Entropy-Augmented Multi-Objective Policy Optimization in Multiagent Systems

Jamie Santos[1] , **Ayhan Alp Aydeniz**[1] , **Raghav Thakar**[1] and **Kagan Tumer**[1]
[1]The Collaborative Robotics and Intelligent Systems (CoRIS) Institute
Oregon State University, Corvallis, Oregon, USA
{santjami, aydeniza, thakarr, kagan.tumer}@oregonstate.edu

Abstract

Autonomous agent teams deployed in settings such as marine and extraterrestrial outposts must coordinate actions to achieve optimal outcomes across multiple competing objectives. Multi-objective evolutionary algorithms such as NSGA-II optimize for diversity in the objective space, but neglect diversity in the behavior space, possibly leading to premature convergence and a collapse in behaviors that may differentiate policies in different external conditions. To address this, we introduce an entropy-augmented policy evaluation strategy that incorporates an entropy bonus into agent fitness scores, discouraging behavioral homogeneity across the evolving population. By augmenting policy evaluation with a behavior-space diversity signal while preserving the underlying Pareto optimization framework, our method is designed to encourage exploration of behaviorally distinct policies in multiagent domains. We evaluate our approach across rover-domain experiments with qualitatively distinct reward structures and observe hypervolume improvements of up to 48% relative to the NSGA-II baseline, suggesting that behavioral diversity is a promising and underexplored direction for improving multi-objective multiagent evolutionary optimization.

1 Introduction

Robot teams have been increasingly deployed in settings such as marine monitoring, space exploration, and disaster response. These teams must often balance competing objectives such as energy usage, task completion, and team safety. Since a single policy rarely optimizes all objectives simultaneously, as these are often in conflict, learning a set of Pareto-optimal team policies is frequently necessary. Generating a robust set of such policies that provide optimal trade-offs across objectives is therefore crucial for supporting a variety of operator preferences and evolving mission conditions.

Multi-Objective Evolutionary Algorithms (MOEAs) such as Strength Pareto Evolutionary Algorithm 2 (SPEA2), Multi-Objective Evolutionary Algorithm Based on Decomposition (MOEA/D), and Non-Dominated Sorting Algorithm II (NSGA-II) optimize sets of Pareto-optimal control policies by maintaining diversity across candidate solutions in the objective space during population-based searches [Zitzler *et al.*, 2001; Zhang and Li, 2007; Deb *et al.*, 2002]. Diversity is typically enforced through mechanisms such as crowding distance, which preferentially retains solutions that are more spread out in the objective space. Multi-Objective Quality-Diversity (MOQD) approaches extend this idea by optimizing a grid of Pareto fronts across predefined behavioral niches, explicitly accounting for behavioral diversity alongside objective performance [Pierrot *et al.*, 2022; Janmohamed and Cully, 2025].

However, these approaches do not integrate behavior-based diversity into the search for the global optimal Pareto front. As a result, behaviorally distinct policies that yield similar objective scores early in optimization may be prematurely discarded. This risk is especially acute under sparse rewards, where behavioral differences are not yet reflected in objective values. For example, rovers learning to visit Points of Interest (POIs) may receive identical rewards whether most of the team sits idle near one POI, or whether most of the team wanders near other POIs, yet the latter strategy may be closer to achieving higher rewards in fewer generations. In Quality-Diversity (QD) approaches, behavior is more explicitly accounted for by optimizing Pareto fronts across behavioral niches, but localized competition can suppress policies that can otherwise emerge as globally dominant.

In this work, we introduce a behavior-space-aware fitness-shaping strategy that incorporates behavioral diversity into global multi-objective policy optimization for teams of agents, built on the widely adopted NSGA-II algorithm. Our key idea is to augment the rollout rewards for each objective with an entropy bonus, measured as the estimated average k-Nearest Neighbors (kNN) entropy of joint agent positions across a policy's trajectory. Rather than treating entropy as a separate objective, it is added directly to each objective's reward and scaled by a tunable parameter β, such that higher values of β favor policies that promote broader exploration during evolution. This approach preserves the global Pareto structure while encouraging behavioral diversity throughout the evolutionary search.

We evaluate our approach in the multi-rover domain across experiments with qualitatively distinct reward structures: an obstacle navigation task requiring agent coordination to ob-

tain higher-valued rewards, and a temporal task in which agents must trade off between near-term and long-term rewards. In these experiments, entropy augmentation yielded an average of 18% higher hypervolume in Experiment 1 and 48% higher hypervolume in Experiment 2 compared to the NSGA-II baseline. These findings provide preliminary evidence that incorporating behavioral diversity into fitness evaluation can improve multi-objective optimization performance in some setting and motivate further investigation of behavior-aware evolutionary search. While the presented results are encouraging, additional evaluation over more trials and a broader range of domains is necessary to determine robustness and generality of the proposed strategy.

2 Background and Related Work

2.1 Multiagent, Multi-Objective Learning

Many tasks, such as cooperative ground exploration and mapping and multi-robot search and rescue are naturally multi-agent, multi-objective tasks. For example, autonomous extraterrestrial rover teams may need to work together to efficiently cover ground for exploration while minimizing completion time and energy consumption [Nayak *et al.*, 2025]. In this setting, 2+ agents are tasked with cooperating in a common environment to achieve several shared goals. Furthermore, agents are expected to act simultaneously and independently, and may not be privileged with global information [Yamauchi, 1998]. An example of where this may occur is in underwater robotics where the properties of water make communication between robots especially difficult [Heidemann *et al.*, 2006].

Formally, the multi-objective multiagent setting is modeled as a Multi-Objective Decentralized Partially Observable Markov Decision Process (MODec-POMDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \vec{r}, T)$ [Felten *et al.*, 2024]. In the MODec-POMDP, $\mathcal{S}$ represents the state space, $\mathcal{A} = a_1 \times a_2 \times ... \times a_i$ the joint action space across all I agents, and T the state transition probability function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. The vector $\vec{r}$ is the immediate reward vector of dimension n, the number of objectives. Actions are taken according to each agent i's local observation space, $o_i \in O$, and a decentralized policy, $\pi(a_i|o_i)$, in an effort to maximize the total accumulated reward vector over the episode horizon [Roijers *et al.*, 2013].

Because agents cannot access global state information during deployment, a common approach to solving multi-agent MDPs is to use a Centralized Training, Decentralized Execution (CTDE) framework. Using CTDE in team-based scenarios mitigates issues that arise from relying on shared intelligence during deployment, such as when individual robots fail [Yamauchi, 1998].

The team is also faced with a dilemma; optimizing returns on one objective may come at the cost of diminishing the outcomes of other objectives. Therefore, there may not be a single optimal solution. Instead, the optimal set of policies in which a single policy cannot be considered objectively better than another, also known as the Pareto front, is the final output [Zitzler *et al.*, 2004]. In the multiagent setting, a single solution along the Pareto front often correlates to a joint-team control policy, such as a set of weights for agents' neural network controllers [Thakar *et al.*, 2025]. Points along the Pareto front in these cases therefore represent tradeoffs between different controllers, where points along different objective axes represent policies that specialize in achieving specific objectives.

2.2 Evolutionary Approaches to Multi-Objective Learning

A classic approach to multi-objective optimization problems is to scalarize the reward by taking a weighted sum of the reward vector $\vec{r}$ (or expected value vector $\vec{V}^{\pi}$ in reinforcement learning approaches):

$$\tilde{r}_t = \vec{w} \cdot \vec{r}_t = \sum_{i=1}^{n} w_i r_{t,i} \tag{1}$$

[Van Moffaert *et al.*, 2014]. While simply hand-crafting a vector of importance weights for each objective is sometimes an effective strategy, doing so assumes in-depth knowledge of the relative importance of objectives and that a single fixed tradeoff is sufficient [Tan *et al.*, 2005]. Furthermore, weighted-sum methods fail to capture the full portfolio of possible solutions in non-convex problems [Roijers *et al.*, 2013; Pappas *et al.*, 2021].

Evolutionary Algorithms (EAs) evolve a population of solutions to an objective function. By extension, MOEAs are an established method of optimizing and maintaining a set of solutions over a multi-dimensional objective space. They aim to minimize functions of the form

$$F(x) = (f_1(x), \ldots, f_m(x))^T \quad \text{s.t.} \quad x \in \Omega$$

in which Ω represents the decision space, x a decision vector, m the number of objective functions $f_i : \Omega \rightarrow R$, and R^m the objective space [Zhou *et al.*, 2011]. The set of non-dominated solutions is what is known as a Pareto front; a solution x is said to dominate solution y when

$$\forall i \in \{1, \ldots, m\},\ f_i(x) \geq f_i(y)$$

and

$$\exists j \in \{1, \ldots, m\} \text{ such that } f_j(x) > f_j(y),$$

i.e., x performs at least as well as y across all objectives, and there exists at least one objective in which x outperforms y [Hu *et al.*, 2023; Zhou *et al.*, 2011].

Strength-based methods such as SPEA2 seek to preserve an even distribution of solutions in the objective space by ranking solutions according to how many others they dominate [Zitzler *et al.*, 2001]. Objective-space diversity is maintained via kNN distance. In contrast, dominance-based methods such as NSGA-II seek to preserve an even spread of solutions across the Pareto front [Deb *et al.*, 2002]. NSGA-II ranks solutions primarily by Pareto dominance, and then by crowding distance, in each evolution generation.

However, as the dimensionality of the objective space increases, i.e., 4 or more objectives, solutions become sparser and crowding distance becomes a less relevant strategy for ranking solutions [Yuan *et al.*, 2014]. Reference-based methods such as NSGA-III address this issue by selecting non-dominated solutions nearest to supplied reference vectors

[Deb and Jain, 2013]. Yet, this performance comes at the cost of additional computational overhead. Therefore, for this work, NSGA-II was selected as the baseline in order to demonstrate improved Pareto optimization in lower dimensions.

These methods are primarily designed to optimize diversity in the objective space, but neglect to directly consider the behavior space [Lehman and Stanley, 2011b]. However, evolutionary algorithms themselves are agnostic to what exactly they are evolving, and modifications to the fitness function(s) enable behavioral diversity in solutions, as well. Quality-diversity methods such as Multi-Objective Map-Elites (MOME) aim to address behavioral diversity by optimizing Pareto fronts across a grid of niches, producing an array of behaviorally distinct Pareto fronts [Pierrot *et al.*, 2022]. While this method explicitly accounts for behavioral diversity, competition occurs primarily within behavioral niches rather than across a single global Pareto front. Furthermore, while MAP-Elites has been extended to both multi-objective [Pierrot *et al.*, 2022] and multiagent [Ingvarsson *et al.*, 2023] settings, combining these ideas remains relatively unexplored. This paper introduces a method of adapting MOEAs, namely NSGA-II, to consider behavioral diversity as well as objective-space diversity when optimizing the global Pareto front in multiagent settings.

2.3 Entropy-Driven Behavioral Diversity

Behavioral diversity has been explored in evolutionary and reinforcement learning as a means to encourage exploration to mitigate premature convergence. Lehman et al. demonstrate that the objective function itself may misguide the policy optimization process, and that searching for behavioral novelty instead can outperform objective-based optimization [Lehman and Stanley, 2011a]. This idea has been extended to multiagent settings; novelty search has demonstrated that rewarding behavioral diversity can encourage exploration and higher team rewards in single-objective cooperative tasks [Aydeniz *et al.*, 2023].

In population-based search settings, entropy is a well-established measure of diversity [Haarnoja *et al.*, 2018]. Maximum-entropy Reinforcement Learning (RL) methods incorporate entropy directly into the objective function:

$$J(\pi) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(s_t,a_t)\sim\rho^\pi} \left[r(s_t, a_t) + \beta \mathcal{H}(\pi(\cdot|s_t)) \right], \quad (2)$$

where s_t represents the states at time t, a the action, r the reward signal, H the expected entropy of the policy, and β the temperature parameter. H is calculated as

$$H(\pi(\cdot|s_t)) = \mathbb{E}_{a_t\sim\pi(\cdot|s_t)} \left[-\log \pi(a_t|s_t) \right] \quad (3)$$

[Dong *et al.*, 2025].

Entropy can also be estimated over agent trajectories to quantify behavioral diversity across a population. A common metric used is kNN entropy, which approximates the entropy of a continuous distribution using a finite set of samples [Kozachenko, 1987]. Analogously, the maximum entropy approach incorporates entropy directly into the objective function; this work proposes introducing entropy directly to the fitness scores of candidate policies in evolutionary learning methods. This work investigates augmenting evolutionary fitness scores using entropy estimates over agent trajectories. By incorporating behavioral diversity directly into the fitness evaluation, the proposed approach seeks to differentiate behaviorally distinct policies without introducing behavioral diversity as an additional optimization objective.

3 Method

In this section we discuss the problem formulation in the multiagent, multi-objective setting and the incorporation of the entropy metric into the fitness evaluation for the NSGA-II algorithm.

3.1 Environment: Multi-Agent Rover Domain

The Rover Domain is a robot-based simulation benchmark for evaluating multi-agent policies under a shared global reward [Tumer *et al.*, 2002; Aydeniz *et al.*, 2023; Thakar *et al.*, 2025]. In this environment, teams of agents are tasked with visiting POIs and are rewarded according to the task's specific conditions. Objectives can be defined according to POI type, or through other metrics such as time or distance covered. Each rover is equipped with sensors that enable each individual to gather information about its local environment. However, agents cannot communicate with each other; this constraint enables testing of policies developed using the CTDE paradigm.

Under these conditions, the Rover Domain can be described as a MODec-POMDP [Chatterjee *et al.*, 2006; Thakar *et al.*, 2025]. The global state consists of the joint coordinate positions of the rover team. For every time step, each rover takes an action based on its local observations of the distance of POIs and nearby rovers, and its current policy. An agent's action corresponds to a movement in the x and y directions up to a specified step maximum and to environmental dynamics. Activation of a POI depends on (a) the number of rovers visiting it simultaneously, i.e., the coupling factor, and (b) the length of time they have been there. The global reward at each time step then depends on which POIs have been successfully activated. POIs may present different reward values depending on the experimental setup.

3.2 Evolutionary Policy Optimization

The NSGA-II algorithm was selected for its relevance to multi-objective optimization and its established use in evolutionary policy search [Ma *et al.*, 2023]. This makes it a natural baseline for evaluating the effects of entropy-influenced policy evaluation relative to conventional objective-space methods.

The policies evolved using NSGA-II are the weights and biases of neural network controllers for each agent. The parameters used are listed in Table 1. Per the standard baseline NSGA-II algorithm, policies are evaluated primarily on their fitness as determined by the rewards gained for each objective in the Rover Domain. NSGA-II primarily preserves policies with the highest ranking, i.e., policies with fitnesses that are not dominated by any other policy's fitness on a single objective, and secondarily by crowding distance in the objective space.

Table 1: Hyperparameters used in NSGA-II algorithm

Hyperparameter	Value
Population Size	75
Generations	10,000
Policy Hidden Layers	[16, 16]
Mutation Rate	0.75
Mutation Scale	0.5
Weight Initialization Limit	0.2
Bias Initialization Limit	0.2
Entropy Neighbors (k)	5
Entropy Scale Factor (β)	$\{0.0, 0.05, 0.1, 0.25, 0.5\}$

However, Pareto ranking cannot decipher if policies have converged, or distinguish between policies that produce largely stagnate behavior and ones where agents are nearer to achieving higher rewards. Consequently, in a sparse reward environment, policies that achieve similar reward vectors but drastically different behaviors will be selected indiscriminately. Thus, the following section explains the novel introduction of entropy-influenced policy evaluation for multi-objective, multiagent joint policies.

3.3 KNN Entropy-Augmented Policy Evaluation

To prevent direct competition of task objectives with entropy, this work augments objective fitness scores with a scaled bonus of the measured entropy rather than introduce entropy as a competing objective. This method is intended to preserve the existing Pareto structure while promoting the discovery of more tradeoff solutions along the Pareto front.

A proxy for entropy, $\hat{H}$, is approximated using a kNN-based estimator [Pappas *et al.*, 2021] computed over joint trajectories within an episode:

$$\hat{H} = \frac{1}{N} \sum_{n=1}^{N} \log \epsilon_n$$

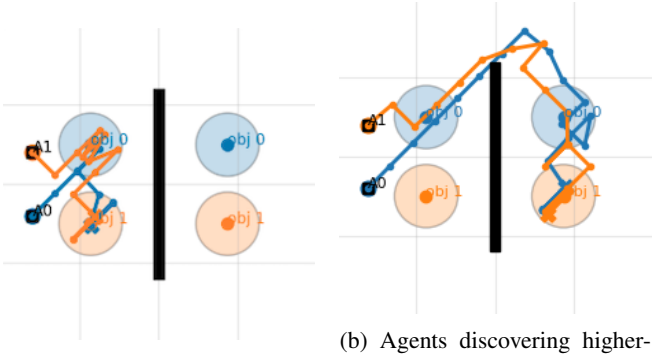
where ϵ_n denotes the distance from sample n to its kth nearest neighbor in the joint state space, and N denotes the length of the episode. The augmented fitness score for any policy, π, of an objective, i, then becomes the approximated kNN entropy, $\hat{H}$ for the joint trajectories obtained using that policy, scaled by a factor, β, and added to the original fitness score, f:

$$\tilde{f}_i(\pi) = f_i(\pi) + \beta \cdot \hat{H}(\pi), \quad \forall i \in 1, \ldots, m$$

By increasing β, policies that tend to exhibit broader exploration of the state space become more likely to be preserved during evolution. However, when β becomes too large, it can distort the true objective function. Ideally, β values should be large enough to differentiate between similarly performing policies, without overshadowing true performance across objective dimensions the algorithm is optimizing for.

3.4 Performance Metrics

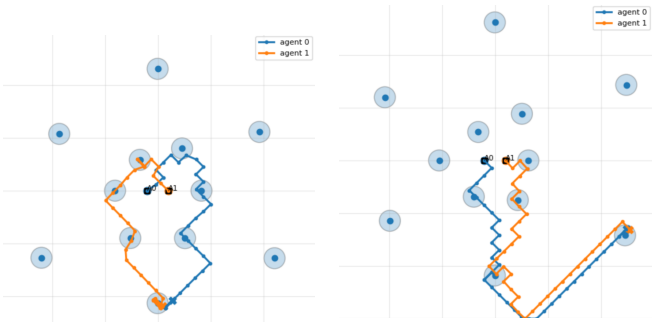
We seek to determine whether the addition of entropy as a secondary policy evaluation criterion can increase the overall hypervolume beneath the Pareto front. Hypervolume is computed after each evolutionary generation to characterize convergence and compare the Pareto fronts obtained under different values of β.



(a) Agents favoring locally safe objectives

(b) Agents discovering higher-value tradeoff solutions through exploration

Figure 1: Example agent trajectories: (a) agents easily discover low value, repeatable POIs and tend to adhere to this safe region; (b) agents may eventually discover higher value POIs and therefore higher rewards on one or both objectives by accepting low rewards early on in the episode



(a) Agents optimizing for easy, fast rewards (inner ring)

(b) Agents trading delayed rewards for higher value POIs (outer ring)

Figure 2: Example agent trajectories from the time experiment: (a) agents collect one-time rewards from close, 1-point POIs before moving onto higher reward, 3-point POIs; (b) agents collect close POIs on the way to farther, higher reward POIs

4 Experiments

In this section, we introduce the experiment setup in which the standard NSGA-II implementation is compared to the introduction of entropy-based bonuses to the policy evaluation procedure.

4.1 Experiment Setups

The purpose of this work is to investigate whether incorporating behavioral diversity into policy evaluation influences the diversity and quality of tradeoff solutions for teams of agents in a multi-objective setting.

Experiment 1: Obstacle Navigation

In this experiment, a two-rover team was tasked with maximizing rewards over two objectives; each POI corresponded with one objective. The rovers were initially placed near low-value POIs, in which they received one point for each time step they visited the POI concurrently. This sub-configuration of two neighboring POIs, in which the agents must "choose" between two mutually exclusive objectives, was configured in this manner to facilitate localized tradeoff exploration [Thakar *et al.*, 2025]. However, an identical setup was then placed behind a simple wall, in which rovers must trade lower rewards in the beginning of the episode in order to achieve higher rewards near the end (5 points per POI per time step). This discovery requires agents to explore beyond their immediate location despite the immediate rewards provided to them. This configuration is shown in Figure 1.

Experiment 2: Temporal Tradeoff

The purpose of this experiment was to evaluate temporal objectives. As opposed to POIs representing different objectives, the POIs in this experiment did not directly represent the objectives. Instead, the objectives were the reward collected across all POIs over the first half of the episode, and then over the total episode. The agent team was rewarded one point each per near (inner ring) POI, and three points each per outer POI. Inner POIs had a coupling factor of one, so agents could activate them individually. Outer POIs were more difficult to activate, with a coupling factor of two. The episode length was short enough such that the most efficient team would be unable to collect all POIs and be forced to make tradeoffs between collecting POIs quickly in the episode, versus sacrificing easy, early rewards for higher value, resource intensive rewards.

4.2 Compared Methods

We compare the standard NSGA-II policy evaluation procedure with entropy-augmented policy evaluation based on scaled kNN entropy. Entropy bonuses represent an introduction of behavior-space influence in policy evaluation and are scaled using the parameter, β = 0.05, 0.1, 0.25, and 0.5. A β value of 0.0 represents objective-space evaluation only.

4.3 Analysis Protocol

In the two-agent experiment, each trial was run for 10,000 time steps. Each β value was run 30 times for Experiment 1, and 3 times for Experiment 2, using the same random seed for each β value and a different random seed for each run. Experiment 2 should be interpreted as a preliminary evaluation due to the limited number of independent trials. All policies were rolled out at every timestep. The average and standard error of each beta value's hypervolumes over all runs and generations were collected and used to compare objective-space policy evaluation to a combined behavior- and objective-space evaluation. Hypervolume was selected as the primary performance metric because it jointly captures convergence and coverage of the Pareto front. For hypervolume calculation, a negative reference point was selected, i.e., (-1, -1) following standard hypervolume evaluation practice for this reward range.

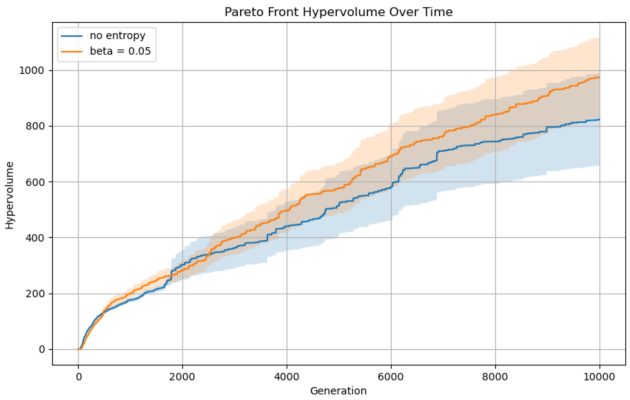


Figure 3: Experiment 1 Pareto front hypervolumes over NSGA-II generations with and without entropy-augmented fitness evaluation. Shaded regions represent the error.

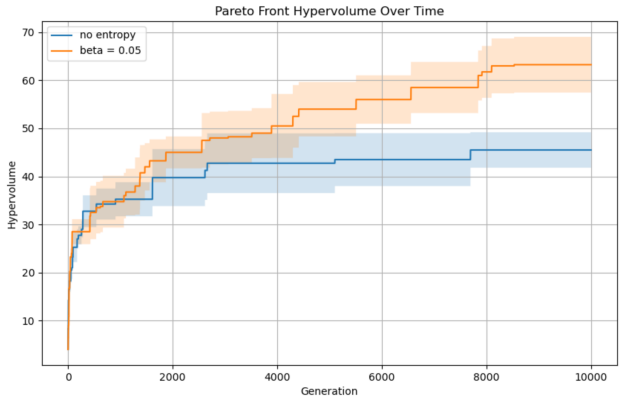


Figure 4: Experiment 2 Pareto front hypervolumes over NSGA-II generations, with shaded regions representing the error. The lower variance relative to Experiment 1 reflects the absence of a strong local optimum in the temporal tradeoff domain.

5 Results and Discussion

This section discusses the results from the study and evaluates the implications of introducing behavior-space awareness into the policy evaluation step of multi-objective evolutionary algorithms. The aim of this study is to evaluate whether entropy-augmented policy evaluation influences the quality of the Pareto fronts generated by NSGA-II.

5.1 Pareto Front Quality

For evaluation, we observe the total hypervolume beneath each Pareto front, which indirectly measures both the quality and diversity of the optimal solution sets. Figures 3 and 4 compare the baseline with the best-performing entropy coefficient identified during parameter tuning. A broader range of β values were tested in initial experiments (Table 1); however, preliminary parameter tuning suggested that the small value of 0.05 provided the most promising results.

In Experiment 1, the entropy-augmented method achieved a hypervolume of 973.34 ± 142 compared to 822 ± 164.79 for the baseline (Table 3), an average increase of 18%. However,

Table 2: Final Pareto-front hypervolume for Experiment 1 across entropy coefficients. Values are reported as mean $\pm$ standard error (SE); n=29.

β	Hypervolume
0.0	822 ± 164.79
0.05	$\mathbf{973.34 \pm 142.43}$

Table 3: Final Pareto-front hypervolume for Experiment 2 across entropy coefficients. Values are reported as mean $\pm$ standard error (SE); n=3.

β	Hypervolume
0.0	42.67 ± 3.33
0.05	$\mathbf{63.25 \pm 5.79}$

as shown in Figure 3, there is high variance in the results, causing uncertainty in the significance of the result. One possible explanation for the observed variance is that the wall creates a strong local optimum attractor as shown in Figure 1 that heavily depends on early random initialization and if the agents happen to explore around the edge simultaneously. Therefore, to reach statistically significant results, this experiment likely requires many more trials to reduce the standard error. However, despite the high variance, entropy augmentation produced higher hypervolume than the baseline in most runs, suggesting that entropy augmentation may be beneficial even in high-variance settings.

In Experiment 2, the entropy-augmented method achieved a final hypervolume of 63.25 ± 5.79, compared with 42.67 ± 3.33 (Table 3). Although these preliminary experiments showed a substantial improvement in average hypervolume, the evaluation was conducted using a limited number of trials. Consequently, these results should be interpreted as exploratory and warrant additional evaluation to determine their robustness.

6 Conclusions

To our knowledge, this is the first work to incorporate both objective- and behavior-space diversity into Pareto optimization for multiagent systems. Rather than introducing entropy as a competing objective, an entropy-based bonus is distributed directly across objective fitness values, preserving the global Pareto structure. Preliminary results suggest that incorporating behavioral diversity into policy evaluation is a promising direction for improving multi-objective evolutionary search. However, additional evaluation across a broader range of random seeds, domains, and problem settings is needed to assess the robustness and generality of the observed improvements.

6.1 Limitations and Future Work

While the proposed fitness shaping strategy is simple to incorporate into existing MOEAs, several challenges remain. The efficacy of the method depends on the relative magnitude of the entropy bonus. If the entropy bonus becomes too large relative to the task rewards, it can dominate the fitness evaluation and diminish the influence of the original objectives. Selecting or adapting the scaling factor remains an important challenge, particularly in sparse-reward domains.

Future work will explore more complex environment configurations, such as incorporating additional agents and POIs, as well as extending episode length to better evaluate the discovery of novel behaviors. In addition, we will investigate the conditions under which entropy augmentation is beneficial, including its sensitivity to problem characteristics, random initialization, and the choice of behavioral characterization. While entropy was selected as an initial implicit summary of team behavior, alternative behavioral summaries and diversity measures may better capture task-relevant coordination patterns while preserving diverse collaborative behaviors during evolutionary search.

References

[Aydeniz *et al.*, 2023] Ayhan Alp Aydeniz, Robert Loftin, and Kagan Tumer. Novelty seeking multiagent evolutionary reinforcement learning. In *Proceedings of the genetic and evolutionary computation conference*, pages 402–410, 2023.

[Chatterjee *et al.*, 2006] Krishnendu Chatterjee, Rupak Majumdar, and Thomas A Henzinger. Markov decision processes with multiple objectives. In *Annual symposium on theoretical aspects of computer science*, pages 325–336. Springer, 2006.

[Deb and Jain, 2013] Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. *IEEE transactions on evolutionary computation*, 18(4):577–601, 2013.

[Deb *et al.*, 2002] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[Dong *et al.*, 2025] Xiaoyi Dong, Jian Cheng, and Xi Sheryl Zhang. Maximum entropy reinforcement learning with diffusion policy. *arXiv preprint arXiv:2502.11612*, 2025.

[Felten *et al.*, 2024] Florian Felten, Umut Ucak, Hicham Azmani, Gao Peng, Willem Röpke, Hendrik Baier, Patrick Mannion, Diederik M Roijers, Jordan K Terry, El-Ghazali Talbi, et al. Momaland: A set of benchmarks for multi-objective multi-agent reinforcement learning. *arXiv preprint arXiv:2407.16312*, 2024.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[Heidemann *et al.*, 2006] John Heidemann, Wei Ye, Jack Wills, Affan Syed, and Yuan Li. Research challenges and applications for underwater sensor networking. In

IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006., volume 1, pages 228–235. IEEE, 2006.

[Hu *et al.*, 2023] Tianmeng Hu, Biao Luo, Chunhua Yang, and Tingwen Huang. Mo-mix: Multi-objective multi-agent cooperative decision-making with deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12098–12112, 2023.

[Ingvarsson *et al.*, 2023] Garar Ingvarsson, Mikayel Samvelyan, Bryan Lim, Manon Flageat, Antoine Cully, and Tim Rocktäschel. Mix-me: Quality-diversity for multi-agent learning. *arXiv preprint arXiv:2311.01829*, 2023.

[Janmohamed and Cully, 2025] Hannah Janmohamed and Antoine Cully. Multi-objective quality-diversity in unstructured and unbounded spaces. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 149–157, 2025.

[Kozachenko, 1987] Leonenko Kozachenko. Sample estimate of the entropy of a random vector. *Probl. Pered. Inform.*, 23:9, 1987.

[Lehman and Stanley, 2011a] Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[Lehman and Stanley, 2011b] Joel Lehman and Kenneth O Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pages 211–218, 2011.

[Ma *et al.*, 2023] Haiping Ma, Yajing Zhang, Shengyi Sun, Ting Liu, and Yu Shan. A comprehensive survey on nsga-ii for multi-objective optimization and applications. *Artificial Intelligence Review*, 56(12):15217–15270, 2023.

[Nayak *et al.*, 2025] Sharan Nayak, Grace Lim, Federico Rossi, Michael Otte, and Jean-Pierre de la Croix. Multi-robot exploration for the cadre mission. *Autonomous Robots*, 49(2):17, 2025.

[Pappas *et al.*, 2021] Iosif Pappas, Styliani Avraamidou, Justin Katz, Baris Burnak, Burcu Beykal, Metin Turkay, and Efstratios N Pistikopoulos. Multiobjective optimization of mixed-integer linear programming problems: a multiparametric optimization approach. *Industrial & engineering chemistry research*, 60(23):8493–8503, 2021.

[Pierrot *et al.*, 2022] Thomas Pierrot, Guillaume Richard, Karim Beguir, and Antoine Cully. Multi-objective quality diversity optimization. In *Proceedings of the genetic and evolutionary computation conference*, pages 139–147, 2022.

[Roijers *et al.*, 2013] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[Tan *et al.*, 2005] Kay Chen Tan, Eik Fun Khor, and Tong Heng Lee. *Multiobjective evolutionary algorithms and applications*. Springer, 2005.

[Thakar *et al.*, 2025] Raghav Thakar, Gaurav Dixit, Siddarth Iyer, and Kagan Tumer. Multiagent credit assignment for multi-objective coordination. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 663–672, 2025.

[Tumer *et al.*, 2002] Kagan Tumer, Adrian K Agogino, and David H Wolpert. Learning sequences of actions in collectives of autonomous agents. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems: Part 1*, pages 378–385, 2002.

[Van Moffaert *et al.*, 2014] Kristof Van Moffaert, Tim Brys, and Ann Nowé. Efficient weight space search in multi-objective reinforcement learning. *Relatório técnico, Vrije Universiteit Brussel*, 2014.

[Yamauchi, 1998] Brian Yamauchi. Frontier-based exploration using multiple robots. In *Proceedings of the second international conference on Autonomous agents*, pages 47–53, 1998.

[Yuan *et al.*, 2014] Yuan Yuan, Hua Xu, and Bo Wang. An improved nsga-iii procedure for evolutionary many-objective optimization. In *Proceedings of the 2014 annual conference on genetic and evolutionary computation*, pages 661–668, 2014.

[Zhang and Li, 2007] Qingfu Zhang and Hui Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, 11(6):712–731, 2007.

[Zhou *et al.*, 2011] Aimin Zhou, Bo-Yang Qu, Hui Li, Shi-Zheng Zhao, Ponnuthurai Nagaratnam Suganthan, and Qingfu Zhang. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and evolutionary computation*, 1(1):32–49, 2011.

[Zitzler *et al.*, 2001] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. Spea2: Improving the strength pareto evolutionary algorithm. *TIK report*, 103, 2001.

[Zitzler *et al.*, 2004] Eckart Zitzler, Marco Laumanns, and Stefan Bleuler. A tutorial on evolutionary multiobjective optimization. *Metaheuristics for multiobjective optimisation*, pages 3–37, 2004.