POGEMA: A Benchmark Platform for Cooperative Multi-Agent Navigation

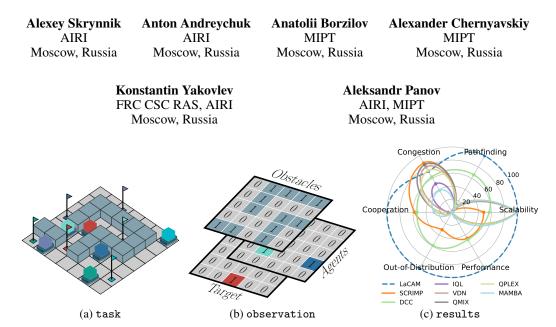


Figure 1: (a) Example of the multi-robot navigation problem considered in POGEMA: each robot must reach its goal, denoted by a flag of the same color. (b) Observation tensor of the red agent. (c) Results of the evaluation of several MARL, hybrid, and search-based solvers on the proposed POGEMA benchmark.

Abstract

Multi-agent reinforcement learning (MARL) has recently excelled in solving challenging cooperative and competitive multi-agent problems in various environments with, mostly, few agents and full observability. Moreover, a range of crucial robotics-related tasks, such as multi-robot navigation and obstacle avoidance, that have been conventionally approached with the classical non-learnable methods (e.g., heuristic search) is currently suggested to be solved by the learning-based or hybrid methods. Still, in this domain, it is hard, not to say impossible, to conduct a fair comparison between classical, learning-based, and hybrid approaches due to the lack of a unified framework that supports both learning and evaluation. To this end, we introduce POGEMA, a set of comprehensive tools that includes a fast environment for learning, a generator of problem instances, the collection of pre-defined ones, a visualization toolkit, and a benchmarking tool that allows automated evaluation. We introduce and specify an evaluation protocol defining a range of domain-related metrics computed on the basics of the primary evaluation indicators (such as success rate and path length), allowing a fair multi-fold comparison. The results of such a comparison, which involves a variety of state-of-the-art MARL, search-based, and hybrid methods, are presented.

1 Introduction

Multi-agent reinforcement learning (MARL) has gained an increasing attention recently and significant progress in this field has been achieved [6, 32, 62]. MARL methods have been demonstrated to generate well-performing agents' policies in strategic games [2, 63], sport simulators [49, 66], multi-component robot control [59], city traffic control [20], and autonomous driving [70]. Currently, several ways to formulate and solve MARL problems exist, based on what information is available to the agents and what type of communication is allowed in the environment [67]. Due to the increased interest in robotic applications, decentralized cooperative learning with minimizing communication between agents has recently attracted a specific attention [45, 68]. Decentralized learning naturally suits the partial observability of the environment in which the robots usually operate. Reducing the information transmitted through the communication channels between the agents increases their degree of autonomy.

The main challenges in solving MARL problems are the non-stationarity of the multi-agent environment, the need to explicitly predict the behavior of the other agents to implement cooperative behavior, high dimensionality of the action space, which grows exponentially with the number of agents, and the sample inefficiency of existing approaches. The existing MARL including model-based and hybrid learnable methods [12, 26] exhibit faster and more stable learning in SMAC-type environments [13] with vector observations and full observability. Currently, the best results are shown by the discrete explicit world models, that use Monte Carlo tree search for planning with various heuristics to reduce the search space [18, 26].

However, in numerous practically inspired applications, like in mobile robot navigation, agents' observations are typically high-dimensional (e.g. stacked occupancy grid matrices or image-based observations as compared to 32-dim vectors in SMAC [13]) and only partially describe the state of the environment, including the other agents [17, 14]. This makes the problem specifically challenging, especially in the environments where a large number of agents are involved. For example, it is not uncommon in robotics to consider settings where up to hundreds of agents are acting (moving) simultaneously in the shared workspace as opposed to 2–10 agents in conventional MARL environments such as SMAC [13] or Google Research Football [21]. Learning to act in such crowded, observation-rich and partially-observable environments is a notable challenge to existing MARL methods.

Conventionally, the problem of multi-robot cooperative navigation (which is very important due to its applications in modern automated warehouses and fulfillment centers [11]) is framed as a search problem over a discretized search space, composed of robots-locations tuples. All robots are assumed to be confined to a graph, typically – a 4-connected grid [38], and at each time step a robot can either move following a graph's edge or stay at the current vertex. This problem setting is known as (Classical) Multi-agent Pathfinding problem [50]. Even in such simplified setting (discretized space, discretized time, uniform-duration actions etc.) obtaining a set of individual plans (one for each robot) that are mutually-conflict-free (i.e. no vertex or edge is occupied by discrinct agents at the same time step) and minimize a common objective such as, for example, the arrival time of the last agent (known as the makespan in the literature) is NP-Hard [65]. Moreover if the underlying graph is directed even obtaining a valid solution is HP-Hard as well [31].

To this end the focus of the multi-agent pathfinding community is recently being shifted towards exploring of how state-of-the-art machine learning techniques, especially reinforcement learning and imitation learning, can be leveraged to increase the efficiency of traditional solvers. Methods like [48, 46, 47, 29, 61, 42, 58, 27, 10] are all hybrid solvers that rely on both widespread search-based techniques and learnable components as well. They all are developed using different frameworks, environments and datasets and are evaluated accordingly, i.e. in the absence of the unifying evaluation framework, consisting of the (automated) evaluation tool, protocol (that defines common performance indicators) and the dataset of the problem instances. Moreover, currently most of the pure MARL methods, i.e. the ones that do not involve search-based modules, such as QMIX [37], MAMBA [12], MAPPO [64] etc., are mostly not included in comparison. The main reason is that to train MARL policies a fast environment is needed, which is suited to cooperative multi-agent navigation.

To close the mentioned gaps we introduce POGEMA, a comprehensive set of tools that includes:

 a fast and flexible environment for learning and planning supporting several variants of the multi-robot navigation problem,

- a generator of problem instances for multi-task and generalization testing,
- a visualization toolkit to create plots for debugging and performance information and to make high-quality animations,
- a benchmarking tool that allows automated evaluation of both learnable, planning, and hybrid approaches.

Moreover, we introduce and specify an evaluation protocol defining a range of domain-related metrics computed on the basics of the primary evaluation indicators (such as success rate and path length), allowing a fair multi-fold comparison of learnable and classical methods. The results of such a comparison, which involves a range of the state-of-the-art MARL, search-based, and hybrid methods, are presented. The source code for POGEMA Benchmark is available at the following links¹²³. Additionally, the POGEMA environment and pogema-toolbox can be installed via PyPI using pip.

2 Related Work

Currently, a huge variety of MARL environments exists that are inspired by various practical applications and encompass a broad spectrum of nuances in problem formulations. Notably, they include a diverse array of computer games [41, 13, 39, 7, 51, 19, 5, 3, 21]. Additionally, they address complex social dilemmas [1] including public goods games, resource allocation problems [35], and multi-agent coordination challenges. Some are practically inspired, showcasing tasks such as competitive object tracking [34], infrastructure management and planning [23], and automated scheduling of trains [30]. Beyond these, the environments simulate intricate, interactive systems such as traffic management and autonomous vehicle coordination [57], multi-agent control tasks [39, 36], and warehouse management [16]. Each scenario is designed to challenge and analyze the collaborative and competitive dynamics that emerge among agents in varied and complex contexts. We summarize the most wide-spread MARL environments in Table 1. A detailed description of each column is presented below.

Navigation Navigation tasks arise in almost all multi-agent environments (e.g. unit navigation in SMAC or robotic warehouse management in RWARE), however only a handful of environments specifically focus on challenging navigation problems: Flatland, Nocturne, RWARE, and POGEMA.

Partially observable Partial observability is an intrinsic feature of a generic multi-agent problem, meaning that an individual agent does not have access to the full state of the environment but rather is able to observe it only locally (e.g. an agent is able to determine the locations of the other agents and/or static obstacles only in its vicinity). Most of the considered environments are partially observable, with the exception of Overcooked.

Python based Most of the current multi-agent environments are implemented in Python, however, some of them include bindings of the other program languages or external dependencies that might complicate their usage. Pure Python implementations generally ensure ease of modification and customization, allowing researchers to readily adapt and extend the environments.

Hardware-agnostic setup means that the environment doesn't require setup for any specific type of training or inferencing hardware.

Performance >10K Steps/s Training and evaluating multi-agent reinforcement learning agents often requires making billions of steps (transitions) in the environment. Thus, it is crucial that each transition is computed efficiently. In general, performing more than 10K steps per second is a good indicator of the environment's efficiency. While XLA versions can provide high performance by vectorizing the environment on GPU, they require modern hardware setups, which can be a barrier for some researchers. In contrast, fast environments like POGEMA or RWARE can achieve high performance without such stringent hardware requirements, making them more accessible and easier to integrate into a variety of research projects.

¹https://github.com/AIRI-Institute/pogema-benchmark

²https://github.com/AIRI-Institute/pogema-toolbox

³https://github.com/AIRI-Institute/pogema

Table 1: Comparison of different multi-agent reinforcement learning environments

Environment	Repository	Navigation	Partially observable	Python based	Hardware-agnostic setup	Performance >10K Steps/s	Procedural generation	Requires generalization	Evaluation protocols	Tests & CI	PyPi Listed	Scalability >1000 Agents	Induced behavior
Flatland [30]	link	1	1	1	1	Х	1	1	1	Х	1	1	Соор
GoBigger [69]	link	1	1	1	/	Х	Х	Х	1	Х	1	Х	Mixed/Coop
Google Research Football [21]	link	/	/	Х	Х	Х	Х	Х	Х	1	Х	Х	Mixed
Griddly [4]	link	1	1	X	Х	1	1	Х	X	1	1	1	Mixed
Hide-and-Seek [3]	link	1	1	1	Х	Х	Х	Х	Х	Х	X	Х	Comp
IMP-MARL [23]	link		1	1	1	X	X	X	1	X	X	1	Coop
Jumanji (XLA) [5]	link	1	✓	✓	Х	1	X	X	✓	1	✓	Х	Mixed
LBF [35]	link		1	1	1	X	X	X	X	1	1	Х	Coop
MAMuJoCo [36]	link		✓	✓	/	X	X	X	X	✓	✓	Х	Coop
MATE [34]	link		✓	✓	1	Х	Х	X	✓	1	Х	Х	Coop
MeltingPot [1]	link		/	X	Х	X	X	✓	✓	/	/	Х	Mixed/Coop
Minecraft MALMO [19]	link		✓	Х	Х	Х	✓	✓	✓	/	Х	Х	Mixed
MPE [28]	link		/	/	/	/	X	X	X	Х	/	Х	Mixed
MPE (XLA) [39]	link		/	/	Х	1	Х	Х	Х	1	1	Х	Mixed
Multi-agent Brax (XLA) [39]	link		/	/	X	/	X	Х	X	/	/	Х	Coop
Multi-Car Racing [44]	link		/	/	1	Х	Х	Х	Х	X	X	X	Comp
Neural MMO [51]	link		/	/	/	X	/	Х	/	/	/	/	Comp
Nocturne [57]	link		/	X	Х	Х	X	X	/	/	X	/	Mixed
Overcooked [7]	link		X	1	/	X	X	✓	/	/	/	Х	Coop
Overcooked (XLA) [39]	link		X	/	Х	/	X	✓	X	/	/	/	Coop
RWARE [35]	link		/	/	/	/	X	Х	X	/	/	Х	Coop
SISL [16]	link		/	1	/	/	X	Х	X	1	/	Х	Coop
SMAC [41]	link		✓.	Х	Х	Х	Х	Х	1	Х	Х	Х	Mixed/Coop
SMAC v2 [13]	link		✓,	X	Х	X	Х	Х	√	X	X	X	Mixed/Coop
SMAX (XLA) [39]	link	/	/	1	X	1	X	X	X	1	1	1	Mixed/Coop
POGEMA (ours)	link	✓	✓	✓	✓	✓	✓	✓	✓	1	✓	✓	Mixed

Procedural generation To improve the ability of RL agents to solve problem instances that are not the same that were used for training (the so-called ability to generalize) procedural generation of the problem instances is commonly used. I.e. the environment does not rely on a predefined set of training instances but rather procedurally generates them to prevent overfitting. As highlighted in the Proceen paper [9], this approach ensures that agents develop robust strategies capable of adapting to novel and diverse situations. Moreover, in multi-agent settings, agents must be able to handle and adapt to a variety of unforeseen agent behaviors and strategies, ensuring robustness and flexibility in dynamic environments [1].

Evaluation protocols means that the environment features a comprehensive evaluation API, including computation of distinct performance indicators and visualization tools. These capabilities allow precise performance measurement and deeper insights into RL agents' behavior, going beyond just reward curves, which can often hide agents exploiting the reward system rather than genuinely solving the tasks.

Tests and CI means that the environment is set up for development with continuous integration and is covered with tests, which are essential for collaborative open-source development.

PyPI listed means that the environment library is listed on PyPI⁴, making it easy to install and integrate into your projects with a simple pip install command.

⁴https://pypi.org

Scalability to >1000 Agents stands for the environment's ability to handle more than 1000 simultaneously acting agents, ensuring robust performance and flexibility for large-scale multi-agent simulations.

Induced behaviour Multi-agent behaviour can be induced by modifying the reward function. Competitive (Comp) behaviour occurs when all agents' rewards sum to a fixed constant at each time step. As an example, minimax games, where rewards sum up to zero, are competitive games. Cooperative/collaborative (Coop) behaviour happens when the agents share a reward function – e.g. when both agents have the same goal in the environment and get rewarded only by its completion. Mixed behaviour (Mixed) does not constrain the reward function in any way and can be described as a combination of cooperative and competitive behaviours. One of the most famous examples of these games is iterated prisoner's dilemma.

As we aim to create a lightweight and easy-to-configure multi-agent environment for reinforcement learning and pathfinding tasks, we consider the following factors essential. First and foremost, our environment is fully compatible with the native Python API: we target pure Python builds independent of hardware-specific software with a minimal number of external dependencies. Moreover, we underline the importance of constant extension and flexibility of the environment. Thus, we prioritize testing and continuous integration as cornerstones of the environment, as well as trouble-free modification of the transition dynamics. Secondly, we highlight that our environment targets generalization and may utilize procedural generation. Last but not least, we target high computational throughput (i.e., the number of environment steps per second) and robustness to an extremely large number of agents (i.e., the environment remains performant under high loads).

There are many environments inducing various types of multi-agent behaviors via different reward structures. Unfortunately, many of them require extensive Python support and rely on APIs of different programming languages (e.g., Lua, C++) for lower latency or depend on hardware-specific libraries such as XLA. Furthermore, many environments do not support generalization and lack procedural generation, especially in multi-agent cases. Additionally, customization of certain environments might be considered an issue without reverse engineering them. That's why we emphasize the superiority of the proposed benchmark.

Despite the diversity of available environments, most research papers tend to utilize only a selected few. Among these, the most popular are the StarCraft Multi-agent Challenge (SMAC), Multi-agent MuJoCo (MAMuJoCo), and Google Research Football (GRF), with SMAC being the most prevalent in top conference papers. The popularity of these environments is likely due to their effective contextualization of algorithms. For instance, to demonstrate the advantages of a method, it is crucial to test it within a well-known environment.

The evaluation protocols in these environments typically feature learning curves that highlight the performance of each algorithm under specific scenarios. For SMAC, these scenarios involve games against predefined bots with specific units on both sides. In MAMuJoCo, the standard tasks involve agents controlling different sets of joints, while in GRF, the scenarios are games against predefined policies from Football Academy scenarios. Proper evaluation of MARL approaches is a serious concern. For SMAC, it's highlighted in the paper [15], which proposes a unified evaluation protocol for this benchmark. This protocol includes default evaluation parameters, performance metrics, uncertainty quantification, and a results reporting scheme.

The variability of results across different studies underscores the importance of a well-defined evaluation protocol, which should be developed alongside the presentation of the environment. In our study, we provide not only the environment but also the evaluation protocol, popular MARL baselines, and modern learnable MAPF approaches to better position our benchmark within the context.

3 POGEMA

POGEMA, which comes from Partially-Observable Grid Environment for Multible Agents, is an umbrella name for a collection of versatile and flexible tools aimed at developing, debugging and evaluating different methods and policies tailored to solve several types of multi-agent navigation tasks.

3.1 POGEMA Environment

POGEMA⁵ environment is a core of POGEMA suite. It implements the basic mechanics of agents' interaction with the world. The environment can be installed using the Python Package Index (PyPI). The environemnt is open-sourced and available at github⁶ under MIT license. POGEMA provides integration with existing RL frameworks: PettingZoo [54], PyMARL [40], and Gymnasium [56].

Basic mechanics The workspace where the agents navigate is represented as a grid composed of blocked and free cells. Only the free cells are available for navigation. At each timestep each agent individually and independently (in accordance with a policy) picks an action and then these actions are performed simultaneously. POGEMA implements collision shielding mechanism, i.e. if an agent picks an action that leads to an obstacle (or out-of-the-map) than it stays put, the same applies for two or more agents that wish to occupy the same cell. POGEMA also has an option when one of the agents deciding to move to the common cell does it, while the others stay where they were. The episode ends when the predefined timestep, episode length, is reached. The episode can also end before this timestep if certain conditions are met, i.e. all agents reach their goal locations if MAPF problem (see below) is considered.

Problem settings POGEMA supports two generic types of multi-agent navigation problems. In the first variant, dubbed MAPF (from Multi-agent Pathfinding), each agent is provided with the unique goal location and has to reach it avoiding collisions with the other agents and static obstacles. For MAPF problem setting POGEMA supports both *stay-at-target* behavior (when the episode successfully ends only if all the agents are at their targets) and *disappear-at-target* (when the agent is removed from the environment after it first reaches its goal). The second variant is a *lifelong* version of multi-agent navigation and is dubbed accordingly – LMAPF. Here each agent upon reaching a goal is immediately assigned another one (not known to the agent beforehand). Thus the agents are constantly moving trough in the environment until episode ends.

Observation At each timestep each agent in POGEMA receives an individual ego-centric observation represented as a tensor – see Fig. 1. The latter is composed of the following $(2R+1) \times (2R+1)$ binary matrices, where R is the observation radius set by the user:

- 1. Static Obstacles 0 means the free cell, 1 static obstacle
- 2. Other Agents 0 means no agent in the cell, 1 the other agent occupies the cell
- 3. Targets projection of the (current) goal location of the agent to the boundary of its field-of-view

The suggested observation, which is, indeed, minimalist and simplistic, can be modified by the user using wrapper mechanisms. For example, it is not uncommon in the MAPF literature to augment the observation with additional matrices encoding the agent's path-to-goal (constructed by some global pathfinding routine) [46] or other variants of global guidance [29].

Reward POGEMA features the most intuitive and basic reward structure for learning. I.e. an agent is rewarded with +1 if it reaches the goal and receives 0 otherwise. For MARL policies that leverage centralized training a shared reward is supported, i.e. $r_t = goals/agents$ where goals is the number of goals reached by the agents at timestep t and agents is the number of agents. Indeed, the user can specify its own reward using wrappers.

Performance indicators The following performance indicators are considered basic and are tracked in each episode. For MAPF they are: *Sum-of-costs* (*SoC*) and *makespan*. The former is the sum of time steps (across all agents) consumed by the agents to reach their respective goals, the latter is the maximum over those times. The lower those indicators are the more effectively the agents are solving MAPF tasks. For LMAPF the primary tracked indicator is the *throughput* which is the ratio of the number of the accomplished goals (by all agents) to the episode length. The higher – the better.

⁵https://pypi.org/project/pogema

⁶https://github.com/AIRI-Institute/pogema

3.2 POGEMA Toolbox

The POGEMA Toolbox is a comprehensive framework designed to facilitate the testing of learning-based approaches within the POGEMA environment. This toolbox offers a unified interface that enables the seamless execution of any learnable MAPF algorithm in POGEMA. Firstly, the toolbox provides robust management tools for custom maps, allowing users to register and utilize these maps effectively within POGEMA. Secondly, it enables the concurrent execution of multiple testing instances across various algorithms in a distributed manner, leveraging Dask⁷ for scalable processing. The results from these instances are then aggregated for analysis. Lastly, the toolbox includes visualization capabilities, offering a convenient method to graphically represent aggregated results through detailed plots. This functionality enhances the interpretability of outcomes, facilitating a deeper understanding of algorithm performance.

POGEMA Toolbox offers a dedicated tool for map generation, allowing the creation of three distinct types of maps: random, mazes and warehouse maps. All generators facilitates map creation using adjustable parameters such as width, height, and obstacle density. Additionally, maze generator includes specific parameters for mazes such as the number of wall components and the length of walls. The maze generator was implemented based on the generator provided in [10]. POGEMA Toolbox⁸ can be installed using PyPI, and licenced under Apache License 2.0.

3.3 Baselines

POGEMA integrates a variety of MARL, hybrid and planning-based algorithms with the environment. These algorithms, recently presented, demonstrate state-of-the-art performance in their respective fields. Table 2 highlights the differences between these approaches. Some, such as LaCAM and RHCR, are centralized search-based planners. Other approaches, such as SCRIMP and DCC, while decentralized, still require communication between agents to resolve potential collisions. Learnable modern approaches for LifeLong MAPF that do not utilize communication include Follower [46], MATS-LP [47], and Switchers [48] (Assistant Switcher, Learnable Switcher). All these approaches utilize independent PPO [43] as the training method.

The following modern MARL algorithms are included as baselines: MAMBA [12], QPLEX [60], IQL [53], VDN [52], and QMIX [37]. For environment preprocessing, we used the preprocessing scheme provided in the Follower approach, enhancing it with the anonymous targets of other agents' local observations. We utilized the official implementation of MAMBA, as provided by its authors⁹, and employed PyMARL2 framework¹⁰ for establishing MARL baselines.

4 Evaluation Protocol

4.1 Dataset

We include the maps of the following types in our evaluation dataset (with the intuition that different maps topologies are necessary for proper assessment):

- Mazes maps that encouter prolonged corridors with 1-cell width that require high level of cooperation between the agent to accomplish the mission. These maps are procedurally generated.
- Random one of the most commonly used type of maps, as they are easy to generate and allow to avoid overfitting to some special structure of the map. POGEMA ontains an integrated random maps generator, that allows to control the density of the obstacles.
- Warehouses this type of maps are usually used in the papers related to LifeLong MAPF. While there is no narrow passages, high density of the agents might significantly reduce the overall throughput, especially when agents are badly distributed along the map. These maps are also can be procedurally generated.

⁷https://github.com/dask/dask

⁸https://pypi.org/project/pogema-toolbox

⁹https://github.com/jbr-ai-labs/mamba

¹⁰https://github.com/hijkzzz/pymarl2

Table 2: This table provides an overview of various baseline approaches supported by POGEMA and
their features in the context of decentralized multi-agent pathfinding.

Algorithm	Decentralized	Partial Observability	Fully Integrated into POGEMA	Supports MAPF	Supports LifeLong MAPF	No Global Obstacles Map	No Communication	Parameter Sharing	Decentralized Learning	Model-Based	No Imitation Learning
MAMBA [12]	✓	✓	✓	✓	✓	X	✓	✓	Х	✓	✓
QPLEX [60]	1	✓	✓	✓	✓	X	1	✓	X	X	\checkmark
IQL [53]	✓	✓	\checkmark	✓	✓	X	1	✓	✓	X	\checkmark
VDN [52]	1	✓	✓	1	\checkmark	X	1	\checkmark	X	X	✓
QMIX [37]	✓	✓	\checkmark	✓	✓	X	1	✓	X	X	\checkmark
Follower [46]	1	✓	✓	X	✓	X	1	✓	✓	X	\checkmark
MATS-LP [47]	✓	✓	✓	X	✓	X	1	✓	✓	✓	✓
Switcher [48]	1	✓	✓	X	✓	✓	1	✓	✓	X	\checkmark
SCRIMP [61]	✓	✓	X	1	X	X	X	✓	X	X	X
DCC [29]	1	✓	✓	✓	X	X	X	✓	X	X	X
LaCAM [33]	X	X	X	✓	X	X	-	-	-	-	-
RHCR [25]	X	X	X	X	✓	X	-	-	-	-	-

- MovingAI a set of maps from the existing benchmark widely used in MAPF community.
 The contained maps have different sizes and structures. It can be used to show how the
 approach deals with single-agent pathfinding and also deals with the maps that have out-ofdistribution structure.
- MovingAI-tiles a modified MovingAI set of maps. Due to the large size of the original maps, it's hard to get high density of the agents on them. To get more crowded maps, we slice the original maps on 16 pieces with 64×64 size.
- Puzzles a set of small hand-crafted maps that contains some difficult patterns that mandate the cooperation between that agents.

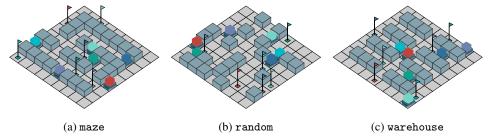


Figure 2: Examples of map generators presented in POGEMA.

Start and goal locations are generated via random generators. They are generated with fixed seeds, thus can be reproduced. It's guaranteed, that each agent has its own goal location and the path to it from its start location exists. Examples of the maps are presented in Figure 2.

4.2 Metrics

The existing works related to solving MAPF problems evaluates the performance by two major criteria – success rate and the primary performance indicators mentioned above: sum-of-costs, makespan, throughput. These are directly obtainable from POGEMA. While these metrics allow to evaluate the algorithms at some particular instance, it's might be difficult to get a high-level conclusion about the performance of the algorithms. Thus, we want to introduce several high-level metrics that covers multiple different aspects:

Performance – how well the algorithm works compared to other approaches. To compute this metric we run the approaches on a set of maps similar to the ones, used during training, and compare the

obtained results with the best ones.

$$Performance_{MAPF} = \begin{cases} SoC_{best}/SoC\\ 0 \text{ if not solved} \end{cases}$$
 (1)

$$Performance_{LMAPF} = throughput/throughput_{best}$$
 (2)

Out-of-Distribution – how well the algorithm works on out-of-distribution maps. This metric is computed in the same way as Performance, with the only difference that the approaches are evaluated on a set of maps, that were not used during training phase and have different structure of obstacles. For this purpose we utilize maps from MovingAI-tiles set of maps.

$$Out_of_Distribution_{MAPF} = \begin{cases} SoC_{best}/SoC \\ 0 \text{ if not solved} \end{cases}$$
 (3)

$$Out_of_Distributionn_{LMAPF} = throughput/throughput_{best}$$
 (4)

Scalability – how well the algorithm scales to large number of agents. To evaluate how well the algorithm scales to large number of agents, we run it on a large warehouse map with increasing number of agents and compute the ratio between runtimes with various number of agents.

$$Scalability = \frac{runtime(agents_1)/runtime(agents_2)}{|agents_1|/|agents_2|}$$
 (5)

Cooperation – how well the algorithm is able to resolve complex situations. To evaluate this metric we run the algorithm on Puzzles set of maps and compare the obtained results with best solutions that were obtained by classical MAPF/LMAPF solvers.

$$Cooperation_{MAPF} = \begin{cases} SoC_{best}/SoC\\ 0 \text{ if not solved} \end{cases}$$
 (6)

$$Cooperation_{LMAPF} = throughput/throughput_{best}$$
 (7)

Congestion – how well the algorithm distributes the agents along the map and reduces redundant waits, collisions, etc. To evaluate this metric we compute the average density of the agents presented in the observations of each agent and compare it to the overall density of the agents on the map.

$$Congestion = \frac{\sum_{i \in agents} agents_density(obs_i)/agents_density(map)}{|agents|}$$
(8)

Pathfinding – how well the algorithm works in case of presence of a single agent on a large map. This metric is tailored to determine the ability of the approach to effectively lead agents to their goal locations. For this purpose we run the approaches on large city maps from MovingAI benchmark sets. The obtained solution cost (in fact - length of the path) should be optimal.

$$Pathfinding = \begin{cases} 1 \text{ if path is optimal} \\ 0 \text{ otherwise} \end{cases}$$
 (9)

4.3 Experimental Results

We have evaluated all the supported baseline algorithms (12 in total) on both MAPF and LMAPF setups on all 6 datasets. The results of this evaluation are presented in Fig.3. The details about number of maps, number of agents, seeds, etc. are given in the supplementary material (as well as details on how these results can be reproduced).

In both setups, i.e. MAPF and LMAPF, the best results in terms of cooperation, out-of-distribution and performance metrics were obtained by centralized planners, i.e. LaCAM and RHCR respectively.

For MAPF tasks, LaCAM outperforms all other approaches on all metrics except congestion. It is hypothesized that in this approach, the even distribution of agents across the environment is not crucial due to its centralized nature, which efficiently resolves complex conflicts. Specialized learnable MAPF approaches, i.e., DCC and SCRIMP, take second place, showing close performance but with different specifics. DCC shows better results on out-of-distribution tasks and pathfinding tasks than SCRIMP, which is better at managing congestion. Surprisingly, the results of SCRIMP are inferior on pathfinding tasks, suggesting a problem with this approach in single-agent tasks that do not require

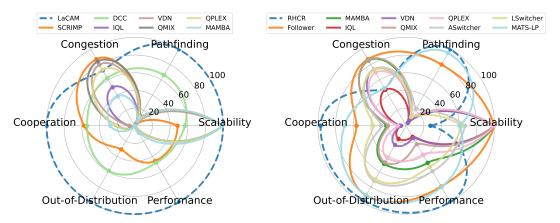


Figure 3: Evaluation of baselines available in POGEMA on (a) MAPF (b) LMAPF instances.

communication, which can be an out-of-distribution setup for this algorithm. MARL algorithms such as QPLEX, VDN, and QMIX underperform in comparison with other approaches, exhibiting a significant gap in the results, which can be attributed to the absence of additional techniques used in hybrid approaches, despite incorporating preprocessing techniques from the Follower approach. This could suggest that the MARL community lacks large-scale approaches and benchmarks for them. Predictably, IQL shows the poorest performance, highlighting the importance of centralized training for multi-agent pathfinding (MAPF) tasks that require high levels of cooperation. The best results out of all MARL approaches are shown by MAMBA, which has non-zero results on the performance metric and better results in terms of the cooperation metric. Its results are still much worse than those obtained by hybrid methods, and it is not able to solve any instances in the out-of-distribution dataset.

For LMAPF tasks, the centralized approach, RHCR, dominates in cooperation, out-of-distribution, performance, and pathfinding metrics. However, it significantly lags behind other approaches in terms of congestion and scalability metrics. Out of the non-centralized approaches, the best results, depending on the metric, are shown by either MATS-LP or Follower. While MATS-LP is better in terms of cooperation and pathfinding, the latter is better in terms of congestion and scalability metrics. It's also worth noting that MATS-LP requires much more runtime than other approaches as it runs MCTS for each of the agents at each step, which takes time (see Appendix F for more details). There are also two more hybrid approaches – ASwitcher and LSwitcher – which differ in the way they switch between planning-based and learning-based parts. One of the reasons for their mediocre results is a total lack of global information, i.e., the Switcher approach assumes that the agents have no information about the global map; thus, each of them needs to reconstruct the map based on their local observations. Again, MARL approaches underperform in these scenarios, with QMIX and QPLEX showing comparable results. QMIX performs better in cooperation and out-of-distribution metrics, while QPLEX excels in performance. As in the MAPF setup, among all the MARL approaches, the best results are demonstrated by MAMBA, which is slightly better than others in terms of cooperation and out-of-distribution metrics and much better in terms of the performance metric.

5 Conclusion and Limitations

This paper presents POGEMA – a powerful suite of tools tailored for creating, assessing, and comparing methods and policies in multi-agent navigation problems. POGEMA encompasses a fast learning environment and a comprehensive evaluation toolbox suitable for pure MARL, hybrid, and search-based solvers. It includes a wide array of methods as baselines. The evaluation protocol described, along with a rich set of metrics, assists in assessing the generalization and scalability of all approaches. Visualization tools enable qualitative examination of algorithm performance. Integration with the well-known MARL API and map sets facilitates the benchmark's expansion. Existing limitations are two-fold. First, a conceptual limitation is that communication between the agents is not currently disentangled in POGEMA environment. Second, the technical limitations include the lack of JAX support and integration with other well-known GPU parallelization tools.

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. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.
- [2] Kai Arulkumaran, Antoine Cully, and Julian Togelius. Alphastar: An evolutionary computation perspective. In *Proceedings of the genetic and evolutionary computation conference companion*, pages 314–315, 2019.
- [3] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020.
- [4] Christopher Bamford. Griddly: A platform for ai research in games. *Software Impacts*, 8:100066, 2021.
- [5] Clément Bonnet, Daniel Luo, Donal John Byrne, Shikha Surana, Paul Duckworth, Vincent Coyette, Laurence Illing Midgley, Sasha Abramowitz, Elshadai Tegegn, Tristan Kalloniatis, et al. Jumanji: a diverse suite of scalable reinforcement learning environments in jax. In *The Twelfth International Conference on Learning Representations*, 2023.
- [6] Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. Applied Sciences, 11(11):4948, 2021.
- [7] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.
- [8] Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared experience actor-critic for multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10707–10717, 2020.
- [9] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020.
- [10] Mehul Damani, Zhiyao Luo, Emerson Wenzel, and Guillaume Sartoretti. Primal _2: Pathfinding via reinforcement and imitation multi-agent learning lifelong. *IEEE Robotics and Automation Letters*, 6(2):2666–2673, 2021.
- [11] Ashutosh Dekhne, Greg Hastings, John Murnane, and Florian Neuhaus. Automation in logistics: Big opportunity, bigger uncertainty. *McKinsey Q*, 24, 2019.
- [12] Vladimir Egorov and Alexei Shpilman. Scalable multi-agent model-based reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 381–390, 2022.
- [13] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 36, 2024.
- [14] Jennifer Gielis, Ajay Shankar, and Amanda Prorok. A critical review of communications in multi-robot systems. *Current robotics reports*, 3(4):213–225, 2022.
- [15] Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems*, 35:5510–5521, 2022.
- [16] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. Autonomous Agents and Multiagent Systems, pages 66–83, 2017.
- [17] Ruihua Han, Shengduo Chen, Shuaijun Wang, Zeqing Zhang, Rui Gao, Qi Hao, and Jia Pan. Reinforcement learned distributed multi-robot navigation with reciprocal velocity obstacle shaped rewards. *IEEE Robotics and Automation Letters*, 7(3):5896–5903, 2022.

- [18] Xiaotian Hao, Jianye Hao, Chenjun Xiao, Kai Li, Dong Li, and Yan Zheng. Multiagent gumbel muzero: Efficient planning in combinatorial action spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12304–12312, 2024.
- [19] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 4246–4247, 2016.
- [20] Máté Kolat, Bálint Kővári, Tamás Bécsi, and Szilárd Aradi. Multi-agent reinforcement learning for traffic signal control: A cooperative approach. Sustainability, 15(4):3479, 2023.
- [21] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4501–4510, 2020.
- [22] Florian Laurent, Manuel Schneider, Christian Scheller, Jeremy Watson, Jiaoyang Li, Zhe Chen, Yi Zheng, Shao-Hung Chan, Konstantin Makhnev, Oleg Svidchenko, et al. Flatland competition 2020: Mapf and marl for efficient train coordination on a grid world. In *NeurIPS* 2020 Competition and Demonstration Track, pages 275–301. PMLR, 2021.
- [23] Pascal Leroy, Pablo G Morato, Jonathan Pisane, Athanasios Kolios, and Damien Ernst. Impmarl: a suite of environments for large-scale infrastructure management planning via marl. Advances in Neural Information Processing Systems, 36, 2024.
- [24] Jiaoyang Li, Zhe Chen, Yi Zheng, Shao-Hung Chan, Daniel Harabor, Peter J Stuckey, Hang Ma, and Sven Koenig. Scalable rail planning and replanning: Winning the 2020 flatland challenge. In *Proceedings of the international conference on automated planning and scheduling*, volume 31, pages 477–485, 2021.
- [25] Jiaoyang Li, Andrew Tinka, Scott Kiesel, Joseph W. Durham, T. K. Satish Kumar, and Sven Koenig. Lifelong multi-agent path finding in large-scale warehouses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11272–11281, May 2021.
- [26] Qihan Liu, Jianing Ye, Xiaoteng Ma, Jun Yang, Bin Liang, and Chongjie Zhang. Efficient multi-agent reinforcement learning by planning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [27] Zuxin Liu, Baiming Chen, Hongyi Zhou, Guru Koushik, Martial Hebert, and Ding Zhao. Mapper: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2020)*, pages 11748–11754, 2020.
- [28] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.
- [29] Ziyuan Ma, Yudong Luo, and Jia Pan. Learning selective communication for multi-agent path finding. *IEEE Robotics and Automation Letters*, 7(2):1455–1462, 2021.
- [30] Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, et al. Flatland-rl: Multi-agent reinforcement learning on trains. arXiv preprint arXiv:2012.05893, 2020.
- [31] Bernhard Nebel. On the computational complexity of multi-agent pathfinding on directed graphs. In *Proceedings of the 30th International Conference on Automated Planning and Scheduling (ICAPS 2020)*, pages 212–216, 2020.
- [32] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.
- [33] Keisuke Okumura. Lacam: Search-based algorithm for quick multi-agent pathfinding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 11655–11662, 2023.
- [34] Xuehai Pan, Mickel Liu, Fangwei Zhong, Yaodong Yang, Song-Chun Zhu, and Yizhou Wang. Mate: Benchmarking multi-agent reinforcement learning in distributed target coverage control. *Advances in Neural Information Processing Systems*, 35:27862–27879, 2022.

- [35] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [36] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.
- [37] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.
- [38] Nicolás Rivera, Carlos Hernández, Nicolás Hormazábal, and Jorge A Baier. The 2[^]k neighborhoods for grid path planning. *Journal of Artificial Intelligence Research*, 67:81–113, 2020.
- [39] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, et al. Jaxmarl: Multi-agent rl environments in jax. *arXiv preprint arXiv:2311.10090*, 2023.
- [40] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft Multi-Agent Challenge. CoRR, abs/1902.04043, 2019.
- [41] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2186–2188, 2019.
- [42] Guillaume Sartoretti, Justin Kerr, Yunfei Shi, Glenn Wagner, TK Satish Kumar, Sven Koenig, and Howie Choset. Primal: Pathfinding via reinforcement and imitation multi-agent learning. *IEEE Robotics and Automation Letters*, 4(3):2378–2385, 2019.
- [43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [44] Wilko Schwarting, Tim Seyde, Igor Gilitschenski, Lucas Liebenwein, Ryan Sander, Sertac Karaman, and Daniela Rus. Deep latent competition: Learning to race using visual control policies in latent space. In *Conference on Robot Learning*, pages 1855–1870. PMLR, 2021.
- [45] Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990, 2022.
- [46] Alexey Skrynnik, Anton Andreychuk, Maria Nesterova, Konstantin Yakovlev, and Aleksandr Panov. Learn to follow: Decentralized lifelong multi-agent pathfinding via planning and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17541–17549, 2024.
- [47] Alexey Skrynnik, Anton Andreychuk, Konstantin Yakovlev, and Aleksandr Panov. Decentralized monte carlo tree search for partially observable multi-agent pathfinding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17531–17540, 2024.
- [48] Alexey Skrynnik, Anton Andreychuk, Konstantin Yakovlev, and Aleksandr I Panov. When to switch: planning and learning for partially observable multi-agent pathfinding. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [49] Yan Song, He Jiang, Haifeng Zhang, Zheng Tian, Weinan Zhang, and Jun Wang. Boosting studies of multi-agent reinforcement learning on google research football environment: the past, present, and future. *arXiv* preprint arXiv:2309.12951, 2023.
- [50] Roni Stern, Nathan R Sturtevant, Ariel Felner, Sven Koenig, Hang Ma, Thayne T Walker, Jiaoyang Li, Dor Atzmon, Liron Cohen, TK Satish Kumar, et al. Multi-agent pathfinding: Definitions, variants, and benchmarks. In *Proceedings of the 12th Annual Symposium on Combinatorial Search (SoCS 2019)*, pages 151–158, 2019.
- [51] Joseph Suarez, David Bloomin, Kyoung Whan Choe, Hao Xiang Li, Ryan Sullivan, Nishaanth Kanna, Daniel Scott, Rose Shuman, Herbie Bradley, Louis Castricato, et al. Neural mmo 2.0: A massively multi-task addition to massively multi-agent learning. *Advances in Neural Information Processing Systems*, 36, 2024.

- [52] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Valuedecomposition networks for cooperative multi-agent learning based on team reward. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pages 2085–2087, 2018.
- [53] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.
- [54] Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.
- [55] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026–5033. IEEE, 2012.
- [56] Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023.
- [57] Eugene Vinitsky, Nathan Lichtlé, Xiaomeng Yang, Brandon Amos, and Jakob Foerster. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world. Advances in Neural Information Processing Systems, 35:3962–3974, 2022.
- [58] Binyu Wang, Zhe Liu, Qingbiao Li, and Amanda Prorok. Mobile robot path planning in dynamic environments through globally guided reinforcement learning. *IEEE Robotics and Automation Letters*, 5(4):6932–6939, 2020.
- [59] Haoyu Wang, Xiaoyu Tan, Xihe Qiu, and Chao Qu. Subequivariant reinforcement learning framework for coordinated motion control. *arXiv preprint arXiv:2403.15100*, 2024.
- [60] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations*, 2020.
- [61] Yutong Wang, Bairan Xiang, Shinan Huang, and Guillaume Sartoretti. Scrimp: Scalable communication for reinforcement-and imitation-learning-based multi-agent pathfinding. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 9301–9308. IEEE, 2023.
- [62] Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6):5023–5056, 2023.
- [63] Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6672–6679, 2020.
- [64] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.
- [65] Jingjin Yu and Steven M LaValle. Optimal multirobot path planning on graphs: Complete algorithms and effective heuristics. *IEEE Transactions on Robotics*, 32(5):1163–1177, 2016.
- [66] Yifan Zang, Jinmin He, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [67] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.
- [68] Lin Zhang, Yufeng Sun, Andrew Barth, and Ou Ma. Decentralized control of multi-robot system in cooperative object transportation using deep reinforcement learning. *IEEE Access*, 8:184109–184119, 2020.

- [69] Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, and Yu Liu. Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [70] Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmellat, Mohsen Rohani, Nicolas Perez Nieves, Yihan Ni, Seyedershad Banijamali, Alexander Cowen Rivers, Zheng Tian, Daniel Palenicek, Haitham bou Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving, 2020.

Appendix Contents

Appendix Sections	Contents
Appendix A	Evaluation Setup Details
Appendix B	Results for MAPF Benchmark
Appendix C	Results for LifeLong MAPF Benchmark
Appendix D	Code examples for POGEMA
Appendix E	POGEMA Toolbox
Appendix F	Extended Related Work
Appendix G	Examples of Used Maps
Appendix H	MARL Training Setup
Appendix I	Resources and Statistics
Appendix J	Community Engagement and Framework Enhancements

A Evaluation Setup Details

POGEMA benchmark contains 6 different sets of maps and all baseline approaches were evaluated on them either on MAPF or on LMAPF instances. Regardless the type of instances, number of maps, seeds and agents were the same. Table 3 contains all information about these numbers. In total, this corresponds to 3,376 episodes for each scenario type. Note that there is no MaxSteps (LMAPF) value for MovingAI set of maps. This set of maps was used only for pathfinding meta-metric, thus all approaches were evaluated only on MAPF instances with a single agent.

Table 3: Details about the instances on different sets of maps.

	Agents	Maps	MapSize	Seeds	MaxSteps (MAPF)	MaxSteps (LMAPF)
Random	[8, 16, 24, 32, 48, 64]	128	17×17 - 21×21	1	128	256
Mazes	[8, 16, 24, 32, 48, 64]	128	$17 \times 17 - 21 \times 21$	1	128	256
Warehouse	[32, 64, 96, 128, 160, 192]	1	33×46	128	128	256
Puzzles	[2, 3, 4]	16	5×5	10	128	256
MovingAI	[1]	8	256×256	10	2048	-
MovingAI-tiles	[64, 128, 192, 256]	128	64×64	1	256	256

B Results for MAPF Benchmark

In this section, we present the extended results of the MAPF benchmark analysis, highlighting the performance, out-of-distribution handling, scalability, cooperation, congestion, and pathfinding capabilities of various approaches. The experiments were conducted on different map types and sizes, employing metrics such as SoC, CSR, and makespan to evaluate effectiveness. Detailed visual and tabular data illustrate how centralized and learnable approaches perform under various conditions.

B.1 Performance

The performance metrics were calculated using Mazes and Random maps of size close to 20×20 . The primary metrics here are SoC and CSR. The results of all the MAPF approaches over different numbers of agents are presented in Figure 4. The superior performance is shown by the centralized approach, LaCAM. The learnable approaches, DCC and SCRIMP, show comparable results. Interestingly, the former has a better SoC metric, despite the latter having better results on CSR. Among the MARL methods, better results are shown by MAMBA for both metrics. However, it narrowly lags behind the specialized approaches, DCC and SCRIMP.

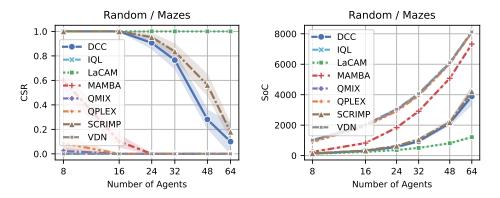


Figure 4: Performance of MAPF approaches on Random and Mazes maps, based on CSR (higher is better) and SoC (lower is better) metrics. The shaded area indicates 95% confidence intervals.

B.2 Out-of-Distribution

Out-of-Distribution metric was calculated on MovingAI-tiles dataset, that consists of pieces of cities maps with 64×64 size. Due to much larger size compared to Mazes and Random maps, the amount of agents was also significantly increased. The results are presented in Figure 5. Here again centralized search-based planner, i.e. LaCAM, demonstrates the best results both in terms of CSR and SoC. Among hybrid methods, DCC completely outperforms the others (including SCRIMP) on the out-of-distribution dataset. MARL approaches are unable to solve any instance even with 64 agents.

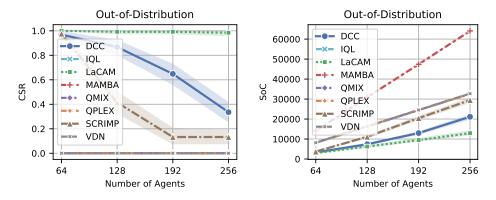


Figure 5: Performance of MAPF approaches on MovingAI-tiles maps. These results were utilized to compute Out-of-Distribution metric. The shaded area indicates 95% confidence intervals.

B.3 Scalability

The results of how well the algorithm scales with a large number of agents are shown in Figure 6. The experiments were conducted on a warehouse map. The plot is log-scaled. The best scalability is achieved with the centralized LaCAM approach, which is a high-performance approach. The worst results in both runtime and scalability are for SCRIMP, with results close to it for DCC. Despite an initially high runtime, the scalability of MAMBA is better than other approaches; however, this could be attributed to the high cost of GPU computation, which is due to the large number of parameters in the neural network and is the limiting factor of this approach.

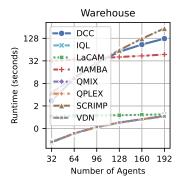


Figure 6: Runtime in seconds for each algorithm. The plot is log-scaled.

B.4 Cooperation

How well the algorithm is able to resolve complex situations on the Puzzles set is reflected in the results presented in Table 4. Surprisingly, the centralized approach LaCAM does not solve all the tasks, showing only a 0.96 CSR. This highlights that this type of task is difficult even for centralized approaches, despite the small map size of 5×5 and the low number of agents (2-4). SCRIMP outperformed DCC in CSR but again showed comparable results in SoC. Among MARL approaches, better cooperation is demonstrated by QMIX, outperforming QPLEX, VDN, IQL, and even MAMBA.

Table 4: Comparison of algorithms' cooperation on Puzzles set. \pm shows confidence intervals 95%. Here, tan boxes highlight the best approach, and teal boxes highlight the best approach with a learnable part.

Algorithm	CSR	SoC
DCC	0.72 ± 0.04	96.33 ± 9.97
IQL	0.27 ± 0.04	316.87 ± 14.47
LaCAM	0.96 ± 0.02	36.29 ± 7.26
MAMBA	0.39 ± 0.04	177.64 ± 14.68
QMIX	0.53 ± 0.05	246.11 ± 17.10
QPLEX	0.33 ± 0.04	250.12 ± 14.80
SCRIMP	0.82 ± 0.04	104.31 ± 13.73
VDN	0.11 ± 0.03	336.99 ± 12.45

B.5 Congestion

Congestion is one of the meta-metrics that estimates how well the agents are distributed along the map. This metric indirectly influences the performance of the approach, as well-distributed agents in case of highly crowded instances allows to reduce the amount of collisions and redundant wait-actions. To compute this metric we utilize the results obtained on Warehouse map with the highest evaluated amount of agents - 192. In Table 5, we present results for a range of agents from 64 to 192 in increments of 32, not just for 192 agents. In contrast to other metrics, that are computed as a ratio to the best obtained results, Congestion metric is computed as the ratio to average density of all agents on the map to average density of the agents in agents' local observations.

Table 5: Average agents density by number of agents across algorithms for Warehouse map. Please note, only column with 192 agents was utilized to compute Congestion metric.

Algorithm	64 Agents	96 Agents	128 Agents	160 Agents	192 Agents
DCC	0.094 ± 0.001	0.132 ± 0.001	$0.170\pm\ 0.001$	0.207 ± 0.001	0.241 ± 0.001
IQL	0.076 ± 0.001	0.114 ± 0.001	0.163 ± 0.002	0.244 ± 0.002	0.299 ± 0.002
LaCAM	0.082 ± 0.001	0.116 ± 0.001	0.149 ± 0.001	0.179 ± 0.001	0.207 ± 0.001
MAMBA	0.101 ± 0.001	0.183 ± 0.001	$0.266 \pm \ 0.002$	0.335 ± 0.002	0.389 ± 0.002
QMIX	0.073 ± 0.001	0.103 ± 0.001	0.130 ± 0.001	0.154 ± 0.001	0.179 ± 0.001
QPLEX	0.077 ± 0.001	0.113 ± 0.001	$0.146 \pm \ 0.001$	0.175 ± 0.001	0.205 ± 0.001
SCRIMP	$0.074 \pm \ 0.001$	$0.104 \pm \ 0.001$	0.127 ± 0.001	0.148 ± 0.001	0.173 ± 0.001
VDN	0.071 ± 0.001	0.101 ± 0.001	0.130 ± 0.001	0.158 ± 0.001	0.188 ± 0.001

B.6 Pathfinding

To compute Pathfidning metric we run the approaches on the instances with a single agent. For this purpose we utilized large MovingAI mapf with 256×256 size, the results are presented in Table 6. While this task seems easy, most of the hybrid and MARL approaches are not able to effectively solve them. Only LaCAM is able to find optimal paths in all the cases, as it utilizes precomputed costs to the goal location as a heuristic. Most of the evaluated hybrid and MARL approaches are also contain a sort of global guidance in one the channels of their observations. However, large maps with out-of-distribution structure, the absence of communication and other agents in local observations are able to lead to inconsistent behavior of the models that are not able to effectively choose the actions that lead the agent to its goal. Please note, SoC and makespan metrics in this case are equal, as there is only one agent in every instance.

Table 6: Comparison of makespan (the lower is better) used for pathfinding metric. tan boxes highlight the best approach, and teal boxes highlight the best approach with a learnable part.

Algorithm	Makespan					
DCC	189.56 ± 28.28					
IQL	1825.95 ± 137.11					
LaCAM	179.82 ± 20.21					
MAMBA	416.45 ± 139.34					
QMIX	955.54 ± 203.76					
QPLEX	933.74 ± 204.21					
SCRIMP	1460.04 ± 176.27					
VDN	1733.20 ± 158.91					

C Results for LifeLong MAPF Benchmark

In this section, we present the extended results of the LifeLong MAPF benchmark analysis, highlighting performance, out-of-distribution handling, scalability, cooperation, congestion, and pathfinding.

C.1 Performance

Performance metric in LMAPF case is based on the ratio of throughput compared to the best obtained one. In contrast to SoC, throughput should be as high as possible. There is also no CSR metric, as there is no need for agents to be at their goal locations simultaneously. As well as in MAPF case, the best results are obtained by centralized search-based approach – RHCR. The best results among decentralized methods demonstrate Follower and MATS-LP, Following them, comparable results are shown by ASwitcher and LSwitcher. Between MARL methods the highest throughput on both Random and Mazes maps is obtained by MAMBA.

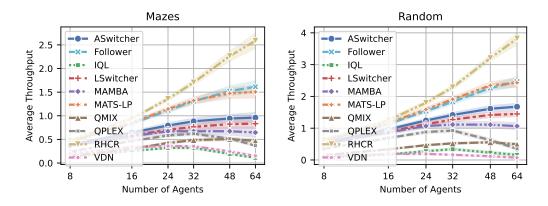


Figure 7: Performance results for LifeLong scenarios on the Mazes and Random maps.

C.2 Out-of-Distribution

The evaluation on out-of-distribution set of maps confirms the results obtained on Random and Mazes maps. The best results demonstrates RHCR. Next best results are obtained by Follower and MATS-LP, which are much closer to RHCR in this experiment. While MATS-LP outperforms Follower on the instances with 64, 128 and 192 agents, Follower is still better on the instances with 256 agents. Such relation is probably explained by the presence of dynamic edge-costs in Follower that allows to better distribute agents along the map and reduce congestion between them.

Table 7: Evaluation on Out-of-Distribution maps.	tan boxes highlight the best approach according to
the average throughput metric, and teal boxes high	alight the best approach with a learnable component.

Algorithm	64 Agents	128 Agents	192 Agents	256 Agents
ASwitcher	1.26 ± 0.08	2.30 ± 0.13	3.14 ± 0.17	3.80 ± 0.20
Follower	1.50 ± 0.08	2.82 ± 0.13	3.95 ± 0.19	4.81 ± 0.22
IQL	0.26 ± 0.02	0.65 ± 0.04	0.78 ± 0.05	0.68 ± 0.05
LSwitcher	1.23 ± 0.07	2.23 ± 0.12	3.06 ± 0.17	3.67 ± 0.20
MAMBA	1.02 ± 0.05	1.42 ± 0.08	2.05 ± 0.12	2.46 ± 0.17
MATS-LP	1.57 ± 0.12	2.98 ± 0.20	4.04 ± 0.33	4.69 ± 0.39
QMIX	0.83 ± 0.03	1.55 ± 0.06	2.01 ± 0.09	2.27 ± 0.11
QPLEX	0.79 ± 0.03	1.48 ± 0.07	1.79 ± 0.10	1.74 ± 0.11
RHCR	1.57 ± 0.08	3.00 ± 0.14	4.22 ± 0.23	5.13 ± 0.34
VDN	0.50 ± 0.03	0.91 ± 0.05	1.03 ± 0.06	0.96 ± 0.06

C.3 Scalability

Figure 8 contains log-scaled plot of average time spent by each of the algorithms to process an instance on Warehouse map with the corresponding amount of agents. Most of the approaches scales almost linearly, except RHCR. This centralized search-based method lacks of exponential grow, as it needs to find a collision-free solution for at least next few steps, rather than just to make a decision about next action for each of the agents. The worst runtime demonstrate MATS-LP, as it runs MCTS and simulates the behavior of the other observable agents. It's still scales better than RHCR as it builds trees for each of the agents independently.

C.4 Cooperation

As well as for MAPF setting, cooperation metric is computed based on the results obtained on Puzzles dataset. Table 8 contains average throughput obtained by each of the evaluated approaches. Here again the best results are obtained by RHCR algorithm. In contrast to Random, Mazes and Warehouse sets of maps, where MATS-LP and Follower demonstrate close results, the ability to simulate the behavior of other agents, provided by MCTS in MATS-LP, allows to significantly outperform Follower on small Puzzles maps. The rest approaches demonstrate much worse results, especially IQL, QPLEX and VDN that have 10 times worse average throughput than RHCR.

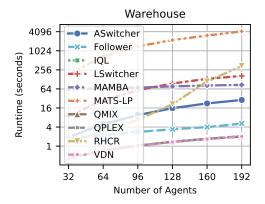


Figure 8: Runtime in seconds for each algorithm. Note that the plot is log-scaled.

Table 8: Average throughput on Puzzles maps that were used to compute Cooperation metric.

Algorithm	Average Throughput
ASwitcher	$0.164 \pm \ 0.015$
Follower	0.319 ± 0.020
IQL	0.036 ± 0.003
LSwitcher	$0.206 \pm \ 0.013$
MAMBA	0.133 ± 0.014
MATS-LP	0.394 ± 0.021
QMIX	0.117 ± 0.010
QPLEX	0.051 ± 0.006
RHCR	0.538 ± 0.021
VDN	0.030 ± 0.004

C.5 Congestion

Table 9 contains average agents density presented in observations. As it was already mentioned in LMAPF:Out-of-Distribution section, Follower contains a mechanism that allows to effectively distribute agents along the map. As a result, the lowest density is demonstrated exactly by this approach. MARL methods, such as MAMBA and QMIX are also demonstrate low average agents density, as they actually utilize the same observation as Follower does.

Table 9: Average Agents Density by Number of Agents

Algorithm	64 Agents	96 Agents	128 Agents	160 Agents	192 Agents
ASwitcher	0.074 ± 0.001	$0.111\pm\ 0.001$	0.146 ± 0.001	0.176 ± 0.001	0.203 ± 0.001
Follower	0.073	0.101	0.126	0.150	0.173
IQL	0.080 ± 0.001	0.127 ± 0.002	$0.188 \pm \ 0.003$	0.257 ± 0.003	0.319 ± 0.002
LSwitcher	0.075 ± 0.001	0.114 ± 0.001	0.149 ± 0.001	0.180 ± 0.001	0.208 ± 0.001
MAMBA	0.073	0.095	0.119 ± 0.001	0.146 ± 0.001	0.176 ± 0.001
MATS-LP	0.110 ± 0.002	0.176 ± 0.002	0.231 ± 0.003	0.274 ± 0.003	0.306 ± 0.003
QMIX	0.074	0.100	0.126	0.150 ± 0.001	0.175 ± 0.001
QPLEX	0.078	0.114 ± 0.001	0.147 ± 0.001	0.176 ± 0.001	0.216 ± 0.002
RHCR	0.088	0.125 ± 0.001	0.170 ± 0.002	0.242 ± 0.004	0.314 ± 0.004
VDN	0.077 ± 0.001	0.109 ± 0.001	0.141 ± 0.001	0.173 ± 0.001	0.204 ± 0.002

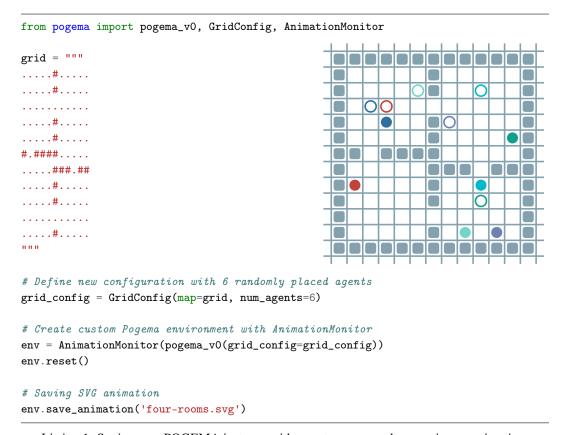
C.6 Pathfinding

Pathfinding metric is tailored to indicate how well the algorithm is able to guide am agent to its goal location. As a result, there is actually no need to run the algorithms on LifeLong instances. Instead, they were run on the same set of instances that were utilized for MAPF approaches. The results of this evaluation are presented in Table 10. Again, the best results were obtained by search-based approach - RHCR. Its implementation was slightly modified to work on MAPF instances, when there is no new goal after reaching the current one. Either optimal or close to optimal paths are able to find Follower and MATS-LP. Followers misses optimal paths due to the integrated technique that changes the edge-costs. MATS-LP adds noise to the root of the search tree that might result in choosing of wrong actions. For the approaches from Switcher family it's actually almost impossible to find optimal paths as they have no information about global map and operate only based on the local observations.

Table 10: Pathfinding results.

Algorithm	Makespan			
ASwitcher	340.56 ± 79.41			
Follower	181.00 ± 20.95			
IQL	1825.95 ± 144.16			
LSwitcher	472.64 ± 119.23			
MAMBA	416.45 ± 136.01			
MATS-LP	179.93 ± 22.45			
QMIX	955.54 ± 200.68			
QPLEX	933.74 ± 199.18			
RHCR	179.82 ± 20.21			
VDN	1733.20 ± 157.96			

D Code examples for POGEMA



Listing 1: Setting up a POGEMA instance with a custom map and generating an animation.

POGEMA is an environment that provides a simple scheme for creating MAPF scenarios, specifying the parameters of GridConfig. The main parameters are: on_target (the behavior of an agent on the target, e.g., restart for LifeLong MAPF and nothing for classical MAPF), seed — to preserve the same generation of the map; agent; and their targets for scenario, size — used for cases without custom maps to specify the size of the map, density — the density of obstacles, num_agents — the

number of agents, obs_radius - observation radius, collision_system - controls how conflicts are handled in the environment (we used a *soft* collision system for all of our experiments). The example of creation such instance is presenten in Listing 1.

Visualization of the agents is a crucial tool for debugging algorithms, visually comparing them, and presenting the results. Many existing MARL environments lack such tools, or have limited visualization functionality, e.g., requiring running the simulator to provide replays, or offering visualizations only in one format (such as videos). In the POGEMA environment, there are three types of visualization formats. The first one is console rendering, which can be used with the default render methods of the environment; this approach is useful for local or server-side debugging. The preferred second option is *SVG* animations. An example of generating such a visualization is presented in the listing above. This approach allows displaying the results using any browser. It provides the ability to highlight high-quality static images (e.g., as the images provided in the paper) or to display results on a website (e. g., animations of the POGEMA repository on GitHub). This format ensures high-quality vector graphics. The third option is to render the results to video format, which is useful for presentations and videos.

E POGEMA Toolbox

The POGEMA Toolbox provides three types of functionality.

The first one is registries to handle custom maps and algorithms. To create a custom map, the user first needs to define it using ASCII symbols or by uploading it from a file, and then register it using the toolbox (see Listing 3). The same approach is used to register and create algorithms (see Listing 2). In that listing, the registration of a simple algorithm is presented, which must includ two methods: act and reset_states. This approach can also accommodate a set of hyperparameters which the Toolbox handles.

```
from pogema import BatchAStarAgent

# Registring A* algorithm
ToolboxRegistry.register_algorithm('A*', BatchAStarAgent)

# Creating algorithm
algo = ToolboxRegistry.create_algorithm("A*")
```

Listing 2: Example of registering the A* algorithm as an approach in the POGEMA Toolbox.

```
from pogema_toolbox.registry import ToolboxRegistry

# Creating cusom_map
custom_map = """
....#.
...#.
...#.
...#.
# Registring custom_map
ToolboxRegistry.register_maps({"custom_map": custom_map})
```

Listing 3: Example of registering a custom map to the Pogema Toolbox.

Second, it provides a unified way of conducting distributed testing using Dask ¹¹ and defined configurations. An example of such a configuration is provided in Listing 4. The configuration is split into three main sections; the first one details the parameters of the POGEMA environment used for

¹¹https://github.com/dask/dask

```
environment: # Configuring Test Environments
  name: Environment
  on_target: 'restart'
  max_episode_steps: 128
  observation_type: 'POMAPF'
  collision_system: 'soft'
  seed:
    grid_search: [ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ]
  num_agents:
    grid_search: [ 8, 16, 24, 32, 48, 64 ]
  map_name:
    grid_search: [
        validation-mazes-seed-000, validation-mazes-seed-001, validation-mazes-seed-002,
        validation-mazes-seed-003, validation-mazes-seed-004, validation-mazes-seed-005,
    ]
algorithms: # Specifying algorithms and it's hyperparameters
  RHCR-5-10:
    name: RHCR
    parallel_backend: 'balanced_dask'
    num_process: 32
    simulation_window: 5
    planning_window: 10
    time_limit: 10
    low_level_planner: 'SIPP'
    solver: 'PBS'
results_views: # Defining results visualization
  01-mazes:
    type: plot
    x: num_agents
    y: avg_throughput
    width: 4.0
    height: 3.1
    line_width: 2
    use_log_scale_x: True
    legend_font_size: 8
    font_size: 8
    name: Mazes
    ticks: [8, 16, 24, 32, 48, 64]
  TabularThroughput:
    type: tabular
    drop_keys: [ seed, map_name]
    print_results: True
```

Listing 4: Example of the POGEMA Toolbox configuration for parallel testing of the RHCR approach and visualization of its results.

testing. It also includes iteration over the number of agents, seeds, and names of the map (which were registered beforehand). The unified grid_search tag allows for the examination of any existing parameter of the environment. The second part of the configurations is a list of algorithms to be tested. Each algorithm has its alias (which will be shown in the results) and name, which specifies

the family of methods. It also includes a list of hyperparameters common to different approaches, e.g., number of processes, parallel backend, etc., and the specific parameters of the algorithm.

The third functionality and third part of the configuration concern views. This is a form of presenting the results of the algorithms. Working with complex testing often requires custom tools for creating visual materials such as plots and tables. The POGEMA toolbox provides such functionality for MAPF tasks out-of-the-box. The listing provides two examples of such data visualization: a plot and a table, which, based on the configuration, provide aggregations of results and present information in a high-quality form, including confidence intervals. The plots and tables in the paper are prepared using this functionality.

F Extended Related Work

StarCraft Multi-Agent Challenge — The StarCraft Multi-Agent Challenge (SMAC) is a highly used benchmark in the MARL community. Most MARL papers that propose new algorithms provide evaluations in this environment. The environment offers a large set of possible tasks where a group of units tries to defeat another group of units controlled by a bot (a predefined programmed policy). Such tasks are partially observable and often require simple navigation. However, the benchmark has several drawbacks, such as the need to use the slow simulator of the StarCraft II engine, deterministic tasks, and the lack of an evaluation protocol.

Nevertheless, some of these drawbacks have already been addressed. SMAX [39] provides a hardware-accelerated JAX version of the environment, but it cannot guarantee full compatibility since the StarCraft II engine is proprietary software. SMAC v2 [13] solves the problem of determinism, highlighting this issue in the original SMAC environments. Moreover, an evaluation protocol for the SMAC environment is proposed in [15]. Despite these efforts, it's hard to say that these tasks require the generalization ability of the agent, since the training and evaluation are conducted on the same scenario.

Multi-agent MuJoCo — In MAMuJoCo, the standard tasks involve agents controlling different sets of joints (or a single joint) within a simulated robot. This set of environments is a natural adaptation of the environment presented in the well-known MuJoCo physics engine [55]. These tasks don't require high generalization abilities or navigation. In the newer version, MuJoCo provides a hardware-accelerated version, forming the basis for Multi-agent BRAX [39], which enhances performance and efficiency.

Google Research Football — Google Research Football [21] is a multi-agent football simulator that provides a framework for cooperative or competitive multi-agent tasks. Despite the large number of possible scenarios in the football academy and the requirement for simple navigation, the tasks are highly specific to the studied domain. Additionally, the number of possible agents is limited. Moreover, the framework offers low scalability, requiring a heavy engine.

Multi-robot warehouse — The multi-robot warehouse environment RWARE [35] simulates a warehouse with robots delivering requested goods. The environment is highly specific to delivery tasks; however, it doesn't support procedurally generated scenarios, thus not requiring generalization abilities or an evaluation protocol. The best-performing solution [8] in this environment is trained on only 4 agents. The benchmark is highly related to multi-agent pathfinding tasks; however, it doesn't provide centralized solution integration, which could serve both as an upper bound for learnable decentralized methods and as a source of expert demonstrations.

Level-Based Foraging — Multi-agent environment LBF [35] simulates food collection by several autonomously navigating agants in a grid world. Each agent is assigned a level. Food is also randomly scattered, each having a level on its own. The collection of food is successful only if the sum of the levels of the agents involved in loading is equal to or higher than the level of the food. The agents are getting rewarded by level of food they collected normalized by their level and overall food level of the episode. The game requires cooperation but also the agents can emerge competitive behavior. The environment is very efficiently designed and very simple to set up; however, it doesn't support procedurally generated scenarios, thus not requiring generalization abilities or an evaluation protocol.

Flatland — The Flatland environment [30] is designed to address the specific problem of fast, conflict-free train scheduling on a fixed railway map. This environment was created for the Flatland Competition [22]. The overall task is centralized with full observation; however, there is an adaptation

to partial observability for RL agents. Unfortunately, during several competitions, despite the presence of stochastic events, centralized solutions [24] from operations research field have outperformed RL solutions by a large margin in both quality and speed. The environment is procedurally generated, which requires high generalization abilities, and the benchmark provides an evaluation protocol. A significant drawback is the extremely slow speed of the environment, which highly restricts large-scale learning.

Overcooked — The Overcooked is a benchmark environment for fully cooperative human-AI task performance, based on the widely popular video game [7]. In the game, agents control chefs tasked to cook some dishes. Due to possible complexity of the cooking process, involving multistep decision-making, it requires emergence of cooperative behaviour between the agents.

Griddly — This is a grid-based game engine [4], allowing to make various and diverse grid-world scenarios. The environment is very performance efficient, being able to make thousands step per second. Moreover, there is test coverage and continuous integration support, allowing open-source development. The engine provides support for different observation setups and maintains state history, making it useful for search based methods.

Multi-player game simulators — Despite the popularity of multi-player games, it's a challenging problem to develop simulators of the games that could be used for research purposes. One of the most popular adaptations are MineCraft MALMO [19] that allows to utilise MineCraft as a configurable research platform for multi-agent research and model various agents' interactions. In spite of the game's flexible functionality, it depends on external runtime, so might be very hard to set up or extremely slow to iterate with. That's why there are several alternatives that prioritise fast iteration over the environment complexity, like Neural MMO [51] that models a simple MMO RPG with agents with a shared resource pool. On top of that, there are even faster implementation, targeting coordination or cooperation, like Hide-and-Seek [3], modelling competition, or GoBigger [69], focusing on competition between cooperating populations.

Multi-agent Driving Simulators — Autonomous driving is one of the important practical applications of MARL, and Nocturne [57] is a 2D simulator, written in C++, that focuses on different scenarios of interactions — e.g. intersections, roundabouts etc. The simulator is based on trajectories collected in real life, so allows modelling practical scenarios. This environment has evaluation protocols and supports open-source development with continuous integration and covered by tests. There are also environments, focusing on particular details of driving, for example, Multi-car Racing [44] that represents racing from bird's eye view.

Suits of multi-agent environments — These multi-agent environments are designed to be very simple benchmarks for specific tasks. Jumanji [5] is a set of environments for different multi-agent scenarios connected to combinatorial optimization and control, for example, routing or packing problems. With the purpose for each environment to be focused on the particular task, the overall suit doesn't test generalization or enable procedural generation. Multi Particle Environments (MPE) [28] is a communication oriented set of partially observable environments where particle agents are able to interact with fixed landmarks and each other, communicating with each other. SISL [16] is a set of three dense reward environments was developed to have a simple benchmark for various cooperative scenarios. For environment suits, testing generalization, MeltingPot [1] comes into place. This set of the environments contains a diverse set of cooperative and general-sum partially observable games and maintains two populations of agents: focal (learning) and visiting (unknown to the environment) to benchmark generalization abilities of MARL algorithms. The set in based on the own game engine and might be extended quite easily.

Real-world Engineering in Practice — Real-world engineering tasks can often be addressed by sophisticated MARL solutions. IMP-MARL [23] provides a platform for evaluating the scalability of cooperative MARL methods responsible for planning inspections and repairs for specific system components, with the goal of minimizing maintenance costs. At the same time, agents must cooperate to minimize the overall risk of system failure. MATE [34] addresses target coverage control challenges in real-world scenarios. It presents an asymmetric cooperative-competitive game featuring two groups of learning agents, cameras and targets, each with opposing goals.

G Examples of Used Maps

The examples of used maps are presented in Figure 9, showing a diverse list of maps. The map types used in the POGEMA Benchmark include: Maze, with prolonged 1-cell width corridors requiring high-level cooperation; Random, easily generated maps to avoid overfitting with controllable obstacle density; MovingAI-tiles, smaller modified slices of MovingAI maps; Puzzle, small hand-crafted maps with challenging patterns necessitating agent cooperation; Warehouse, used in LifeLong MAPF research, featuring high agent density and throughput challenges; and MovingAI, benchmark maps with varying sizes and structures for single-agent pathfinding.

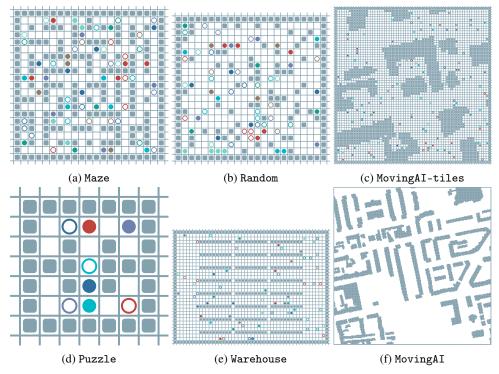


Figure 9: Examples of maps presented in the POGEMA Benchmark. The MovingAI map (on which the pathfinding metric was tested) is shown without grid lines and agents for clarity.

H MARL Training Setup

For training MARL approaches, such as MAMBA, QMIX, QPLEX, and VDN, we used the default hyperparameters provided in the corresponding repositories ¹², and employed the PyMARL2 framework ¹³ to establish MARL baselines. As input, we apply preprocessing from the Follower approach, which is the current state-of-the-art for decentralized LifeLong MAPF. We attempted to add a ResNet encoder, as used in the Follower approach; however, this addition worsened the results, thus we opted for vectorized observation and default MLP architectures. For centralized methods that work with the state of the environment (e.g., QMIX or QPLEX), we utilized the MARL integration of POGEMA, which provides agent positions, targets, and obstacle positions in a format similar to the SMAC environment (providing their coordinates).

Our initial experiments on training this approach with a large number of agents, similar to the Follower model, showed very low results. We adjusted the training maps to be approximately 16×16 , which proved to be more effective and populated them with 8 agents. This setup shows better results. We continued training the approaches until they reached a plateau, which for most algorithms is under 1 million steps.

¹²https://github.com/jbr-ai-labs/mamba

¹³https://github.com/hijkzzz/pymarl2

I Resources and Statistics

To evaluate all the presented approaches integrated with POGEMA we have used two workstations with equal configuration, that includes 2 NVidia Titan V GPU, AMD Ryzen Threadripper 3970X CPU and 256 GB RAM. The required computation time is heavily depends on the approach by itself.

Table 11: Total time (in hours) required by each of the algorithms to run all MAPF instances on the corresponding datasets.

	Random	Mazes	Warehouse	MovingAI-tiles	Puzzles	MovingAI
DCC	2.11	2.46	11.07	22.70	0.09	0.02
IQL	0.05	0.04	0.13	0.13	0.01	0.01
LaCAM	0.20	0.29	0.24	0.23	0.37	0.01
MAMBA	6.62	6.47	8.36	12.27	2.59	3.40
QMIX	0.04	0.04	0.14	0.13	0.01	0.01
QPLEX	0.05	0.04	0.13	0.13	0.01	0.01
SCRIMP	1.66	2.20	16.54	21.63	0.08	0.21
VDN	0.05	0.04	0.13	0.13	0.01	0.01

Table 12: Total time (in hours) required by each of the algorithms to run all LMAPF instances on the corresponding datasets.

	Random	Mazes	Warehouse	MovingAI-tiles	Puzzles	MovingAI
ASwitcher	1.03	0.47	2.95	1.76	0.31	0.04
Follower	0.48	0.23	0.69	0.77	0.26	0.89
IQL	0.08	0.04	0.26	0.24	0.02	0.01
LSwitcher	6.18	2.61	17.30	10.70	0.81	0.21
MAMBA	13.82	6.69	15.81	11.07	7.83	3.40
MATS-LP	77.31	35.34	163.68	129.78	3.80	0.14
QMIX	0.08	0.04	0.26	0.25	0.02	0.01
QPLEX	0.08	0.04	0.26	0.25	0.02	0.01
RHCR	0.57	0.25	17.04	6.28	0.01	0.01
VDN	0.08	0.04	0.25	0.25	0.02	0.01

The statistics regarding the spent time on solving MAPF and LMAPF instances are presented in Table 11 and Table 12 respectively. Please note, that all these approaches were run in parallel in multiple threads utilizing dask, that significantly reduces the factual spent time.

We used pretrained models for all the hybrid methods, such as Follower, Switcher, MATS-LP, SCRIMP, and DCC, thus, no resources were spent on their training. RHCR and LaCAM are pure search-based planners and do not require any training. MARL methods, such as MAMBA, QPLEX, QMIX, IQL, and VDN, were trained by us. MAMBA was trained for 20 hours on the MAPF instances, resulting in 200K environment steps, and for 6 days on LifeLong MAPF instances, resulting in 50K environment steps, which corresponds to the same amount of GPU hours. For MARL approaches, we trained them for 1 million environment steps, which corresponds to an average of 5 GPU hours for each algorithm.

J Community Engagement and Framework Enhancements

Our team is committed to maintaining an open and accountable POGEMA framework. Since 2021, we have continuously improved POGEMA, including the addition of the POGEMA Toolbox and the recent introduction of POGEMA Benchmark. We ensure transparency in our operations and encourage the broader AI community to participate. Our framework includes a fast learning environment, problem instance generator, visualization toolkit, and automated benchmarking tools, all guided by a clear evaluation protocol. We have also implemented and evaluated multiple strong baselines that simplify further comparison. We practice rigorous software testing and conduct regular code reviews. We promptly address issues reported on GitHub and welcome any feedback and contributions through GitHub.