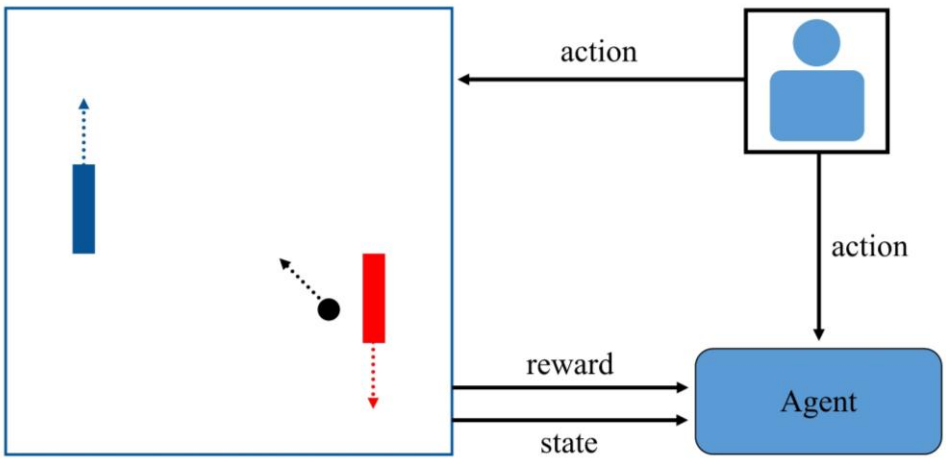

Figure 2. Imitation Learning from Demonstration in 1v1 Pong Game

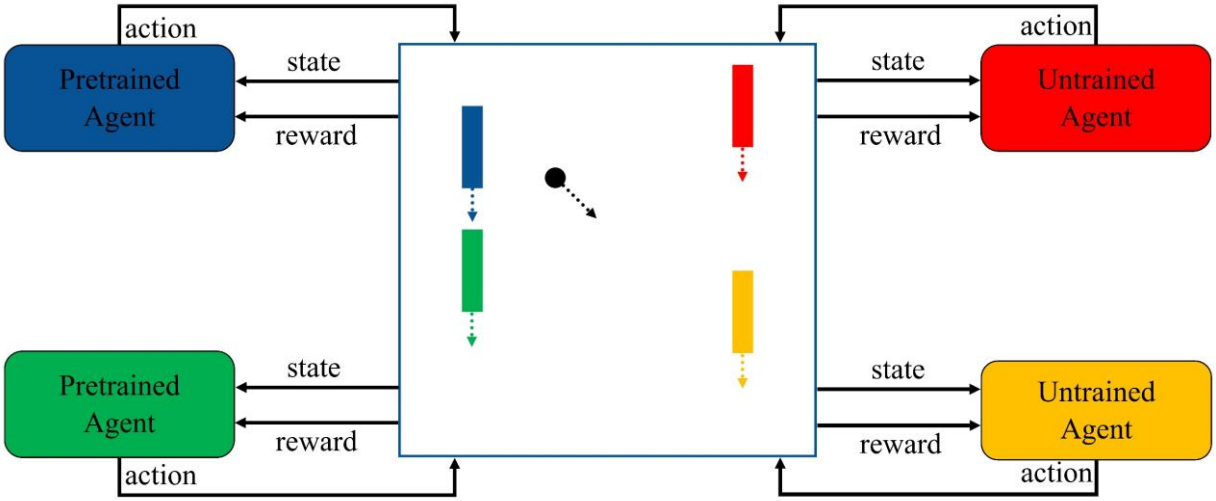

Figure 3. Multi-agent Reinforcement Learning of 2v2 Pong Game

state = [top left paddle position, top left paddle up direction, top left paddle stop,
bottom left paddle position, bottom left paddle up direction, bottom left paddle stop,
top right paddle position, top right paddle up direction, top right paddle stop,
bottom right paddle position, bottom right paddle up direction, bottom right paddle stop,
ball x position, ball y position,
ball right direction, ball up direction].

3. Results and Discussion

In our experiment we use a fully connected neural network that consists of five layers with 7 nodes for input layers, three layers of 64 nodes, output layer that consists of 3 nodes, for training the reward function we use $\alpha = 0.0005$ and $\gamma = 0.95$. In imitation learning, the number of learning steps from demonstrations are 100, 400, 700, and 1000. We observe the sum of reward and score from each agent in the pretrained and untrained teams to compare the performance between the teams. To see the fluctuations of their performance, the reward and score are taken five times, hence the average value as well as distribution between maximum and minimum value can be further analysed. In 2v2 pong game we record the first 1000 steps to examine the effectiveness of 1v1 pong game imitation learning in the dynamics of 2v2 pong game MARL.

Figure 4 shows the reward value of pretrained and untrained teams when playing 1000 steps of 2v2 pong game at learning steps of 100, 400, 700, and 1000. The pretrained team consistently achieves higher rewards than the untrained team across all learning steps variations. This performance gain is attributed to prior imitation learning, which initializes the reward function based on demonstrations from the 1v1 pong game. The average rewards at the final step of 2v2 pong game for the pretrain team at learning steps of 100, 400, 700, and 1000 are 19.73 ± 2.71, 20.90 ± 1.41, 23.12 ± 5.68, and 22.92 ± 2.96 respectively (see Table 1). In contrast, the untrained team achieves lower average rewards of 12.89 ± 1.12, 9.90 ± 1.10, 9.66 ± 1.03, and 10.06 ± 2.08 respectively. Figure 5 presents the average score values of pretrained and untrained teams under the same experimental settings. A similar trend to the reward results is observed, where the pretrained agents' scores generally increase with additional learning steps and slightly decrease when the learning step exceeds 700. The average final scores of the pretrained team for learning steps of 100, 400, 700, and 1000 are 5.00 ± 1.58, 5.60 ± 1.14, 6.60 ± 3.65, and 6.40 ± 1.14 respectively (see Table 2). Meanwhile, the untrained team achieves lower average scores of 1.20 ± 0.45, 0.40 ± 0.55, 0.20 ± 0.45, and 0.40 ± 0.55 respectively. Overall, increasing the learning steps improves the reward and score of the pretrained team due to a more informative reward initialization obtained from longer demonstration training. However, when the learning step exceeds 700, the pretrained agents tend to overfit to the operator's playing style, which limits further reinforcement learning adaptation in the 2v2 Pong environment. This behavior is reflected in Table 1 and Table 2, where the learning step of 700 yields the highest average reward and score, along with the largest standard deviation, compared to other learning step settings.

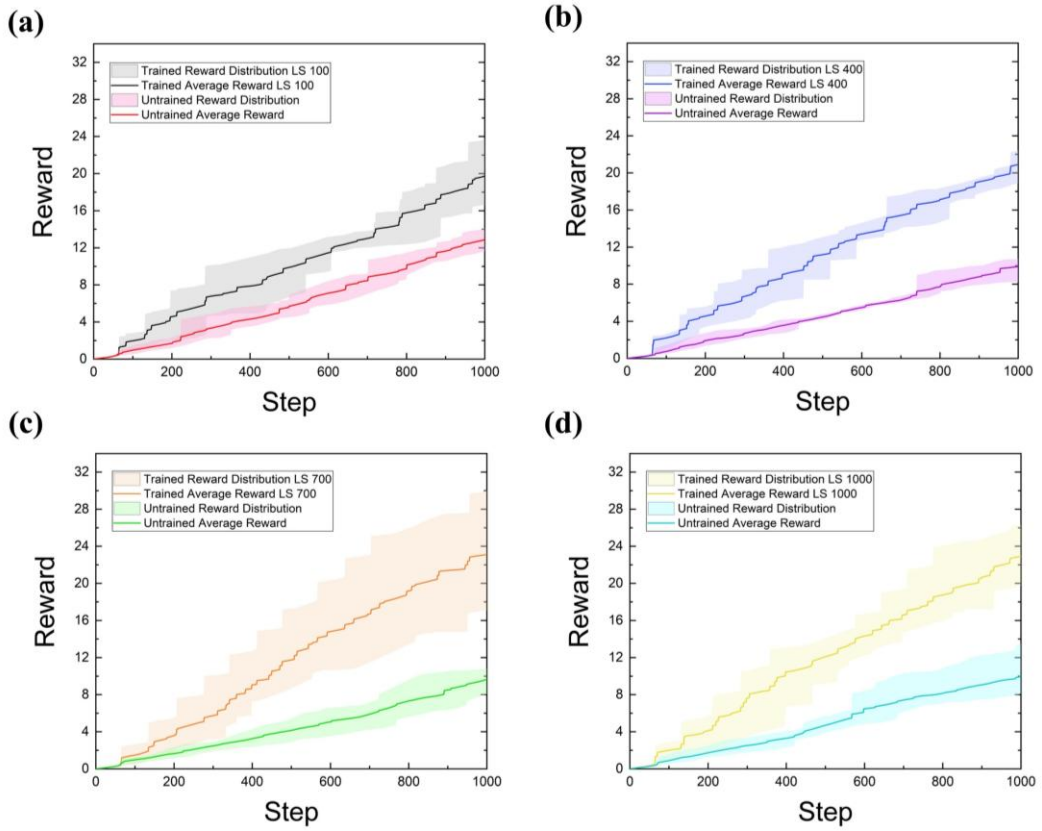

Figure 4. Pretrained vs untrained reward distribution of (a) 100, (b) 400, (c) 700, and (d) 1000 learning steps (LS)

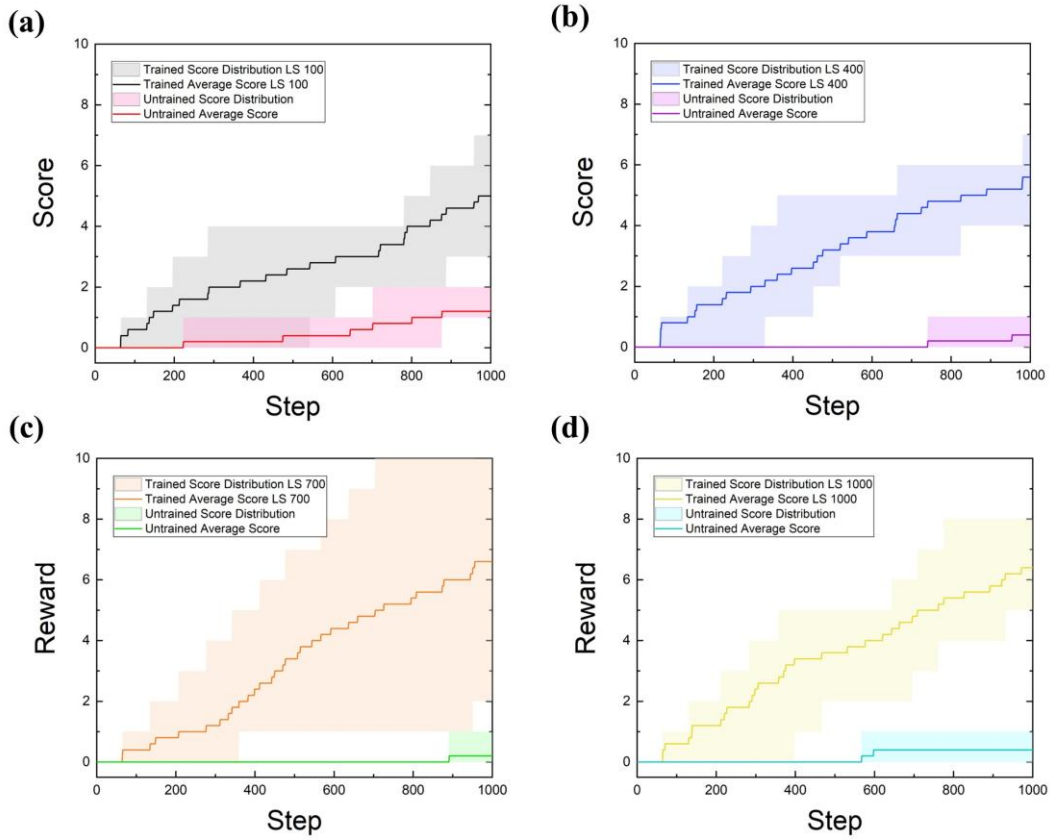

Figure 5. Pretrained vs untrained score distribution of (a) 100, (b) 400, (c) 700, and (d) 1000 learning steps

Figure 6 compares the proposed imitation learning method with the zonation method, which has been previously studied to improve learning efficiency in MARL [25]. The results show that the imitation learning method achieves superior performance in both reward accumulation and score compared to the zonation method when the two pretrained teams are evaluated against each other after 700 learning steps of pretraining for both methods. Although the zonation pretrained team performs better than the untrained team, its performance remains inferior to that of the imitation learning approach. These results indicate that imitation learning provides a more effective strategy for improving learning efficiency in collaborative MARL environments. To quantitatively compare learning efficiency, we conducted additional experiments to determine the number of learning steps required by the zonation method and by regular reinforcement learning with pretraining to achieve performance levels comparable to imitation learning with 700 learning steps. As shown in Figure 7, the zonation method requires approximately 2100 learning steps to reach average reward and score values similar to those obtained by imitation learning with only 700 learning steps. This result indicates that the imitation learning method achieves approximately 300% higher learning efficiency compared to the zonation method. Furthermore, for comparison with reinforcement learning without any pretraining, Figure 8 shows that the agents require approximately 4000 learning steps to achieve performance comparable to imitation learning with 700 learning steps. This finding suggests that imitation learning provides approximately 571% higher learning efficiency than reinforcement learning without a pretraining strategy. Detailed quantitative results for the zonation method and regular reinforcement learning with pretraining are summarized in Table 3.

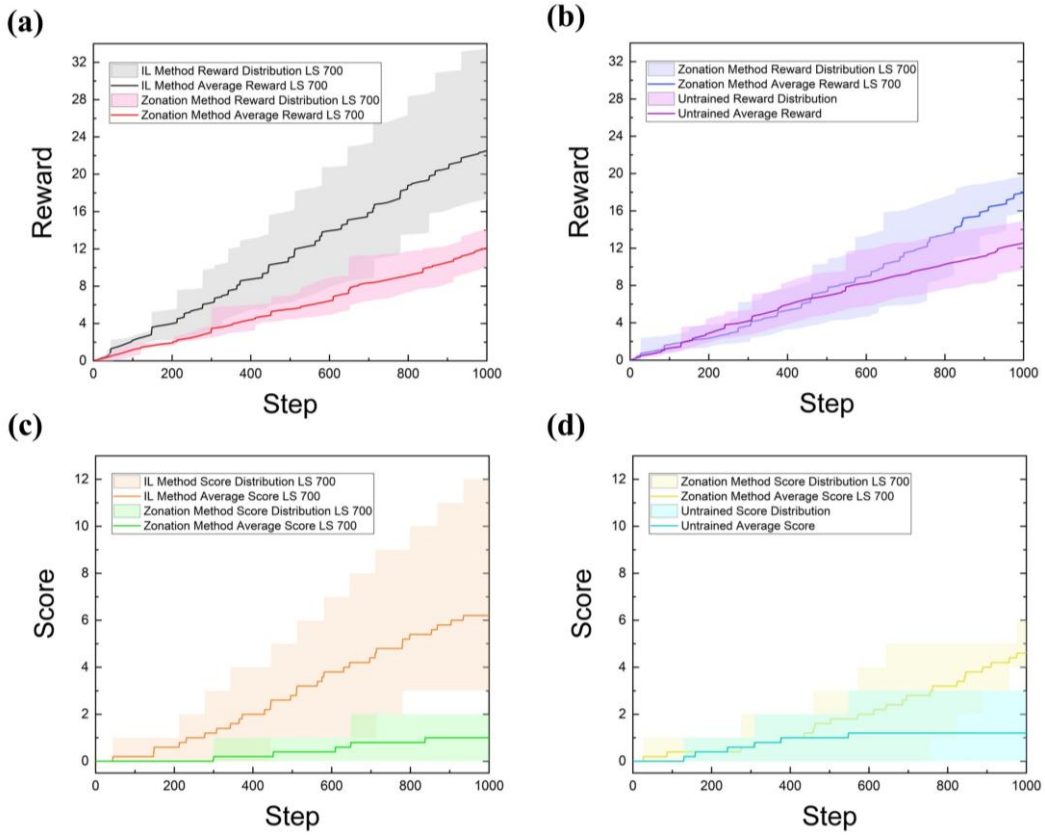

Figure 6. Performance Comparison between Imitation Learning (IL) and the Zonation Method with 700 Learning Steps. (a) Reward and (c) Score Distribution of IL vs the Zonation Method. (b) Reward and (d) Score Distribution of the Zonation Method vs the Untrained Agents

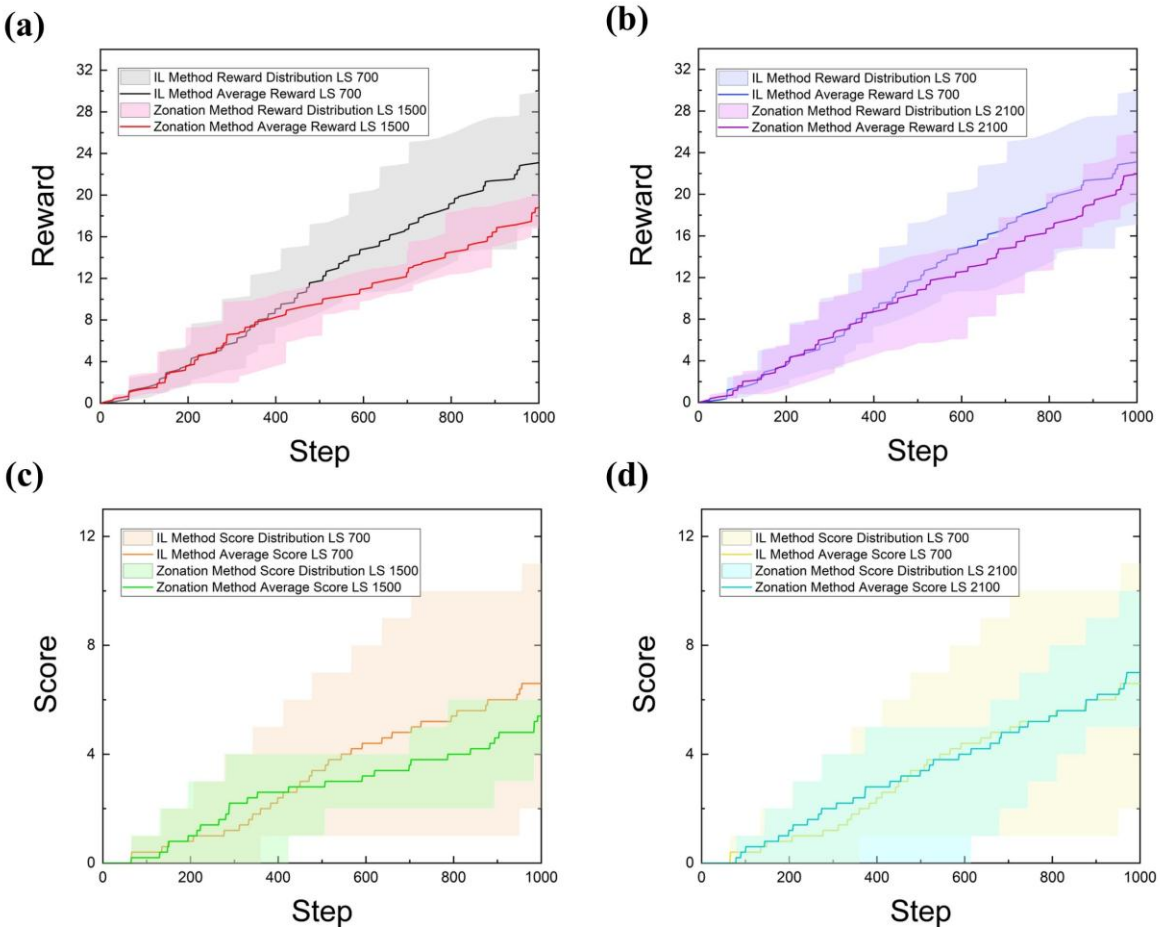

Figure 7. Performance Comparison between the IL Method with 700 Learning Steps and the Zonation Method with Different Learning Steps: (a) Reward and (c) Score Distribution of the Zonation Method with 1500 Learning Steps, and (b) Reward and (d) Score Distribution of the Zonation Method with 2100 Learning Steps

Table 1. Summary of Rewards Statistics for Pretrained and Untrained Teams

Steps	Pretrained			Untrained		
	Avg ± std	Min	Max	Avg ± std	Min	Max
100	19.73 ± 2.71	16.68	23.72	12.89 ± 1.12	11.63	14.10
400	20.90 ± 1.41	18.94	22.40	9.90 ± 1.10	8.26	10.73
700	23.12 ± 5.68	17.10	30.02	9.66 ± 1.03	8.10	10.82
1000	22.92 ± 2.96	19.79	26.31	10.06 ± 2.08	8.24	13.49

Table 2. Summary of Scores Statistics for Pretrained and Untrained Teams

Steps	Pretrained			Untrained		
	Avg ± std	Min	Max	Avg ± std	Min	Max
100	5.00 ± 1.58	3	7	1.20 ± 0.45	1	2
400	5.60 ± 1.14	4	7	0.40 ± 0.55	0	1
700	6.60 ± 3.65	2	11	0.20 ± 0.45	0	1
1000	6.40 ± 1.14	5	8	0.40 ± 0.55	0	1

Table 3. Summary of Learning Efficiency Across Different Methods

Methods	Learning Steps	Reward (avg ± std)	Score (avg ± std)	IL Relative Learning Efficiency
Imitation learning (IL) method	700	23.12 ± 5.68	6.60 ± 3.65	-
Zonation method	2100	21.98 ± 2.68	7 ± 1.87	300%
Regular reinforcement learning (pretrained)	4000	21.77 ± 0.67	7 ± 0.71	571%

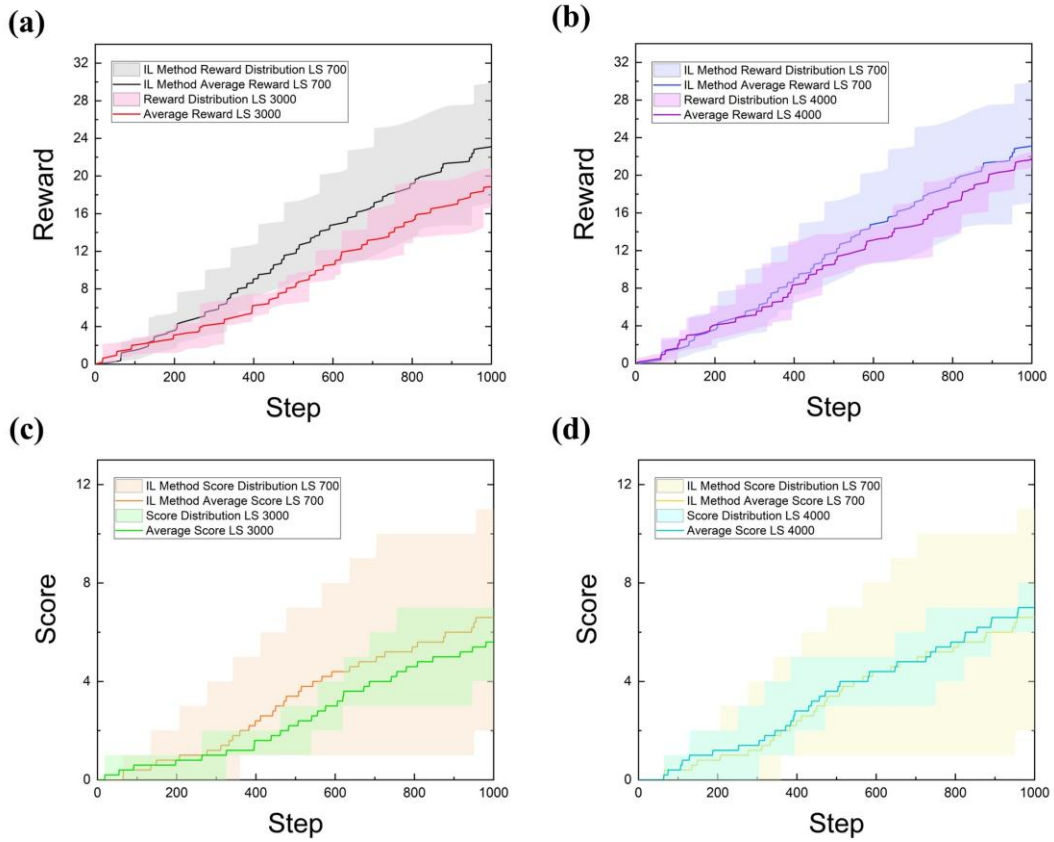

Figure 8. Performance Comparison between the IL Method with 700 Learning Steps and Regular Reinforcement Learning with Pretraining at Different Learning Steps: (a) Reward and (c) Score Distribution with 3000 Learning Steps, and (b) Reward and (d) Score Distribution with 4000 Learning Steps

4. Conclusion

In summary, this work demonstrates that imitation learning from demonstrations collected in a 1v1 pong game can effectively accelerate the training process of MARL in a 2v2 pong environment. By transferring an initial policy learned from demonstrations, the agents are able to reduce sample inefficiency caused by exploration and achieve faster performance improvement compared to training from scratch, even under different interaction dynamics. Based on experiments with 100, 400, 700, and 1000 demonstration learning steps, the reward and score of the MARL agents generally increase as the number of demonstration learning steps increases. However, the optimal performance in the 2v2 pong game is achieved at 700 learning steps. When the demonstration learning steps exceed this value, excessive memorization of the demonstration gameplay limits generalization and inhibits further reinforcement learning, resulting in reduced performance. In comparison with other learning booster strategies, the proposed imitation learning approach consistently outperforms the zonation method and standard reinforcement learning pretraining. Specifically, imitation learning with 700 learning steps achieves performance comparable to the zonation method with approximately 2100 learning steps and to regular reinforcement learning pretraining with approximately 4000 learning steps, corresponding to learning efficiency improvements of about 300% and 571%, respectively. These results indicate that imitation learning provides a more effective and data efficient mechanism for accelerating cooperative MARL training than existing baseline methods. These findings offer a practical approach for transferring knowledge from a simpler single agent setting to a cooperative MARL task and empirically demonstrate an improvement in learning efficiency. This approach is particularly beneficial in scenarios where data collection, computational resources, and training time are tightly constrained, as is often the case in robotics applications. Moreover, the results open new opportunities for advancing game-based MARL research by establishing a transferable learning framework across environments with different interaction dynamics.

Acknowledgement

The authors wish to express their gratitude to Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas Kristen Krida Wacana (UKRIDA) for providing financial support for this publication.

References

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Process Mag*, vol. 34, no. 6, pp. 26–38, Nov. 2017. https://doi.org/10.1109/MSP.2017.2743240

[2] R. Nian, J. Liu, and B. Huang, "A review On reinforcement learning: Introduction and applications in industrial process control," *Comput Chem Eng*, vol. 139, p. 106886, Aug. 2020. https://doi.org/10.1016/j.compchemeng.2020.106886

[3] S. S. Mousavi, M. Schukat, and E. Howley, "Deep Reinforcement Learning: An Overview," *arXiv*, Jun. 2018. https://doi.org/10.1007/978-3-319-56991-8_32

[4] J. Buckman, D. Hafner, G. Tucker, E. Brevdo, and H. Lee, "Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal: Curran Associates, Inc., 2018. Accessed: Jul. 02, 2025.

[5] J. Zhang, J. Kim, B. O'Donoghue, and S. Boyd, "Sample Efficient Reinforcement Learning with REINFORCE," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10887–10895, May 2021. https://doi.org/10.1609/aaai.v35i12.17300

[6] V. Kain *et al.*, "Sample-efficient reinforcement learning for CERN accelerator control," *Physical Review Accelerators and Beams*, vol. 23, no. 12, p. 124801, Dec. 2020, doi: 10.1103/PhysRevAccelBeams.23.124801.

[7] S. Raza, S. Haider, and M.-A. Williams, "Teaching coordinated strategies to soccer robots via imitation," in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, Dec. 2012, pp. 1434–1439. https://doi.org/10.1109/ROBIO.2012.6491170

[8] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation Learning: A Survey of Learning Methods," *ACM Comput Surv*, vol. 50, no. 2, pp. 1–35, Mar. 2018. https://doi.org/10.1145/3054912

[9] J. Ho and S. Ermon, "Generative Adversarial Imitation Learning," in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona: Curran Associates, Inc., 2016.

[10] B. Piot, M. Geist, and O. Pietquin, "Bridging the Gap Between Imitation Learning and Inverse Reinforcement Learning," *IEEE Trans Neural Netw Learn Syst*, vol. 28, no. 8, pp. 1814–1826, Aug. 2017. https://doi.org/10.1109/TNNLS.2016.2543000

[11] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligenc*, JMLR Workshop and Conference Proceedings, 2011, pp. 627–635. https://doi.org/10.48550/arXiv.1011.0686

[12] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, "A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges," *IEEE Trans Cybern*, vol. 54, no. 12, pp. 7173–7186, Dec. 2024. https://doi.org/10.1109/TCYB.2024.3395626

[13] T. Hester *et al.*, "Deep Q-learning From Demonstrations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. https://doi.org/10.1609/aaai.v32i1.11757

[14] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement Learning from Imperfect Demonstrations," in *Proceedings of the 35th International Conference on Machine Learning*, May 2019. https://doi.org/10.48550/arXiv.1802.05313

[15] T. Viet Bui, T. Mai, and T. Hong Nguyen, "Mimicking To Dominate: Imitation Learning Strategies for Success in Multiagent Competitive Games," in *38th Conference on Neural Information Processing Systems (NeurIPS 2024*, Aug. 2023. https://doi.org/10.48550/arXiv.2308.10188

[16] P. Brackett, S. Liu, and Y. Liu, "SC-MAIRL: Semi-Centralized Multi-Agent Imitation Reinforcement Learning," *IEEE Access*, vol. 11, pp. 57965–57976, 2023. https://doi.org/10.1109/ACCESS.2023.3282168

[17] Z. Li, Q. Ji, X. Ling, and Q. Liu, "A Comprehensive Review of Multi-Agent Reinforcement Learning in Video Games," *IEEE Trans Games*, pp. 1–21, 2025. https://doi.org/10.1109/TG.2025.3588809

[18] L. Le Mero, D. Yi, M. Dianati, and A. Mouzakitis, "A Survey on Imitation Learning Techniques for End-to-End Autonomous Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14128–14147, Sep. 2022. https://doi.org/10.1109/TITS.2022.3144867

[19] S. Li and W. Guo, "Supervised Reinforcement Learning for ULV Path Planning in Complex Warehouse Environment," *Wirel Commun Mob Comput*, vol. 2022, pp. 1–12, Oct. 2022. https://doi.org/10.1155/2022/4384954

[20] R. P. Bhattacharyya, D. J. Phillips, B. Wulfe, J. Morton, A. Kuefler, and M. J. Kochenderfer, "Multi-Agent Imitation Learning for Driving Simulation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Oct. 2018, pp. 1534–1539. https://doi.org/10.1109/IROS.2018.8593758

[21] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," *arXiv*, Apr. 2021. https://doi.org/10.48550/arXiv.1911.10635

[22] P. K. Sharma, E. G. Zaroukian, R. Fernandez, A. Basak, and D. E. Asher, "Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, Jul. 2021. https://doi.org/10.48550/arXiv.2107.14316

[23] J. L. Adler and V. J. Blue, "A cooperative multi-agent transportation management and route guidance system," *Transp Res Part C Emerg Technol*, vol. 10, no. 5–6, pp. 433–454, Oct. 2002. https://doi.org/10.1016/S0968-090X(02)00030-X

[24] Y. Zhou *et al.*, "Is Centralized Training with Decentralized Execution Framework Centralized Enough for MARL?," arXiv, May 2025. https://doi.org/10.48550/arXiv.2305.17352

[25] M. Y. Hadiyanto, B. Harsono, and I. Karnadi, "Zonation Method for Efficient Training of Collaborative Multi-Agent Reinforcement Learning in Double Snake Game," *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, p. 02401011, Dec. 2023. https://doi.org/10.26877/asset.v6i1.17562