Bidirectional Distillation: A Mixed-Play Framework for Multi-Agent Generalizable Behaviors

Extended Abstract

Lang Feng*
Zhejiang University
Hangzhou, China
langfeng@zju.edu.cn

Jiahao Lin*
Zhejiang University
Hangzhou, China
22221159@zju.edu.cn

Dong Xing
Zhejiang University
Hangzhou, China
dongxing@zju.edu.cn

Li Zhang
Zhejiang University
Hangzhou, China
zhangli85@zju.edu.cn

De Ma
Zhejiang University
Hangzhou, China
made@zju.edu.cn

Gang Pan†
Zhejiang University
Hangzhou, China
gpan@zju.edu.cn

ABSTRACT

Population-population generalization is a challenging problem in multi-agent reinforcement learning (MARL), particularly when agents encounter unseen co-players. However, existing self-play-based methods are constrained by the limitation of inside-space generalization. In this study, we propose Bidirectional Distillation (BiDist), a novel mixed-play framework, to overcome this limitation in MARL. BiDist leverages knowledge distillation in two alternating directions: forward distillation, which emulates the historical policies' space and creates an implicit self-play, and reverse distillation, which systematically drives agents towards novel distributions outside the known policy space in a non-self-play manner. Our results highlight its remarkable generalization ability across a variety of cooperative, competitive, and social dilemma tasks, and reveal that BiDist significantly diversifies the policy distribution space.

KEYWORDS

Multi-agent reinforcement learning, population-population generalization, unseen co-players, mixed-play framework

ACM Reference Format:
Lang Feng, Jiahao Lin, Dong Xing, Li Zhang, De Ma, and Gang Pan. 2025. Bidirectional Distillation: A Mixed-Play Framework for Multi-Agent Generalizable Behaviors: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION & RELATED WORK

Multi-agent reinforcement learning (MARL) [9, 13, 18, 21] still struggles with generalization [4] like zero-shot co-player generalization, where agents trained together must maintain performance when some are replaced with unseen ones [1, 8]. This paper focuses on population-population generalization in MARL, where multiple agents interact with unseen co-players. This scenario presents a greater challenge compared to ad-hoc teamplay [16, 19, 20], which requires only a single agent to generalize with multiple partners

Standard MARL fosters coordination only among agents' *current* policies during training, limiting interaction diversity. Self-play [2, 5, 11, 12, 15] mitigates this by leveraging historical policies, ideally being able to cover the entire distribution space of both *current* and *historical* training policies. While this improves "inside-space" generalization, it struggles with "outside-space" generalization where zero-shot policies include those never encountered during training. These distributions cannot be captured merely by reusing historical training policies, and in such cases, self-play-based methods fall short.

In this study, we propose a novel mixed-play method that enriches interaction diversity by strategically deviating from historical policy space, especially in the outside space. BiDist alternates between two phases: forward distillation and reverse distillation. The forward phase distills fictitious population policies from historical ones, simulating implicit self-play without costly memorization. The reverse phase, by contrast, pushes fictitious policies away from historical patterns, promoting non-self-play exploration. Empirical evaluations across cooperation, competition, and social dilemma games show BiDist outperforms baselines, generalizes well, and expands policy distribution.

2 BIDIRECTIONAL DISTILLATION

We introduce Bidirectional Distillation (BiDist) to address inside-space and outside-space generalization in MARL. BiDist alternates between forward distillation, which retains self-play knowledge, and reverse distillation, which pushes beyond historical boundaries to explore new behaviors. This approach systematically enhances agent generalization beyond conventional self-play methods.

Fictitious population. To encourage diverse agent interactions, we introduce a *fictitious* population $\tilde{g}$ as an *imaginary background population* g during training. It consists of agents randomly detached from the trained population f. The assignment is determined by a binary vector $\boldsymbol{v} = v_1, \ldots, v_N \in \{0, 1\}^N$, where $v_i = 1$ denotes agent in the trained population and $v_i = 0$ in the fictitious population. The vector $\boldsymbol{v}$ is sampled from the Bernoulli distribution. Agents in $\tilde{g}$ use distilled policies $\{\pi_{\phi_i}\}_{i=1}^N$ instead of learning

*Equal contribution.
†Corresponding author.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Table 1: The normalized focal per-capita returns of different algorithms on the testing scenarios for each substrate. SC denotes the scenario. The returns are min-max normalized within each scenario.

	Pure Coordination					Coop Mining					Chicken					Coins					Prisoners Dilemma				
Method	SC0	SC1	SC2	SC3	SC4	SC0	SC1	SC2	SC3	SC4	SC0	SC1	SC2	SC3	SC4	SC0	SC1	SC2	SC3	SC4	SC0	SC1	SC2	SC3	SC4
MAPPO	0.50	0.61	0.41	0.67	0.32	0.40	0.50	0.00	0.61	0.79	0.66	0.60	0.72	0.71	0.71	0.97	0.87	0.87	0.97	0.73	0.64	0.60	0.75	0.88	0.89
RanNet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
OPRE	0.40	0.58	0.26	0.86	0.03	0.52	0.74	0.93	0.78	0.72	0.54	0.62	0.61	0.72	0.72	0.93	0.92	0.75	**1.00**	**1.00**	0.00	0.27	0.14	0.20	0.29
PP	0.65	0.43	0.81	0.72	0.64	0.25	0.26	0.22	0.18	0.04	0.47	0.43	0.36	0.53	0.59	0.86	0.64	0.76	0.21	0.75	0.66	0.92	0.63	0.77	0.52
RPM	0.77	0.65	0.79	0.80	0.71	0.18	0.29	0.37	0.02	0.05	0.52	0.52	0.48	0.48	0.49	0.88	0.81	0.72	0.09	0.80	0.80	**1.00**	0.79	0.56	0.68
BiDist	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.97	0.99	**1.00**	0.96	**1.00**	**1.00**	**1.00**

policies $\{\pi_{\theta_i}\}_{i=1}^{N}$, gathering trajectories as:

$$\tau \sim \texttt{GatherTrajectories}(\mathcal{Z}, \underbrace{\{\pi_{\theta_i}\}_{v_i=1}}_{\text{Population } f}, \underbrace{\{\pi_{\phi_i}\}_{v_i=0}}_{\text{Population } \tilde{g}}), \quad (1)$$

where $\mathcal{Z}$ denotes the environment and τ denotes the trajectory.

Forward distillation. We employ knowledge distillation [3] to implicitly retain historical learning policies while minimizing resource consumption. This involves using the distilled policy networks $\{\pi_{\phi_i}\}_{i=1}^{N}$ to approximate the distributions of the learning policy networks $\{\pi_{\theta_i}\}_{i=1}^{N}$ throughout training

$$\mathcal{L}^{\mathrm{KL}}(\boldsymbol{\phi}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{o_i\sim\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi_{\theta_i}(\cdot|o_i)||\pi_{\phi_i}(\cdot|o_i)\right)\right], \quad (2)$$

with parameter updates:$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \eta_f \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}^{\mathrm{KL}}(\boldsymbol{\phi})$. Distillation occurs periodically at an interval k_d, maintaining a lagged representation of learning policies.

Reverse distillation. To promote exploration, reverse distillation shifts preferences in $\tilde{g}$ beyond historical distributions. This is achieved by maximizing the KL divergence: $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \eta_r \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}^{\mathrm{KL}}(\boldsymbol{\phi})$. By doing so, the distilled policies can capture action preferences that differ from the learning policies, enabling them to explore and generalize beyond the inside-space distribution. Throughout the training, forward and reverse distillations alternate at intervals of k_d iterations. This dynamic interplay balances retention and innovation, ensuring that not only inherit the strengths of self-play but also adapt to novel outside-space scenarios.

3 EXPERIMENT

In our experiment, we carry out 5 different generalization tasks across cooperation, competition, and social dilemmas, on DeepMind's Melting Pot [1]. Our baseline methods include MAPPO [21], a multi-agent version of the PPO series [10, 14], as well as RandNet [7], OPRE [17], population-based self-play (PP), and RPM [12], a self-play MARL generalization approach. The results are presented in Table 1. We report the min-max normalized focal per-capita returns of five test scenarios (scenarios 0-4) for each substrate.

Main results. We observe that RanNet struggles to achieve population-population generalization. In comparison, OPRE, PP, and RPM achieve competitive outcomes in certain tasks. However, their generalization ability displays discontinuity in different scenarios. For example, RPM ranges from a peak of 1.00 in the prisoner's dilemma to a low of 0.56, whereas OPRE varies from 0.86 in pure coordination scenarios to a mere 0.03 at its weakest. These inconsistencies highlight their difficulties in effectively incorporating

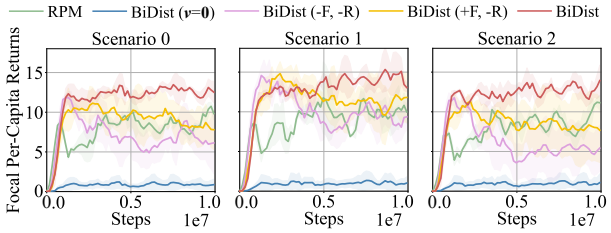

Figure 1: Ablation studies of BiDist on *Pure Coordination*.

preference shifts to address the outside-space generalization. Consequently, while their performance may shine in certain scenarios, they falter in others where agents hold disparate preferences. In contrast, our BiDist consistently achieves high normalized focal per-capita returns across varying dynamics. By integrating preference shifts into the training data, BiDist more effectively addresses both inside-space and outside-space challenges.

Ablations. Figure 1 examines the ablations of BiDist by setting $\boldsymbol{v} = \mathbf{0}$ (BiDist ($\boldsymbol{v} = \mathbf{0}$)), removing reverse distillation (BiDist (+F, -R)) and then further removing forward distillation (BiDist (-F, -R)). The results show a significant drop in BiDist ($\boldsymbol{v} = \mathbf{0}$), highlighting the importance of fictitious agents. Both BiDist (+F, -R) and BiDist (-F, -R) perform poorly, emphasizing the need for both forward and reverse distillations.

4 CONCLUSION & FUTURE WORK

In this paper, we addressed the challenge of population-population generalization in MARL and proposed a concise and effective approach, called BiDist. Central to our contribution is the formulation of a mixed-play framework that leverages the power of diversity in agent interactions. BiDist consists of two alternating phases: forward and reverse distillations, which work together to effectively achieve historical knowledge retention and preference shifts for the fictitious population, thereby enhancing the diversity in agent interactions. One direction for future work lies in long-term knowledge retention and adaptation, for example, through continuous learning [6].

ACKNOWLEDGMENT

This work was supported in part by the STI 2030 Major Projects under Grant 2021ZD0200400, in part by the National Natural Science Foundation of China (61925603, 62376247, U20A20220, and 62334014), and in part by the grants from Key R&D Program of Zhejiang (2022C01048).

REFERENCES

[1] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. 2022. Melting Pot 2.0. *arXiv preprint arXiv:2211.13746* (2022).

[2] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In *International conference on machine learning*. PMLR, 805–813.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[4] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research* 67 (2020), 757–795.

[5] Yuhua Jiang, Qihan Liu, Xiaoteng Ma, Chenghao Li, Yiqin Yang, Jun Yang, Bin Liang, and Qianchuan Zhao. 2024. Learning Diverse Risk Preferences in Population-Based Self-Play. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 12910–12918.

[6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[7] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. 2019. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396* (2019).

[8] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*. PMLR, 6187–6199.

[9] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems* 30 (2017).

[10] Wenjia Meng, Qian Zheng, Gang Pan, and Yilong Yin. 2023. Off-policy proximal policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9162–9170.

[11] Dung Nguyen, Hung Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. 2024. Diversifying Training Pool Predictability for Zero-shot Coordination: A Theory of Mind Approach. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 166–174.

[12] Wei Qiu, Xiao Ma, Bo An, Svetlana Obraztsova, YAN Shuicheng, and Zhongwen Xu. 2023. RPM: Generalizable Multi-Agent Policies for Multi-Agent Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*.

[13] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4295–4304.

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

[15] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[16] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 1504–1509.

[17] Alexander Vezhnevets, Yuhuai Wu, Maria Eckstein, Rémi Leblond, and Joel Z Leibo. 2020. Options as responses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 9733–9742.

[18] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *9th International Conference on Learning Representations*. OpenReview.net.

[19] Dong Xing, Pengjie Gu, Qian Zheng, Xinrun Wang, Shanqi Liu, Longtao Zheng, Bo An, and Gang Pan. 2023. Controlling type confounding in ad hoc teamwork with instance-wise teammate feedback rectification. In *International Conference on Machine Learning*. PMLR, 38272–38285.

[20] Dong Xing, Qianhui Liu, Qian Zheng, Gang Pan, and Z Zhou. 2021. Learning with Generated Teammates to Achieve Type-Free Ad-Hoc Teamwork.. In *IJCAI*. 472–478.

[21] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2021. The surprising effectiveness of ppo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955* (2021).