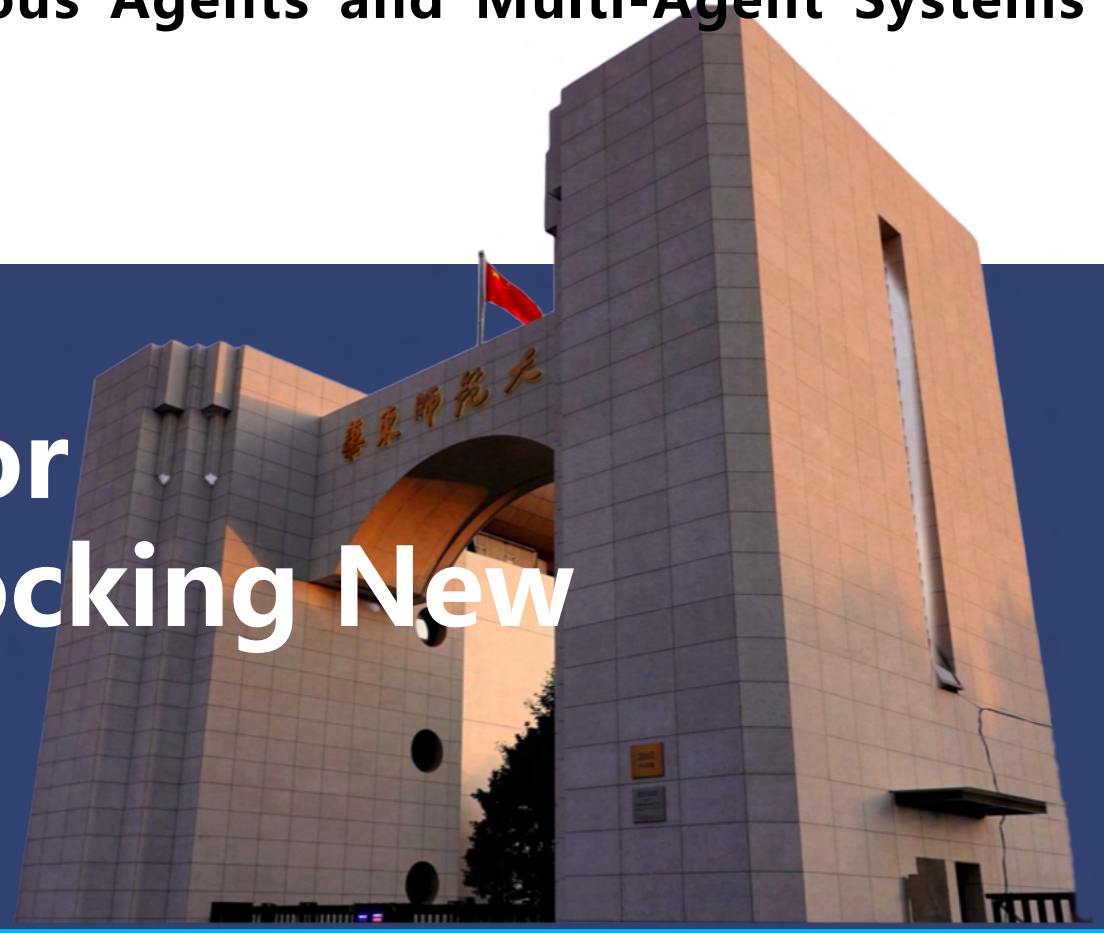
The 23rd International Conference on Autonomous Agents and Multi-Agent Systems

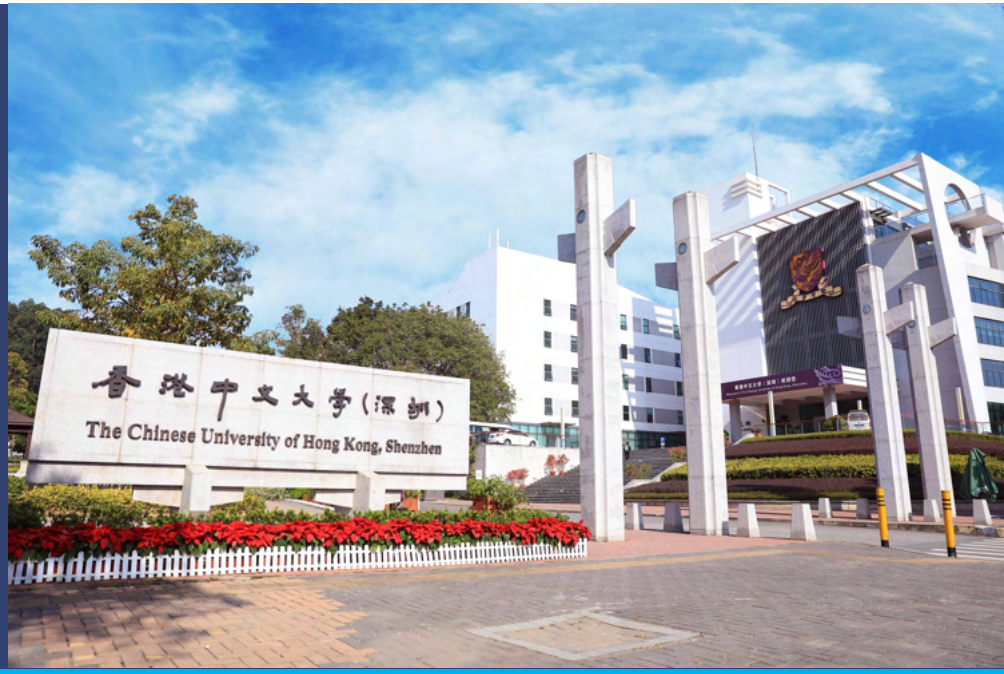


Reinforcement Learning for Operations Research: Unlocking New Possibilities (Part III)

Xiangfeng Wang, Junjie Sheng (East China Normal University)
Wenhao Li (The Chinese University of Hong Kong, Shenzhen)
Contact email: liwenhao@cuhk.edu.cn

The 23rd International Conference on Autonomous Agents and Multi-Agent Systems

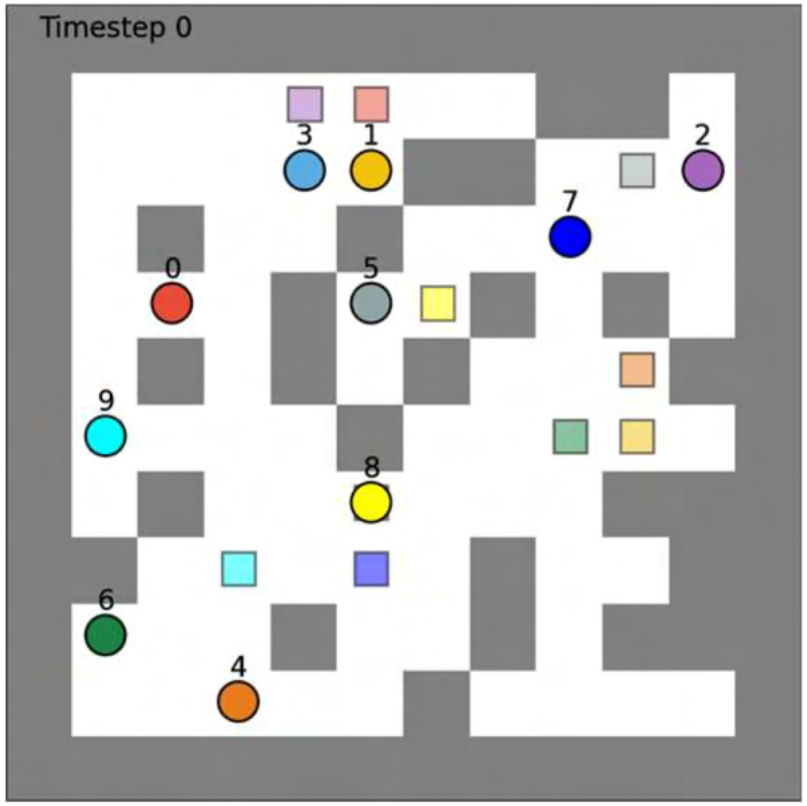
Reinforcement Learning for Multi-Agent Pathfinding



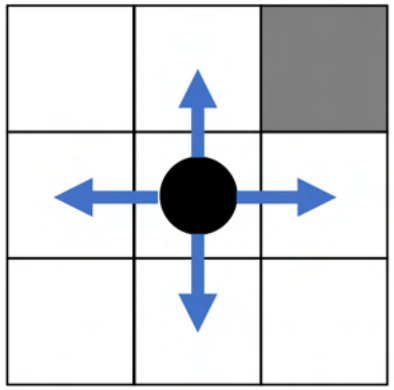
Xiangfeng Wang, Junjie Sheng (East China Normal University)
Wenhao Li (The Chinese University of Hong Kong, Shenzhen)
Contact email: liwenhao@cuhk.edu.cn

Multi-Agent Pathfinding (MAPF)

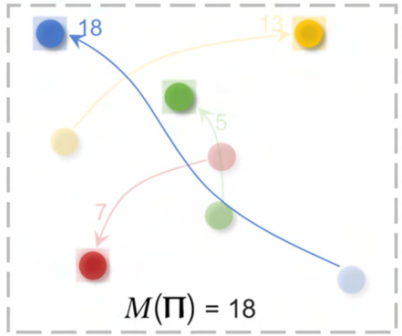
Optimization problem with the objective to minimize task-completion time (called makespan) or the sum of travel times (called flowtime)

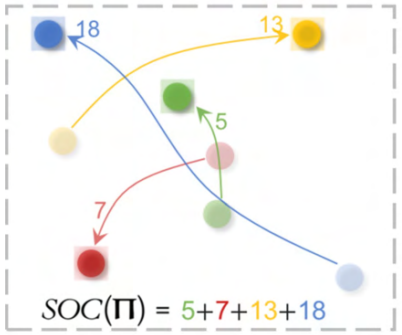


makespan



flowtime

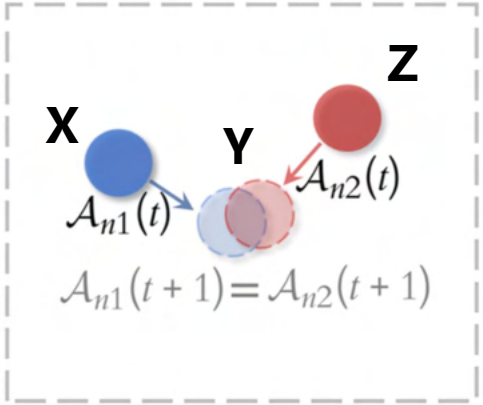


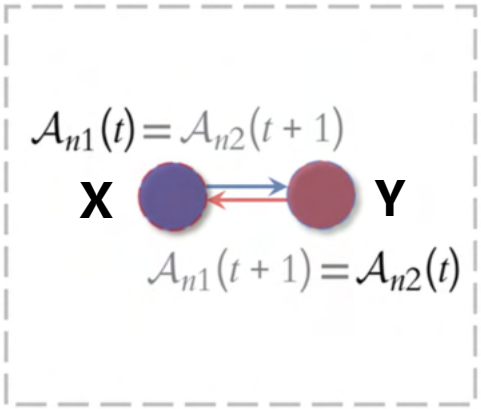


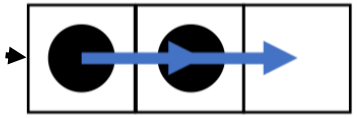
1. Jiaoyang Li, et al. "AAAI-22 Tutorial on Recent Advances in Multi-Agent Path Finding."
2. Chung, Jaehoon, et al. "Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding." Artificial Intelligence Review 57.2 (2024): 41.

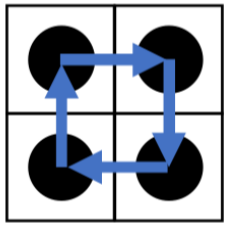
Multi-Agent Pathfinding: Assumptions

- Each agent moves N, E, S or W to an adjacent unblocked cell or waits, in unit time.
- Not allowed ("vertex collision") X Y Z
  - Agent 1 moves from X to Y
  - Agent 2 moves from Z to Y
- Not allowed ("edge collision")
  - Agent 1 moves from X to Y
  - Agent 2 moves from Y to X
- Allowed



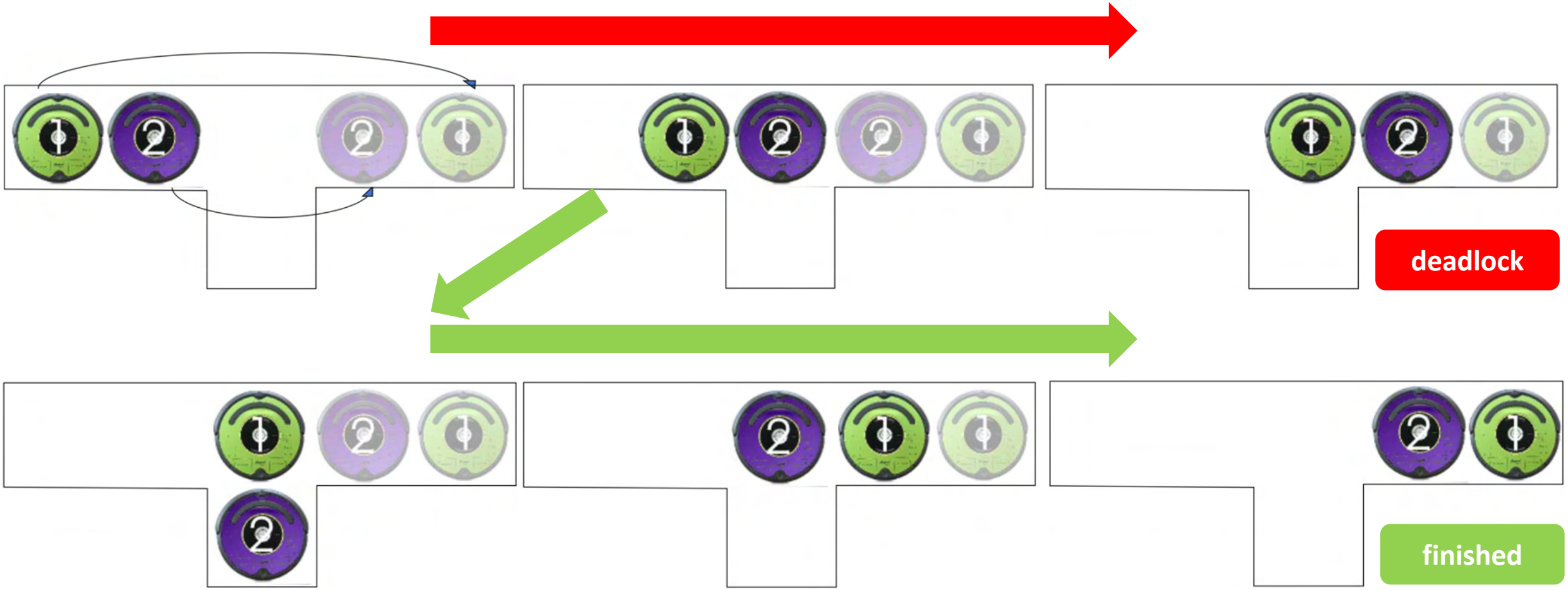






1. Jiaoyang Li, et al. "AAAI-22 Tutorial on Recent Advances in Multi-Agent Path Finding."
2. Chung, Jaehoon, et al. "Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding." Artificial Intelligence Review 57.2 (2024): 41.

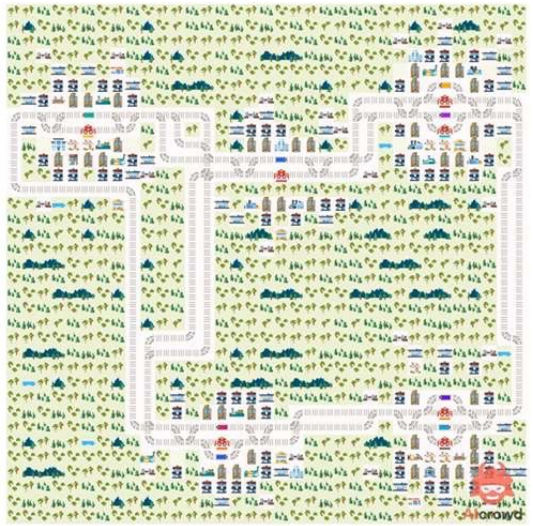
Multi-Agent Pathfinding: Toy Example



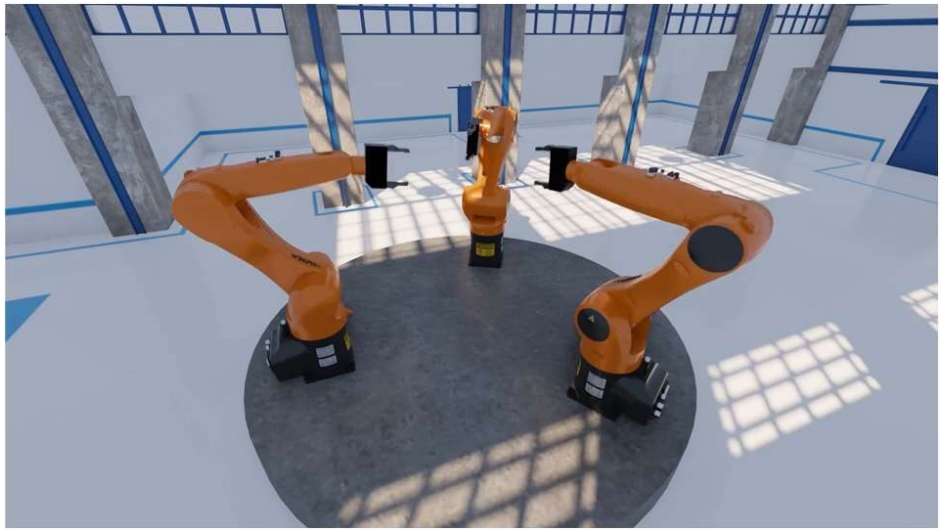
1. Jiaoyang Li, et al. "AAAI-22 Tutorial on Recent Advances in Multi-Agent Path Finding."

Multi-Agent Pathfinding: Applications

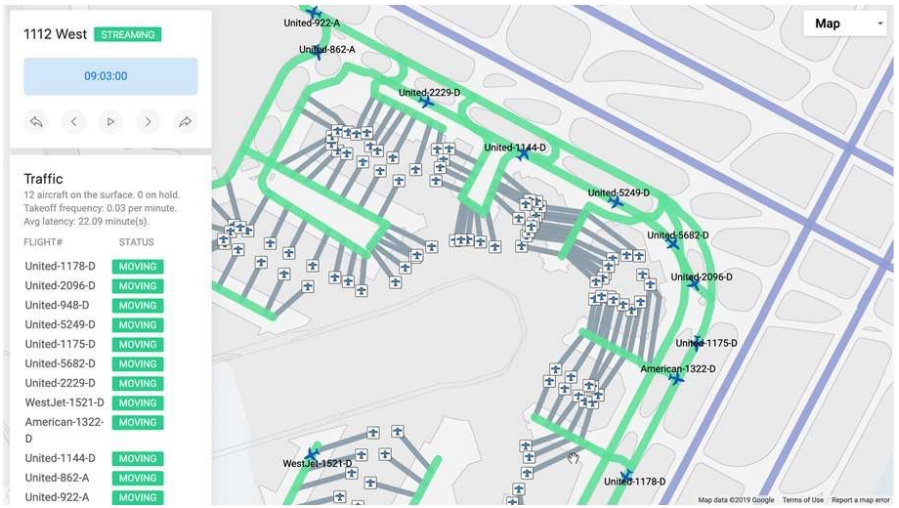
Automated Warehousing



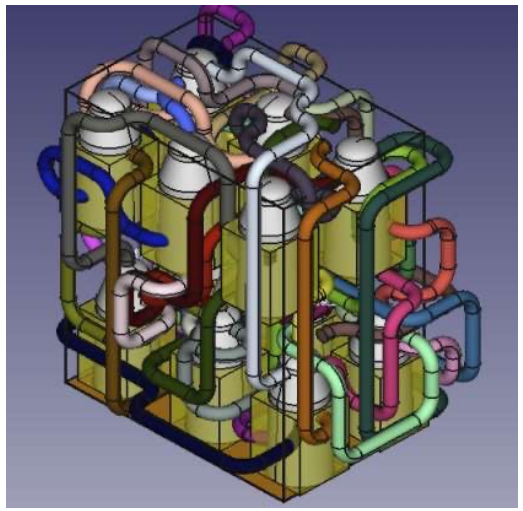
Rail Scheduling



Multi-Arm Assembly



Airport Surface Operation



Pipe Routing

Moving in Formation for Video Games

1. Jiaoyang Li, et al. "AAAI-22 Tutorial on Recent Advances in Multi-Agent Path Finding."

Multi-Agent Pathfinding: Complexity

- **Suboptimal MAPF planning** (= finding any MAPF plan)
  - Undirected graphs
    - **Polynomial time** to find a suboptimal MAPF plan [1]
  - Directed graphs
    - **NP-hard** to find a suboptimal MAPF plan on directed graphs [2]
    - **Polynomal time** to find a suboptimal MAPF plan on directed graphs that are **strongly biconnected and have at least two unoccupied vertices** [3]
- **Optimal MAPF planning**
  - **Polynomial** time to find a **makespan-optimal** MAPF plan with **anonymous agents** (= assignable goal locations) [4]
  - **NP hard** to find a makespan- or flow-time optimal MAPF plan [5]

[1] J. Yu and D. Rus, “Pebble Motion on Graphs with Rotations: Efficient Feasibility Tests and Planning Algorithms”, WAFR, 2014.
[2] B. Nebel, “On the Computational Complexity of Multi-Agent Pathfinding on Directed Graphs”, ICAPS, 2020.
[3] A. Botea et al., “Solving Multi-agent Path Finding on Strongly Biconnected Digraphs”, Journal of Artificial Intelligence Research,
[4] J. Yu and S. LaValle, “Multi-Agent Path Planning and Network Flow”, WAFR, 2012.
[5] J. Yu and S. LaValle, “Structure and Intractability of Optimal Multi-Robot Path Planning on Graphs”, AAAI, 2013.

Multi-Agent Pathfinding: Complexity

Classic solvers

(e.g., A*、M*、integer programming)

(sub)-optimal solution

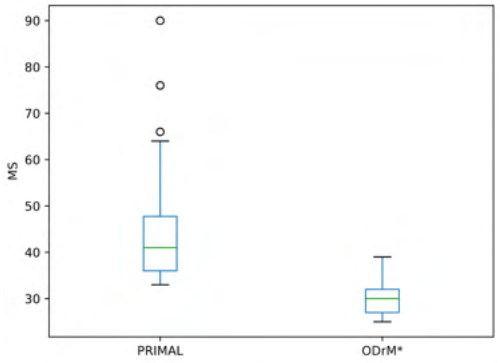
It takes a long time

Greedy methods

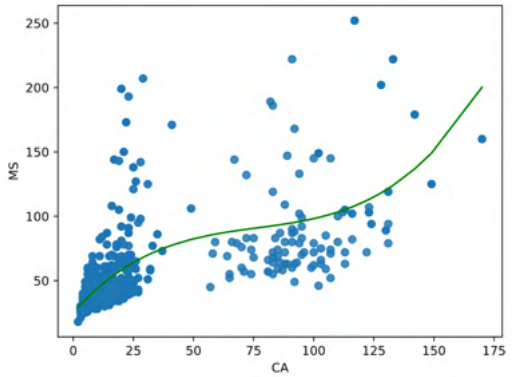
(e.g., data-driven, end-to-end learning)

scaling law, real-time

Deadlocks, congestions



(a) Makespans.



(b) Collisions vs. Makespans.

Wenhao, Li et al. Multi-Agent Path Finding with Prioritized Communication Learning. ICRA 2022

Multi-Agent Pathfinding: Complexity

Classic solvers

(e.g., A*、M*、integer programming)

(sub)-optimal solution

It takes a long time

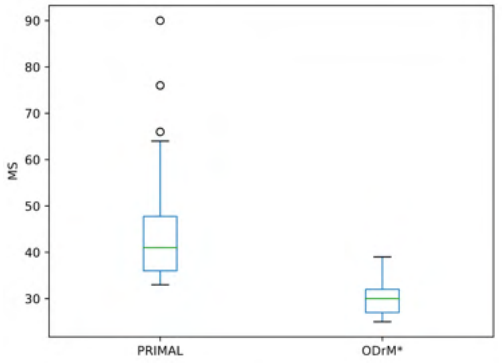
+

Greedy methods

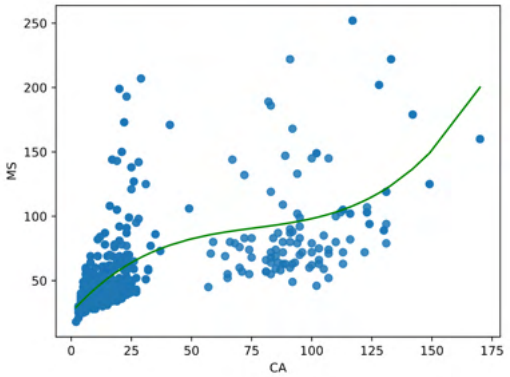
(e.g., data-driven, end-to-end learning)

scaling law, real-time

Deadlocks, congestions



(a) Makespans.



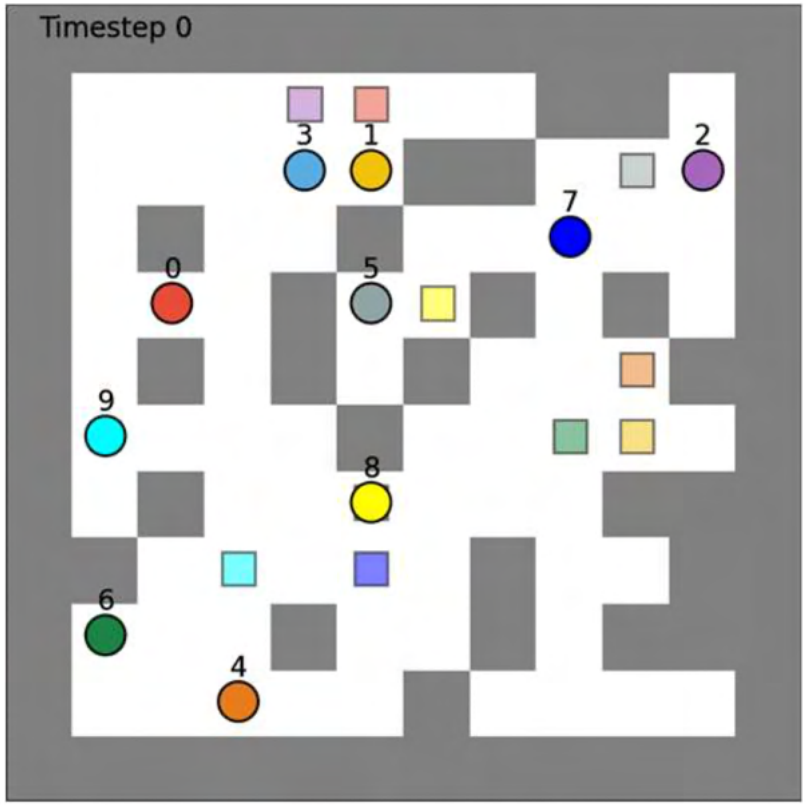
(b) Collisions vs. Makespans.

Wenhao, Li et al. Multi-Agent Path Finding with Prioritized Communication Learning. ICRA 2022

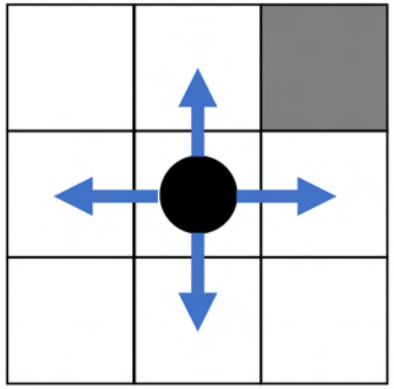
Multi-Agent Pathfinding: MDP Formulation

Optimization problem with the objective to minimize task-completion time (called makespan) or the sum of travel times (called flowtime)

state



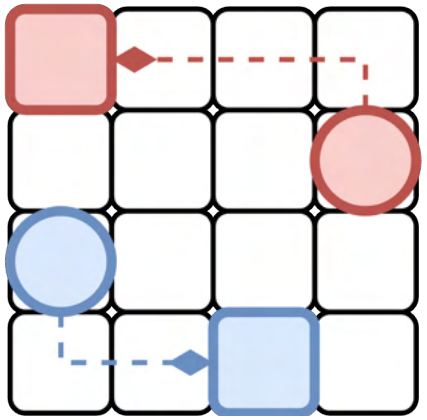
reward



action

Problem Settings

Discrete MAPF



PRIMAL (Sartoretti et al. 2019)

MAPPER (Liu et al. 2020)

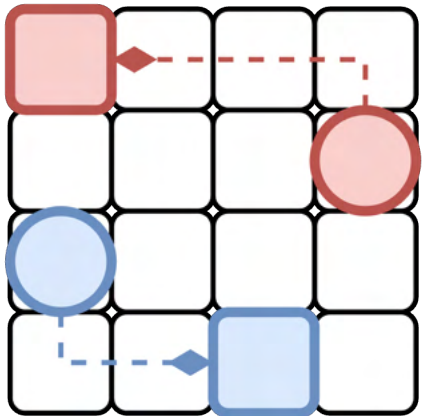
DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

Problem Settings

Discrete MAPF



PRIMAL (Sartoretti et al. 2019)

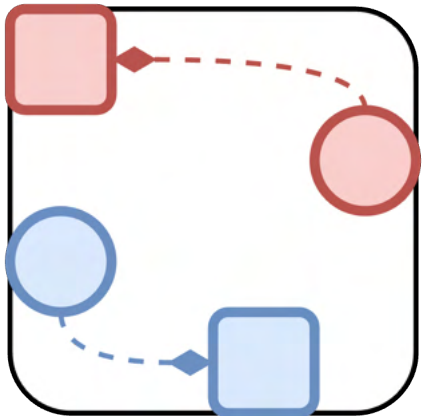
MAPPER (Liu et al. 2020)

DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

Continuous MAPF

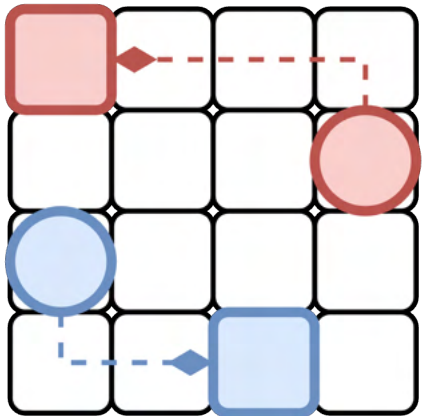


(Qiu et al. 2020)

(Fan et al. 2020)

Problem Settings

Discrete MAPF



PRIMAL (Sartoretti et al. 2019)

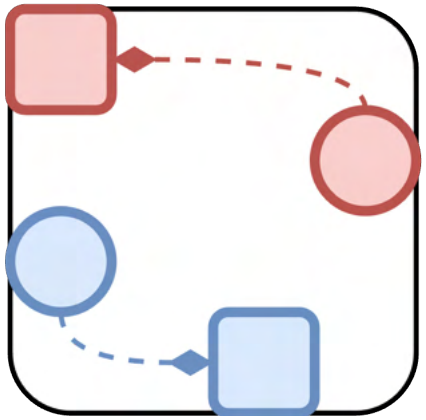
MAPPER (Liu et al. 2020)

DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

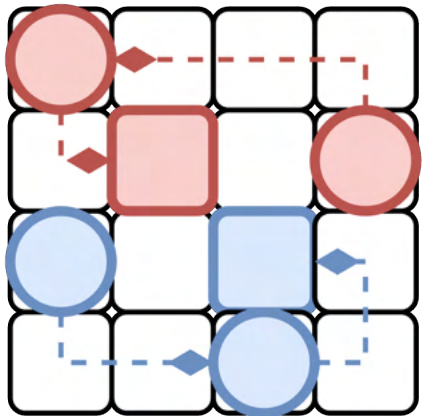
Continuous MAPF



(Qiu et al. 2020)

(Fan et al. 2020)

Lifelong MAPF

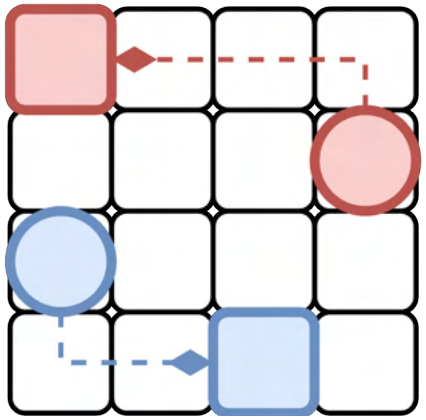


PRIMAL2 (Damani et al. 2021)

(Chen et al. 2023)

Problem Settings

Discrete MAPF



PRIMAL (Sartoretti et al. 2019)

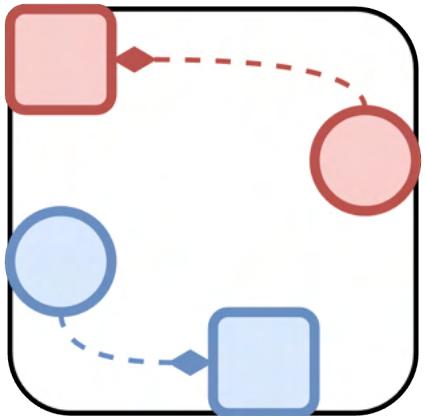
MAPPER (Liu et al. 2020)

DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

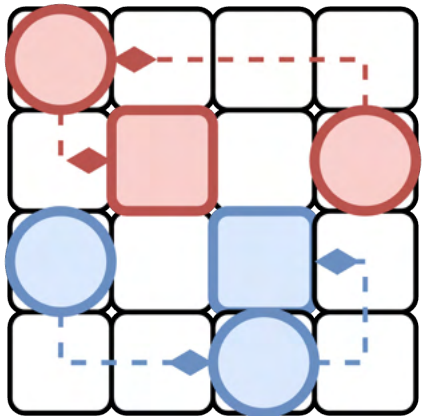
Continuous MAPF



(Qiu et al. 2020)

(Fan et al. 2020)

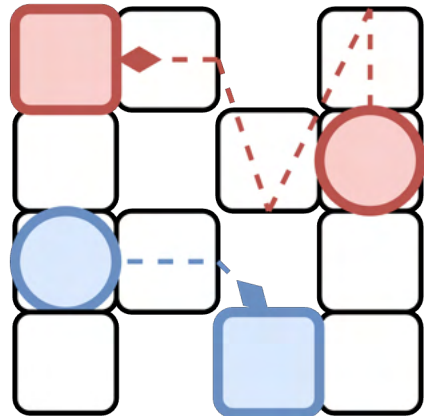
Lifelong MAPF



PRIMAL2 (Damani et al. 2021)

(Chen et al. 2023)

Graphic MAPF

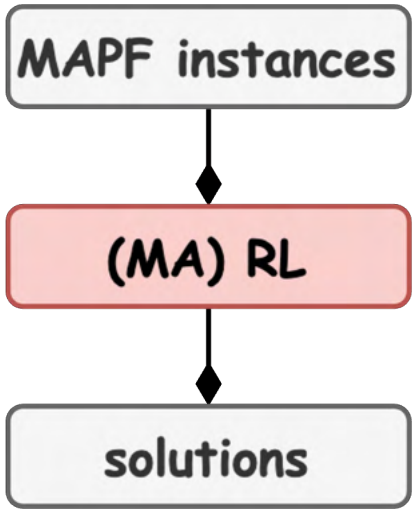


Flatland-RL (Mohanty et al. 2020)

(Knippenberget al. 2021)

Reinforcement Learning Paradigms

(MA)RL as the solver



PRIMAL (Sartoretti et al. 2019)

MAPPER (Liu et al. 2020)

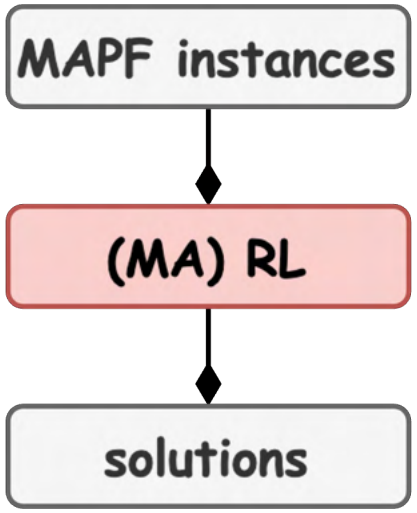
DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

Reinforcement Learning Paradigms

(MA)RL as the solver



PRIMAL (Sartoretti et al. 2019)

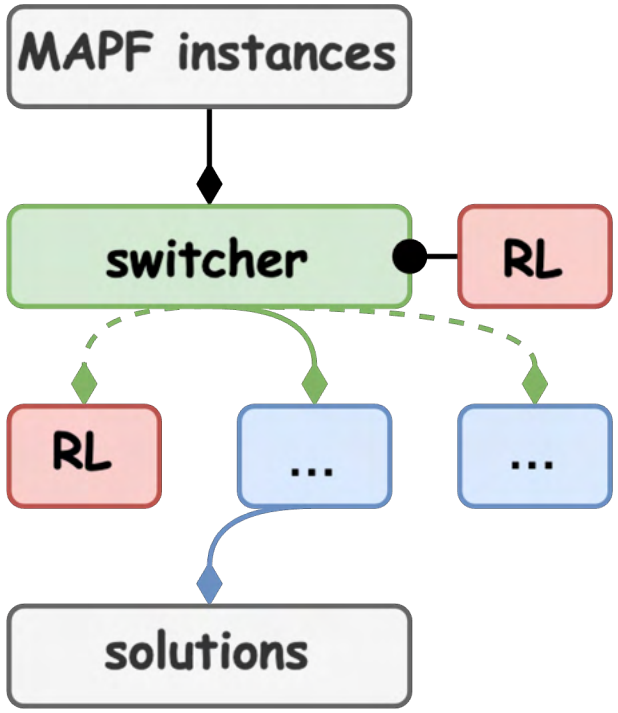
MAPPER (Liu et al. 2020)

DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

(MA)RL as the candidate



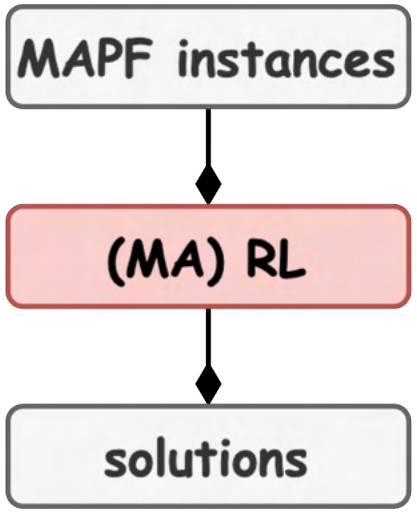
(Qiu et al. 2020)

(Skrynnik et al. 2023)

(Chen et al. 2023)

Reinforcement Learning Paradigms

(MA)RL as the solver



PRIMAL (Sartoretti et al. 2019)

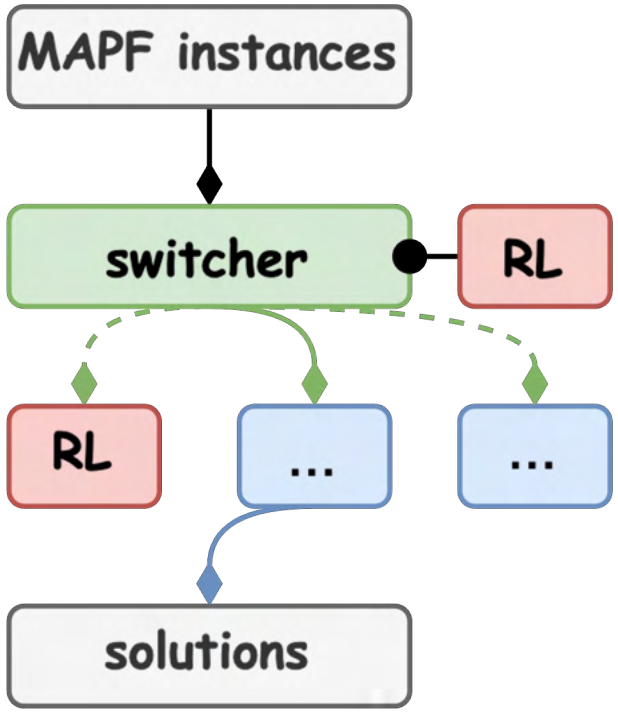
MAPPER (Liu et al. 2020)

DHC (Ma et al. 2021)

PICO (Li et al. 2022)

SCRIMP (Wang et al. 2023)

(MA)RL as the candidate

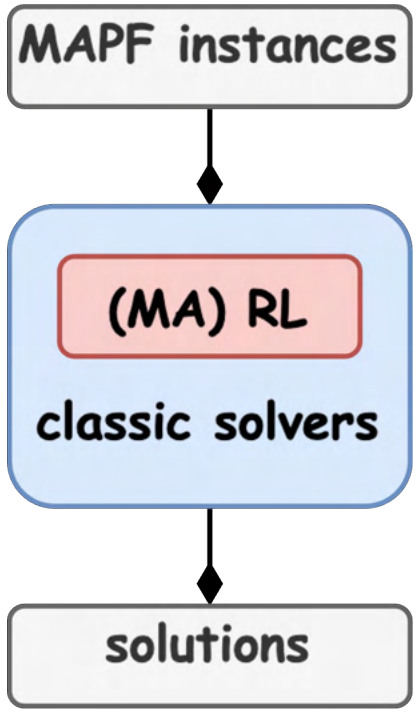


(Qiu et al. 2020)

(Skrynnik et al. 2023)

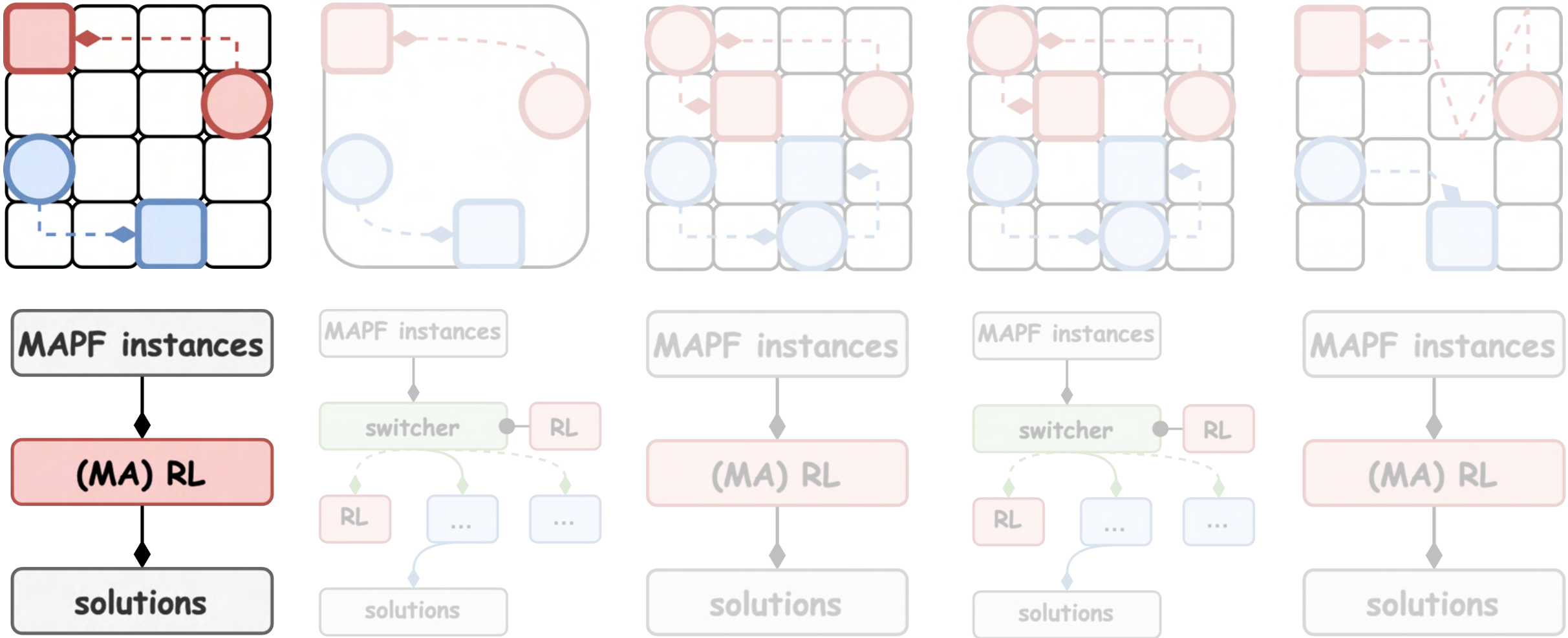
(Chen et al. 2023)

(MA)RL as the assistant

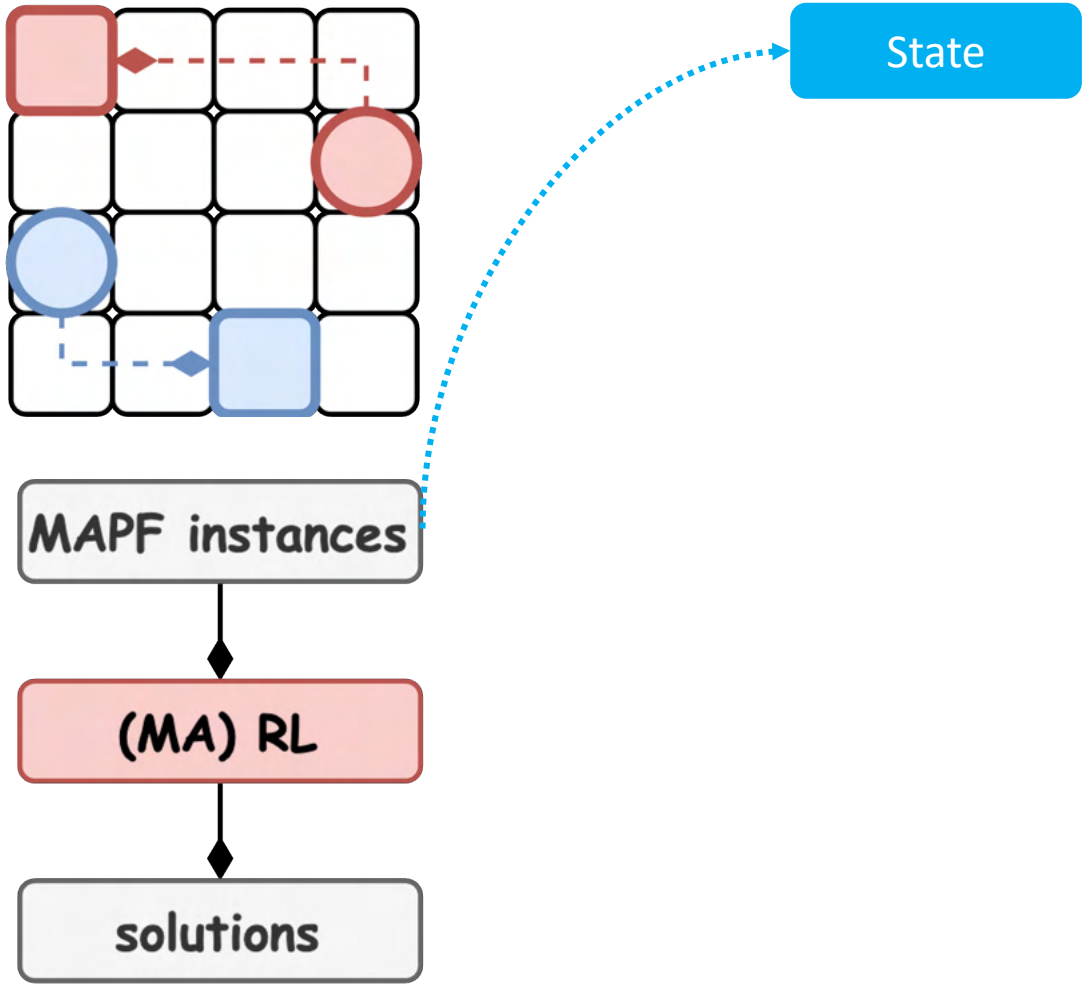


S2AN (Yang et al. 2024)

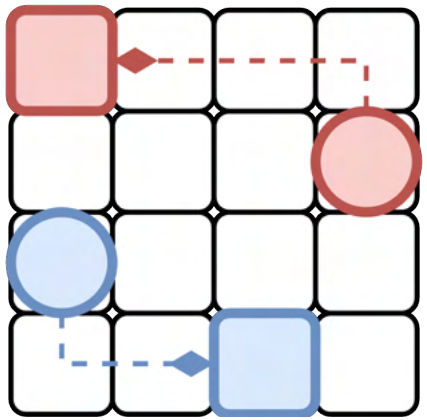
Reinforcement Learning as the Discrete MAPF Solver

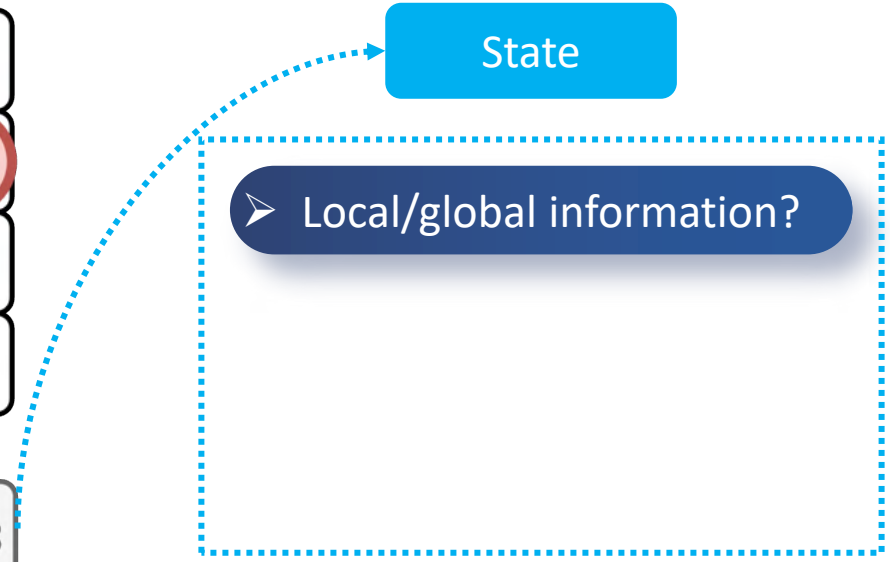


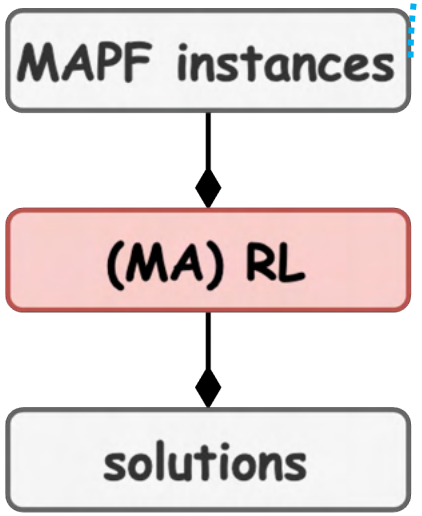
Reinforcement Learning as the Discrete MAPF Solver



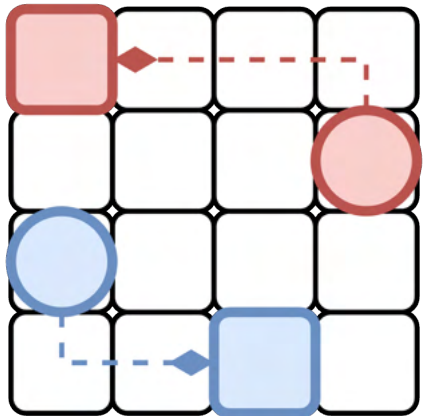
Reinforcement Learning as the Discrete MAPF Solver







Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

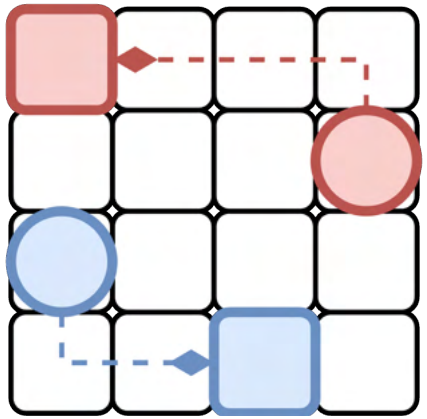
(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

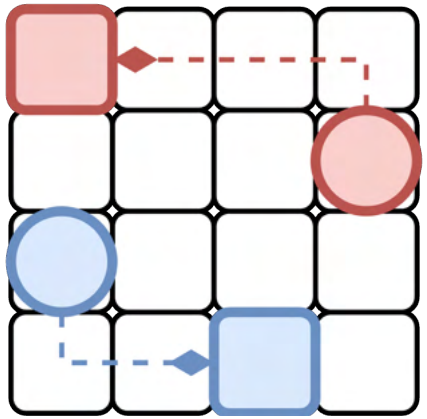
(MA) RL

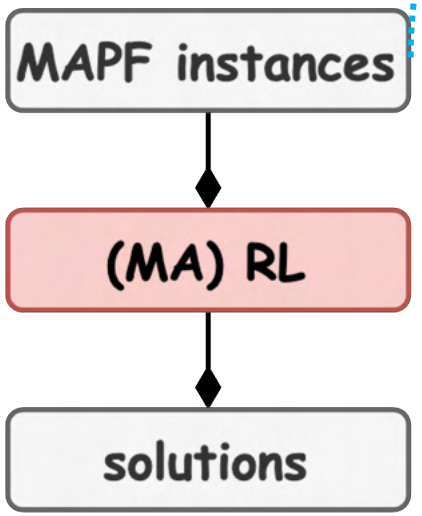
solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?

Reinforcement Learning as the Discrete MAPF Solver

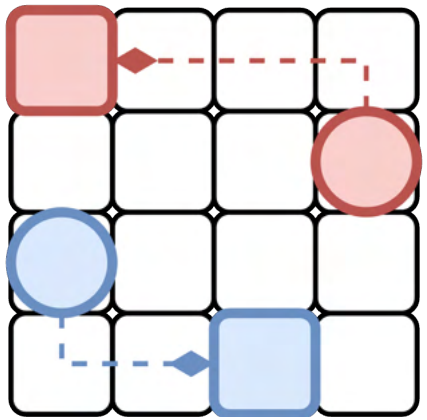




State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

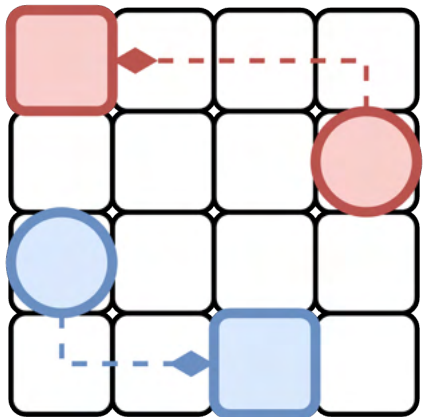
solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reward

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

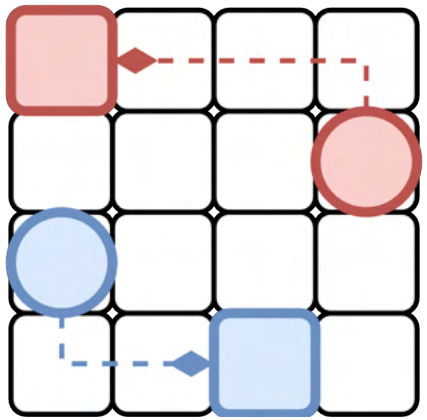
State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reward

- Success or not?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

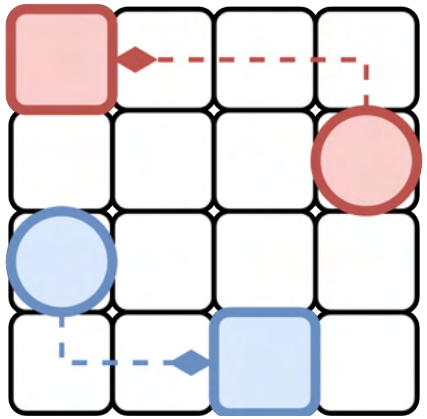
State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reward

- Success or not?
- Cooperation?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

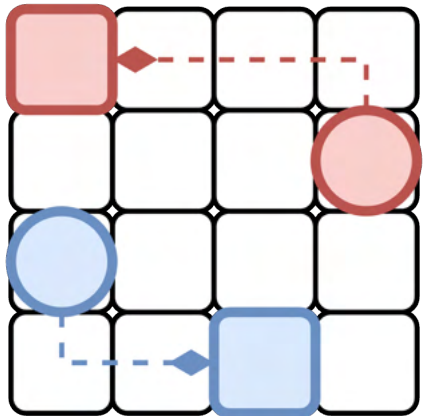
State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

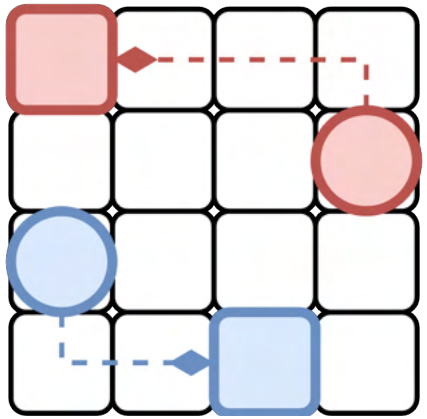
- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

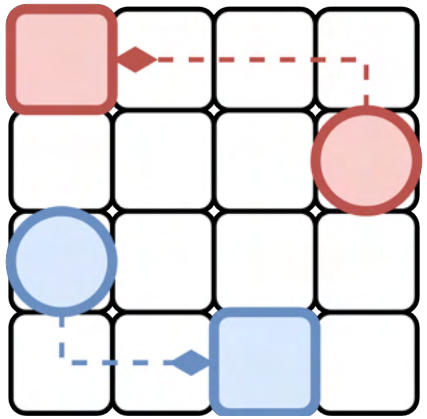
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

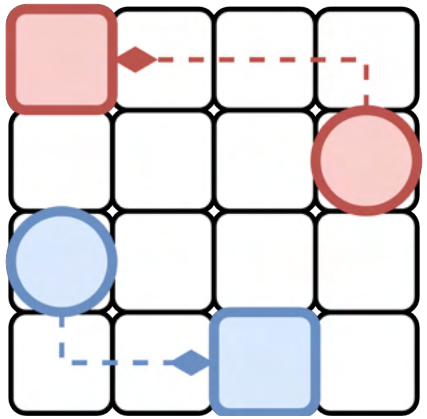
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

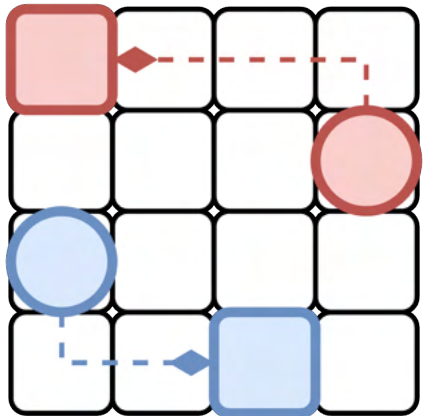
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

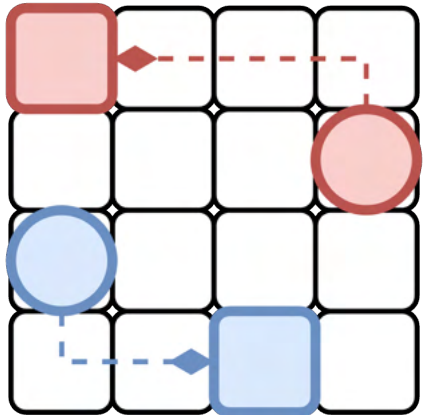
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

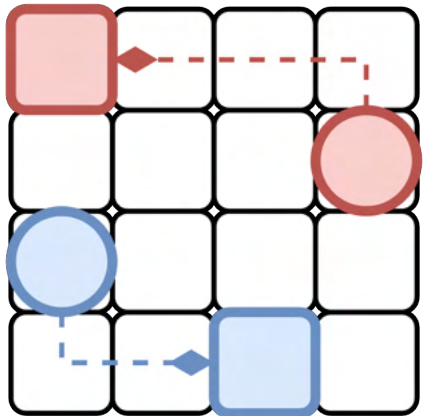
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

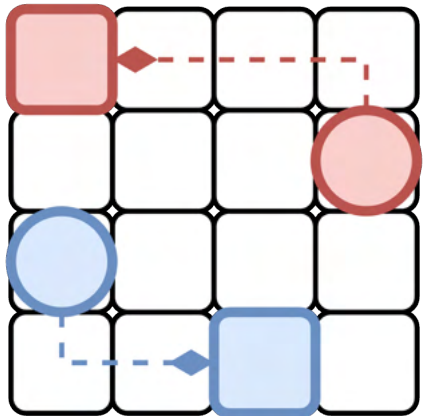
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?
- + evolution computation?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

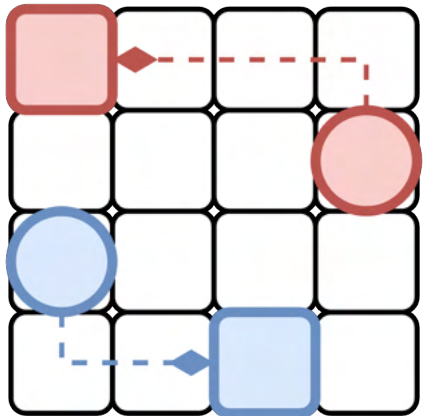
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?
- + evolution computation?
- + curriculum learning?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

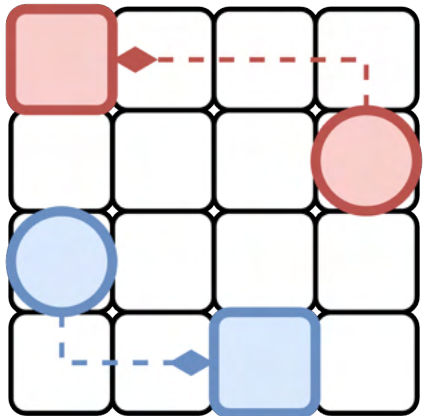
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?
- + evolution computation?
- + curriculum learning?
- + representation learning?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

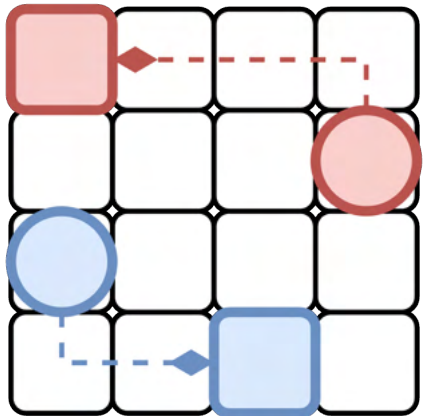
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?
- + evolution computation?
- + curriculum learning?
- + representation learning?
- + postprocessing?

Reinforcement Learning as the Discrete MAPF Solver



MAPF instances

(MA) RL

solutions

State

- Local/global information?
- Static/dynamic information?
- Heuristic knowledge?
- Expert demonstrations?

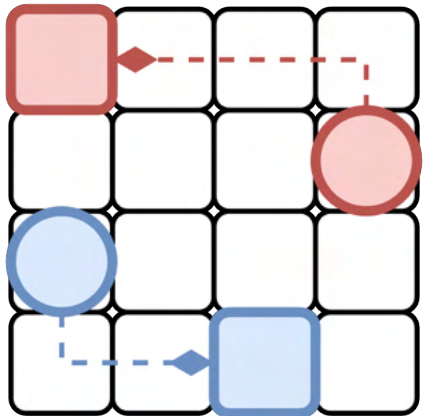
Reward

- Success or not?
- Cooperation?
- Expert knowledge?

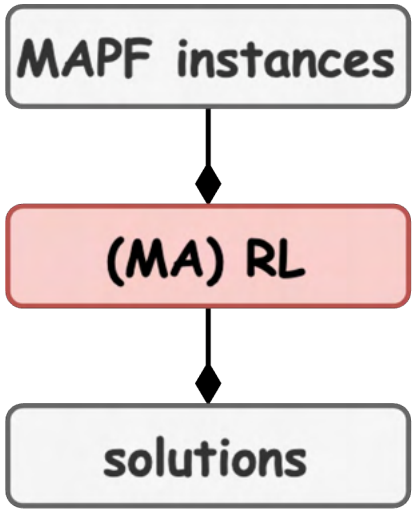
Learning

- Independent learning?
- + imitation learning?
- + communication?
- + cooperation?
- + hierarchical learning?
- + evolution computation?
- + curriculum learning?
- + representation learning?
- + postprocessing?

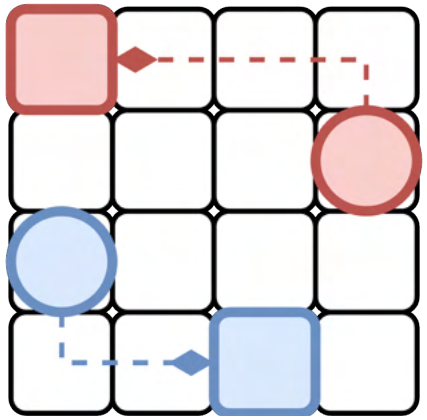
Roadmap



State



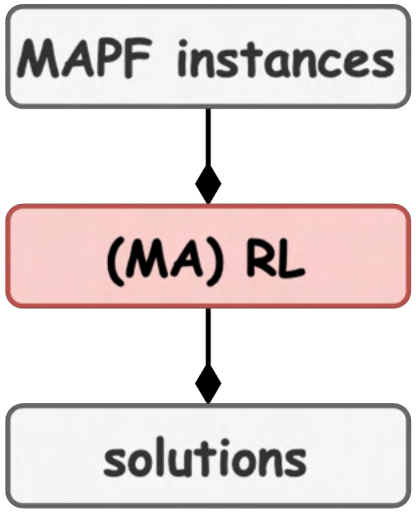
Roadmap



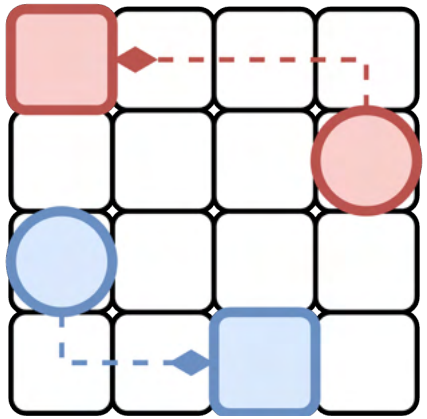
State

local, static

all of them



Roadmap



MAPF instances

(MA) RL

solutions

State

local, static

all of them

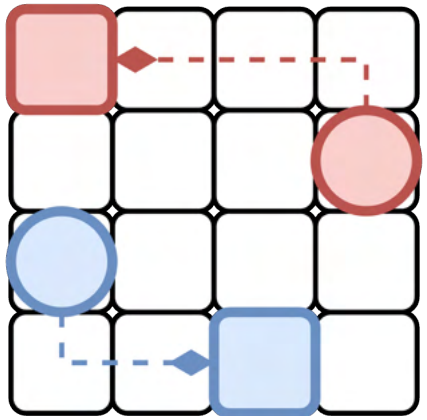
+ expert

MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

Roadmap



MAPF instances

(MA) RL

solutions

State

local, static

all of them

+ expert

MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

+ dynamic

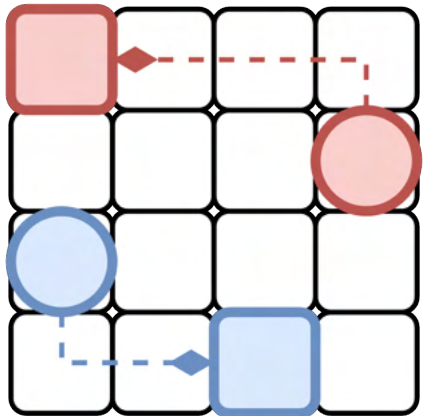
MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

ALPHA (He et al. 2023)

Roadmap



MAPF instances

(MA) RL

solutions

State

local, static

all of them

+ expert

MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

+ dynamic

MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

ALPHA (He et al. 2023)

+ heuristic

DHC (Ma et al. 2021)

DCC (Ma et al. 2021)

Chen et al. 2023a/b

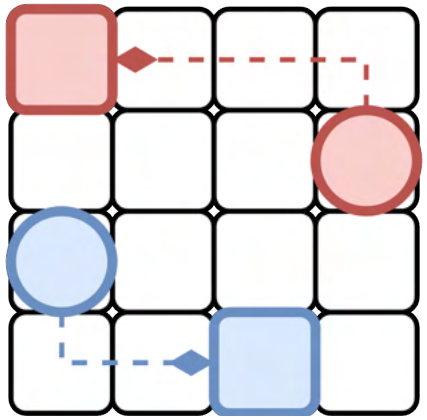
SCRIMP (Wang et al. 2023)

SACHA (Lin et al. 2023)

HELSA (Song et al. 2023)

Cheng et al. 2023

Roadmap



MAPF instances

(MA) RL

solutions

State

local, static — all of them

+ expert — MAPPER (Liu et al. 2020) | G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021)

+ dynamic — MAPPER (Liu et al. 2020) | G2RL (Wang et al. 2020)

AB-MAPPER (Guan et al. 2021) | ALPHA (He et al. 2023)

+ heuristic — DHC (Ma et al. 2021) | DCC (Ma et al. 2021) | Chen et al. 2023a/b

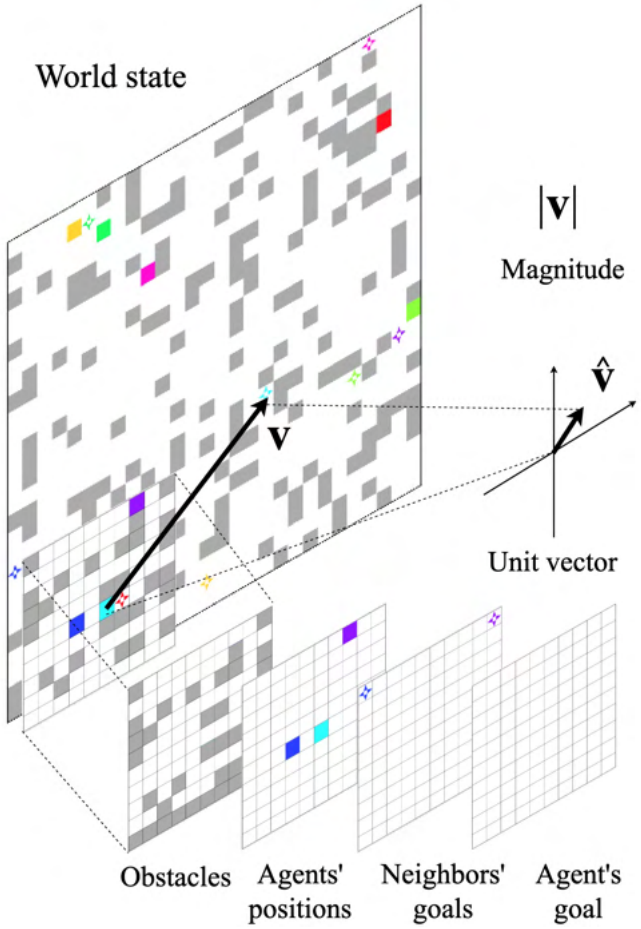
SCRIMP (Wang et al. 2023) | SACHA (Lin et al. 2023)

HELSA (Song et al. 2023) | Cheng et al. 2023

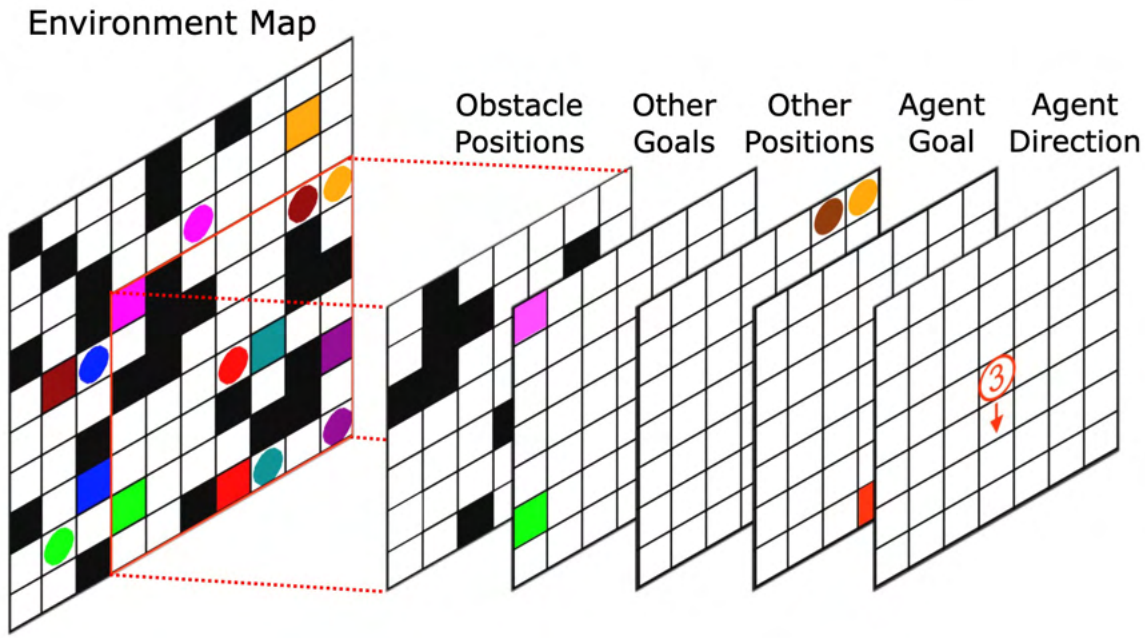
+ global — ALPHA (He et al. 2023)

Local, static observation with

expert ➡ dynamic ➡ heuristic ➡ global information



PRIMAL (Sartoretti el al. 2019)

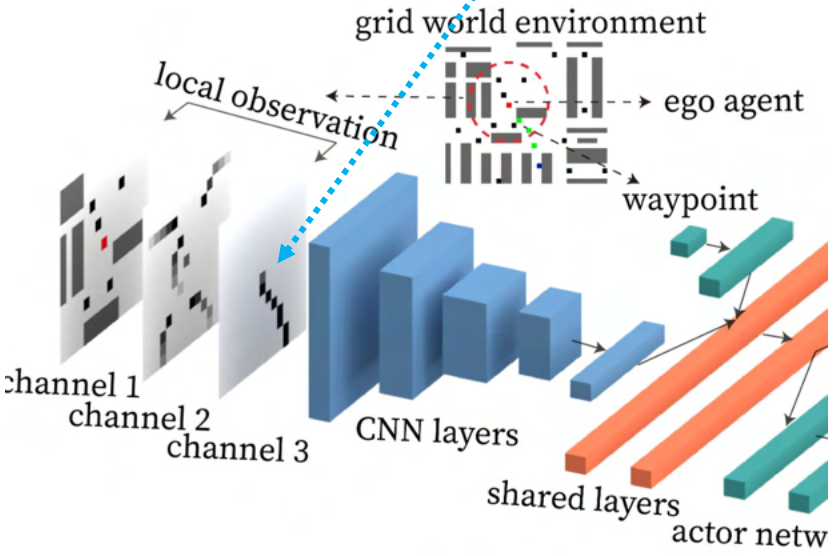


CACTUS (Phan et al. 2024)

Local, static observation with

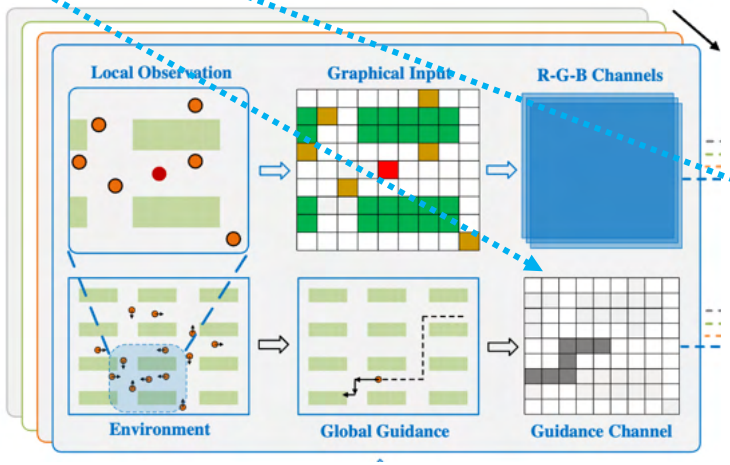
expert ➡ dynamic ➡ heuristic ➡ global information

Reference path planned by A*

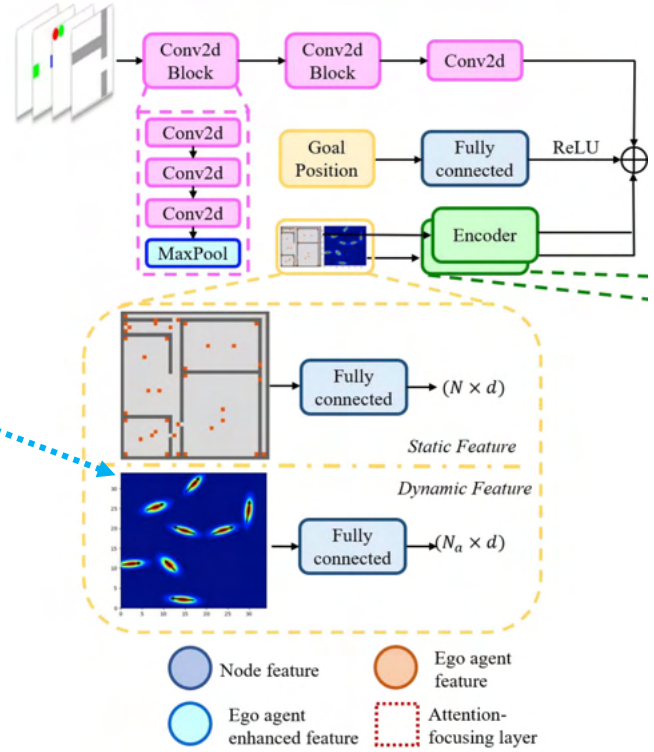


MAPPER (Liu et al. 2020)

AB-MAPPER (Guan et al. 2021)



G2RL (Wang et al. 2020)

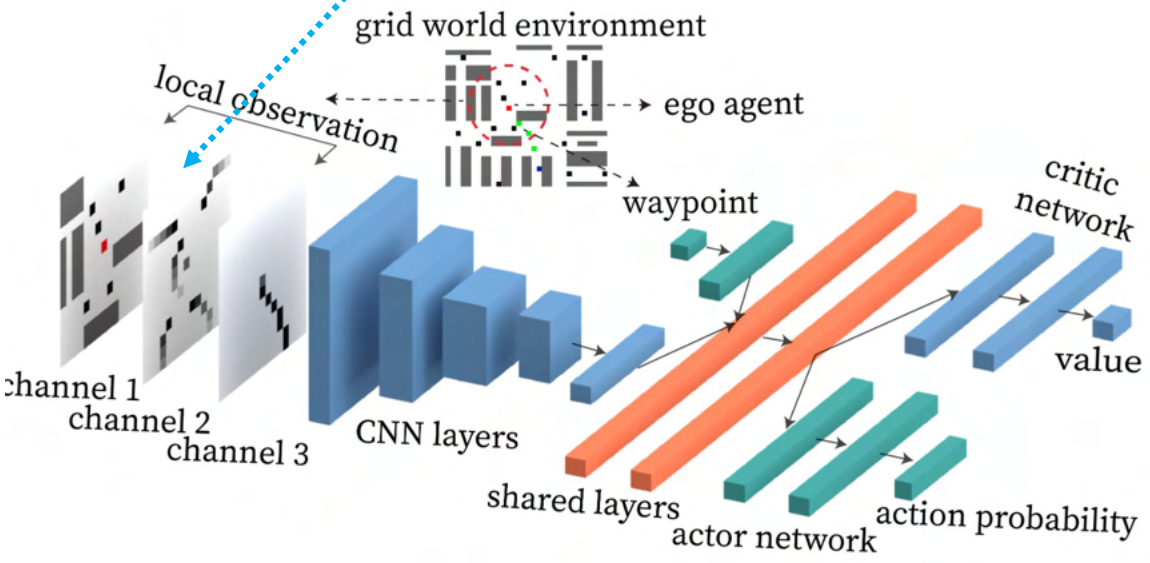


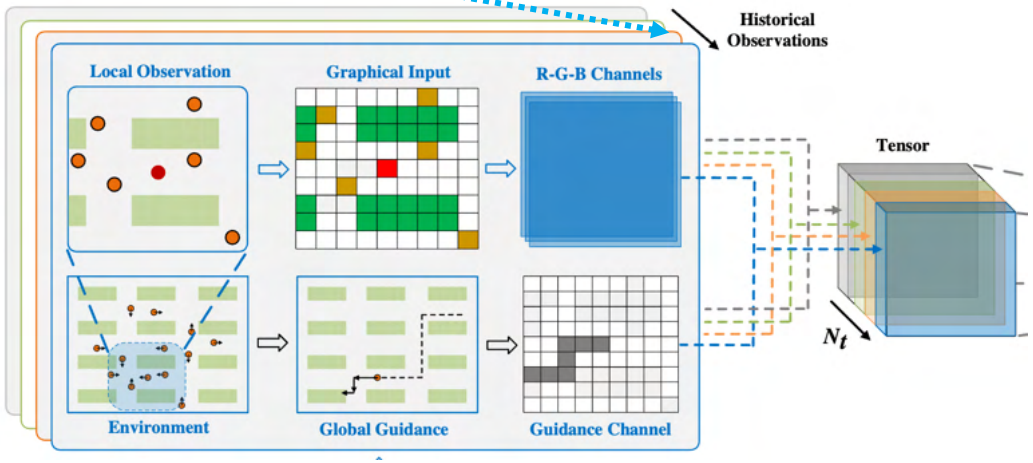
ALPHA (He et al. 2023)

Local, static observation with

expert ➡ dynamic ➡ heuristic ➡ global information

Historical information





MAPPER (Liu et al. 2020)

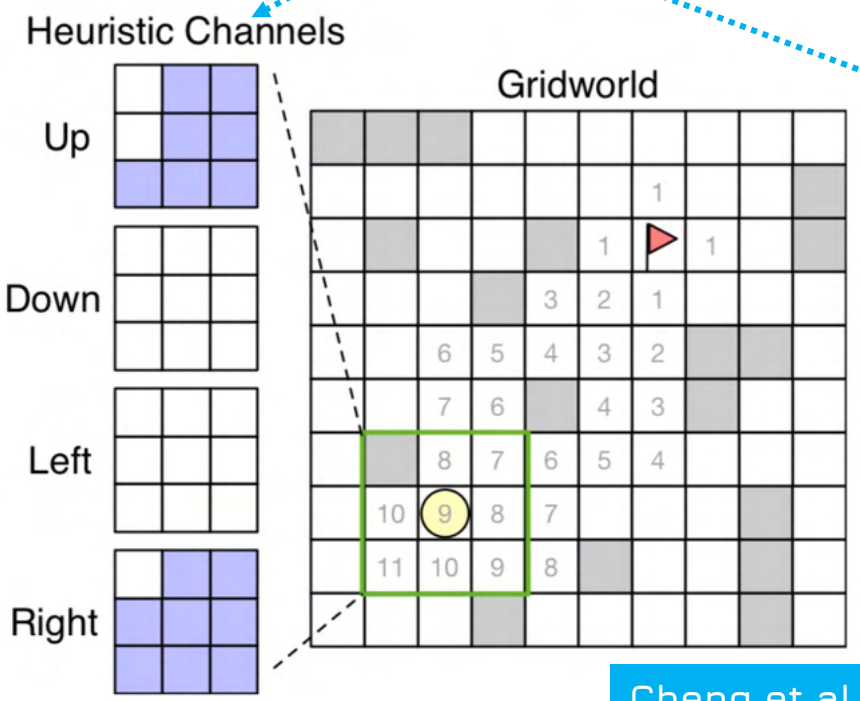
AB-MAPPER (Guan et al. 2021)

G2RL (Wang et al. 2020)

Local, static observation with

expert ➡ dynamic ➡ heuristic ➡ global information

Goal-guided heuristics



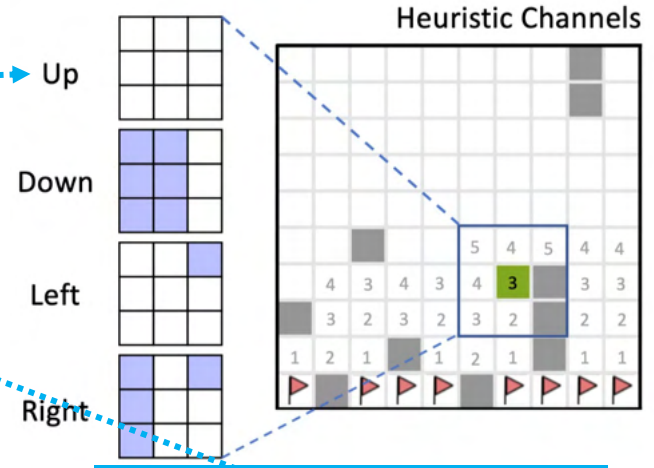
Cheng et al. 2023

DHC (Ma et al. 2021)

DCC (Ma et al. 2021)

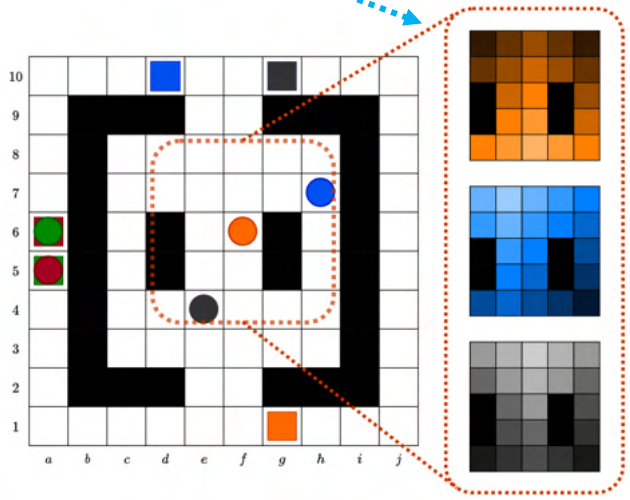
Chen et al. 2023a/b

SCRIMP (Wang et al. 2023)



HELSA (Song et al. 2023)

Goal to sub-goal



SACHA (Lin et al. 2023)

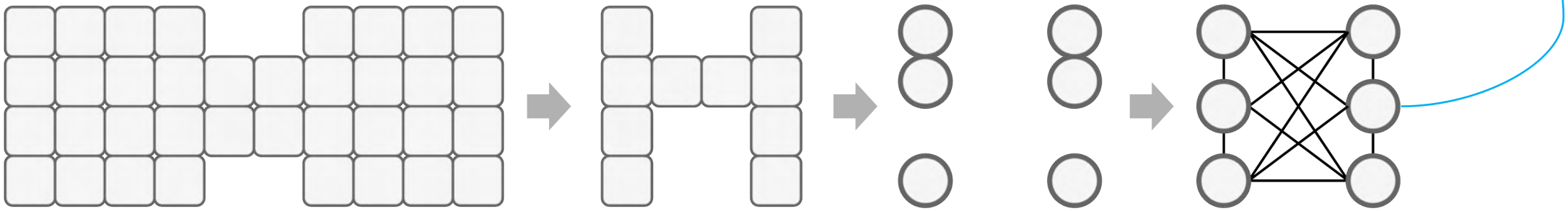
Ego to others

Local, static observation with

expert ➡ dynamic ➡ heuristic ➡ global information

Global static observations

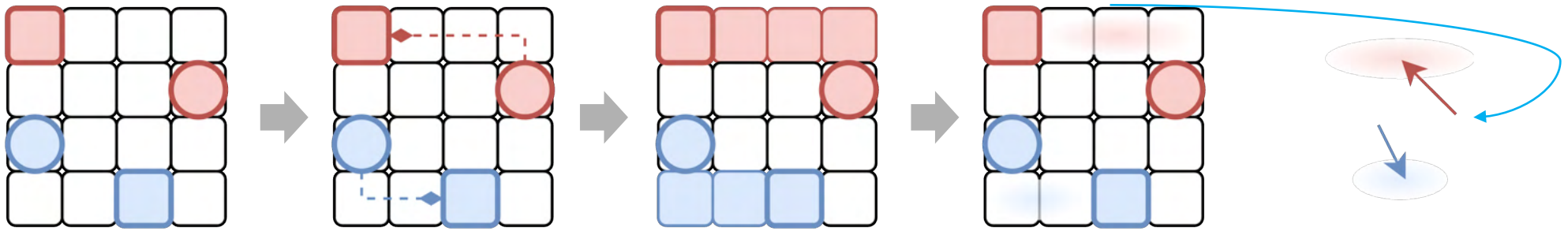
ALPHA (He et al. 2023)



skeletonization neighborhood analysis edge generation

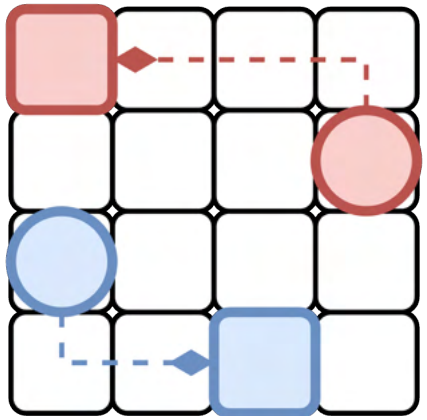
Global dynamic observations

ALPHA (He et al. 2023)

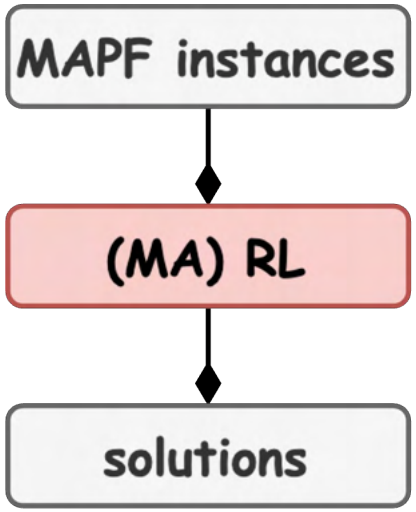


pathfinding with A* fitting a Gaussian

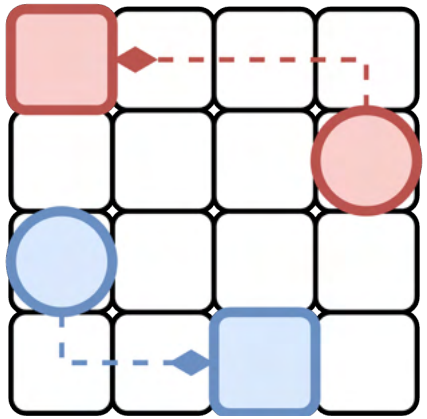
Roadmap



Reward



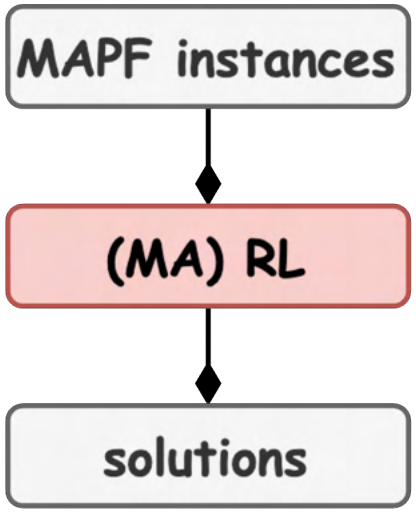
Roadmap



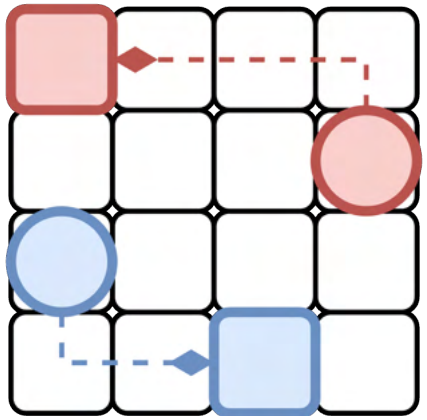
Reward

Success or not

Almost all of them



Roadmap



MAPF instances

(MA) RL

solutions

Reward

Success or not

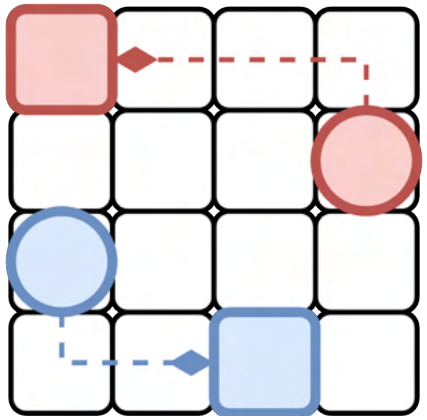
Almost all of them

+ cooperation

CPL (Zhao et al. 2023)

CACTUS (Phan et al. 2024)

Roadmap



MAPF instances

(MA) RL

solutions

Reward

Success or not

+ cooperation

+ expert

Almost all of them

CPL (Zhao et al. 2023)

CACTUS (Phan et al. 2024)

MAPPER (Liu et al. 2020)

G2RL (Wang et al. 2020)

Davydov et al. 2021

AB-MAPPER (Guan et al. 2021)

SACHA (Lin et al. 2023)

ALPHA (He et al. 2023)

CRAMP (Pham et al. 2024)

Success or not reward with

cooperation ➡ expert constraints

Action	Reward
Move [N/E/S/W]	-0.3
Agent Collision	-2.0
No Movement (on/off goal)	0.0 / -0.5
Finish Episode	+20.0

PRIMAL (Sartoretti el al. 2019)

Actions	Reward
Move (Up/Down/Left/Right)	-0.075
Stay (on goal, away goal)	0, -0.075
Collision (obstacle/agents)	-0.5
Finish	3

DHC (Ma et al. 2021)

Reward Type	Reward
Stay idle	-0.5
Move forward	-0.3
Rotate 90° clockwise/ anticlockwise	-0.5
Collision	-2.0
Reach goal	0.0
Finish Episode	+35

MAPF-ROT (Chan et al. 2022)

Action	Reward
Move(Up/Down/Left/Right)	-0.3
Stay(On goal, Away goal)	0.0,-0.3
Collision	-2
Block	-1

SCRIMP (Wang et al. 2023)

Success or not reward with cooperation ➡ expert constraints

CPL (Zhao et al. 2023)

When the entire MAPF task is completed, the team will receive a **team reward of value +20 * k**, where k is the number of agents

Individual Action	Reward
Move to the goal at the first time	+20.0
Agent collision	-2.0
Common move	-0.3
Stay off the goal	-0.5
Stay on the goal	0.0

CACTUS (Phan et al. 2024)

The individual reward is defined by +1 if agent i **reaches its goal**, 0 when agent i is **staying at its goal location**, and -1 otherwise

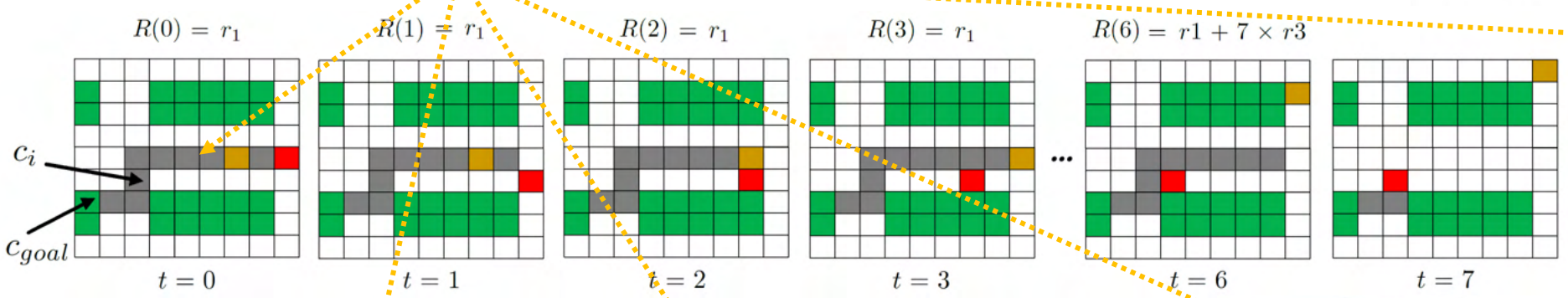
Success or not reward with cooperation ➡ expert constraints

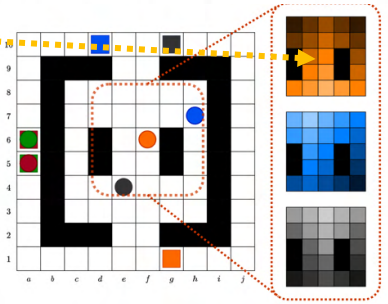
SACHA (Lin et al. 2023)

Expert/goal-guided reward

G2RL (Wang et al. 2020)

$$\tilde{r}_i(s,a) = r_i(s,a) + (1-\lambda)\gamma h_i(s')$$

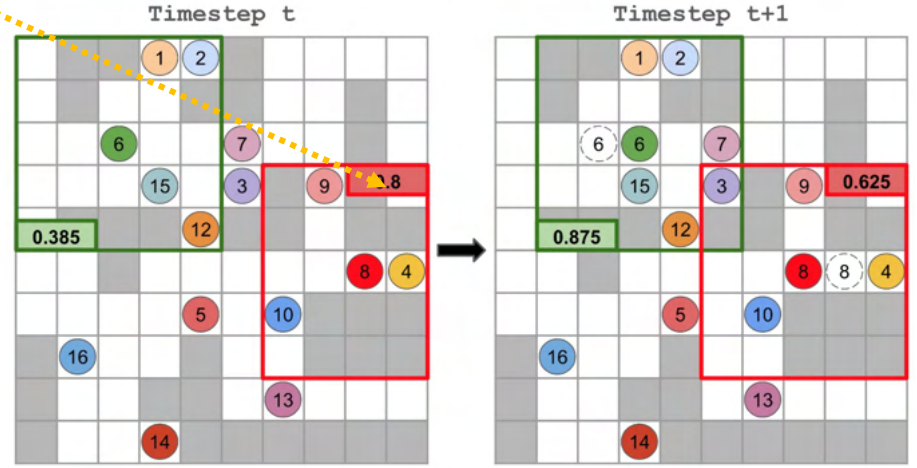




Reward	Value
step penalty r_s	-0.1 (move) or -0.5 (wait)
collision penalty r_c	-5
oscillation penalty r_o	-0.3
off-route penalty r_f	$-\min_{p \in \mathcal{S}} \|\|p_a - p\|\|_2$
goal-reaching reward r_g	30

MAPPER (Liu et al. 2020)

AB-MAPPER (Guan et al. 2021)

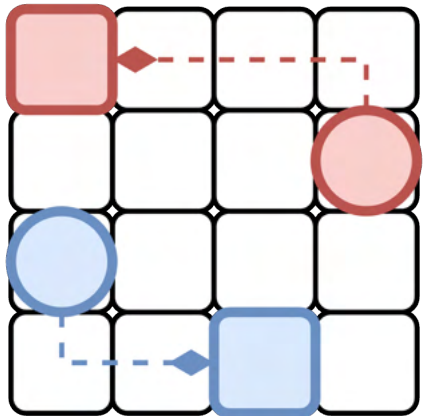


CRAMP (Pham et al. 2024)

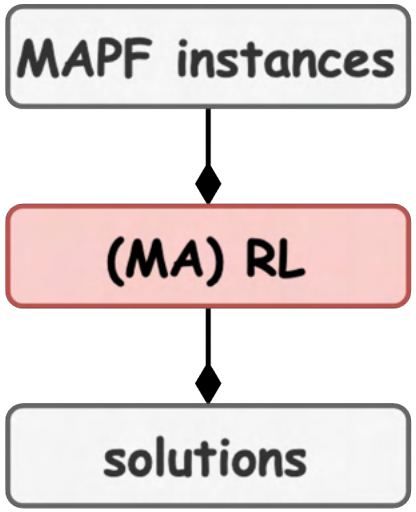
Davydov et al. 2021

Agents receive 0.5 if they **follow** one of the optimal routes, −1 if the agent has **increased** her distance to the target, and −0.5 if the agent **stays** in place

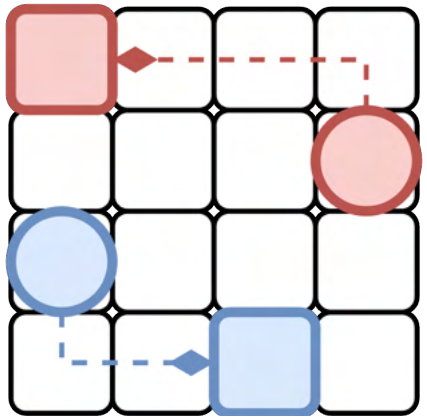
Roadmap

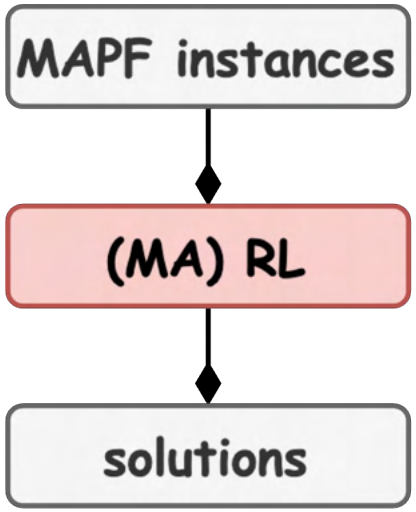


Learning



Roadmap





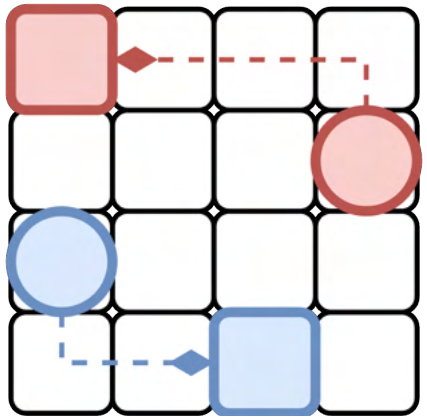
Learning

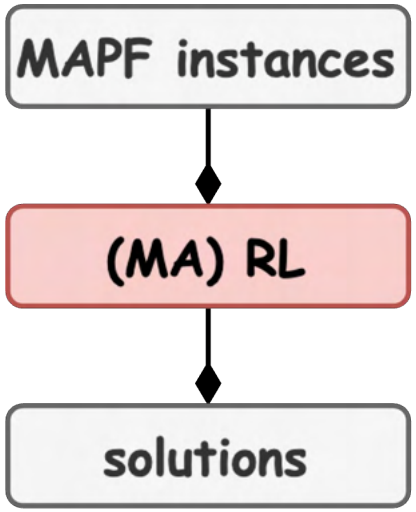
Independent learning

Cruz et al. 2014

G2RL (Wang et al. 2020)

Roadmap





Learning

Independent learning

imitation learning

Cruz et al. 2014

G2RL

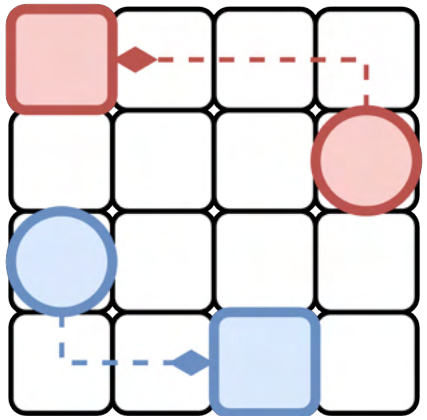
PRIMAL (Sartoretti el al. 2019)

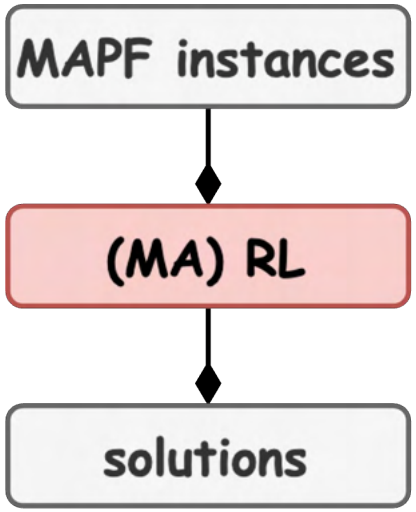
Yang et al. 2020

Chen et al. 2023b

ALPHA (He et al. 2023)

Roadmap





Learning

Independent learning

imitation learning

communication

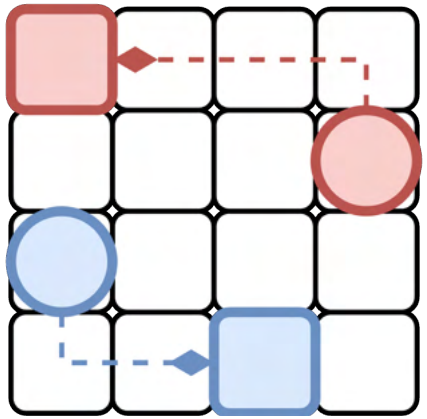
Cruz et al. 2014 G2RL

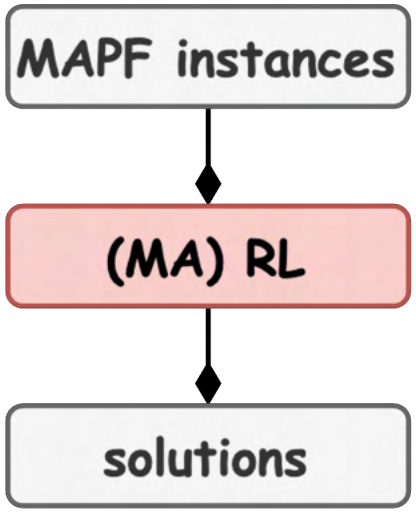
PRIMAL Yang et al. 2020 Chen et al. 2023b ALPHA

DHC (Ma et al. 2021) DCC (Ma et al. 2021) PICO (Li et al. 2022)

Cheng et al. 2023 CRAMP (Pham et al. 2024)

Roadmap





Learning

independent learning

imitation learning

communication

cooperation

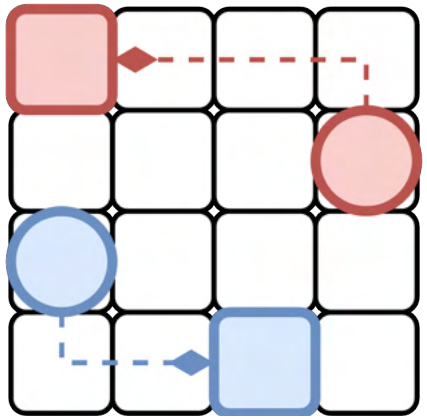
Cruz et al. 2014 G2RL

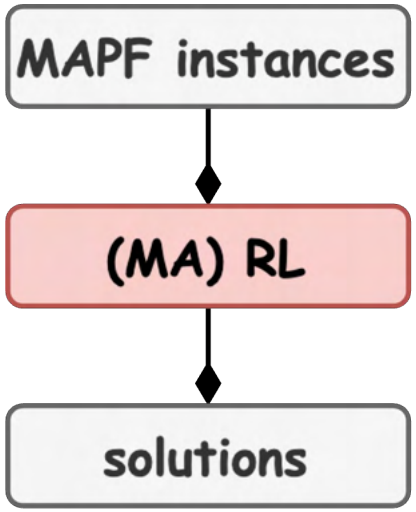
PRIMAL Yang et al. 2020 Chen et al. 2023b ALPHA

DHC DCC PICO Cheng et al. 2023 CRAMP

Davydov et al. 2021 Hu et al. 2023 SACHA (Lin et al. 2023)

Roadmap

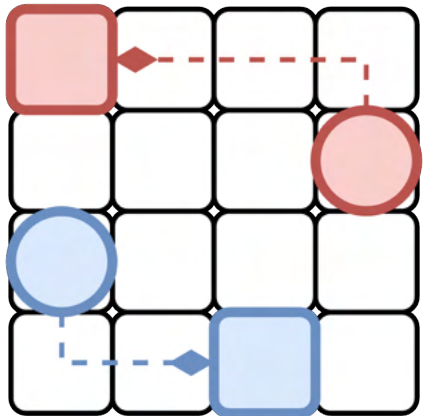


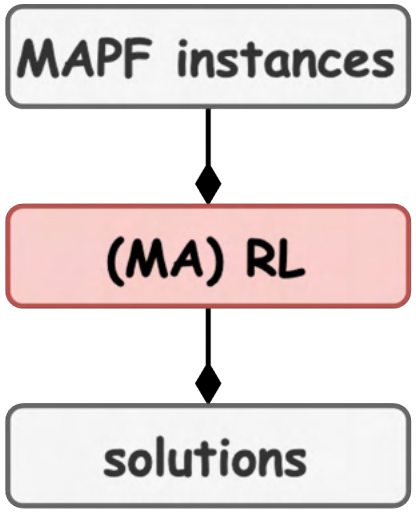


Learning

independent learning	Cruz et al. 2014, G2RL
imitation learning	PRIMAL, Yang et al. 2020, Chen et al. 2023b, ALPHA
communication	DHC, DCC, PICO, Cheng et al. 2023, CRAMP
cooperation	Davydov et al. 2021, Hu et al. 2023, SACHA
hierarchical learning	HELSA (Song et al. 2023)

Roadmap

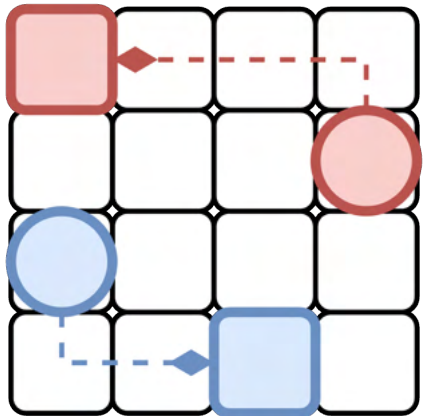


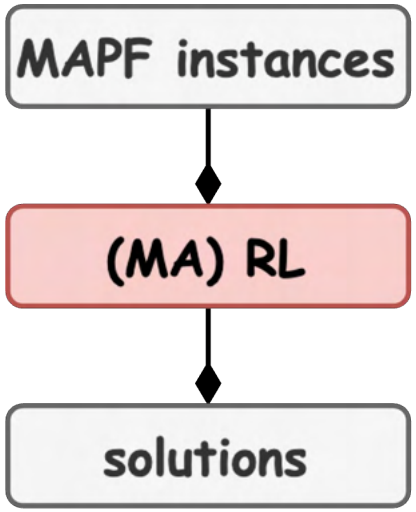


Learning

independent learning	Cruz et al. 2014, G2RL
imitation learning	PRIMAL, Yang et al. 2020, Chen et al. 2023b, ALPHA
communication	DHC, DCC, PICO, Cheng et al. 2023, CRAMP
cooperation	Davydov et al. 2021, Hu et al. 2023, SACHA
hierarchical learning	HELSA
evolutionary computing	MAPPER (Liu et al. 2020), AB-MAPPER (Guan et al. 2021)

Roadmap

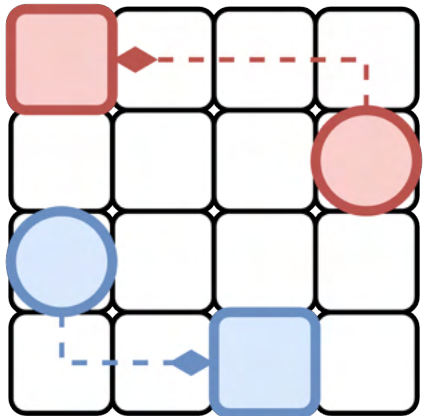


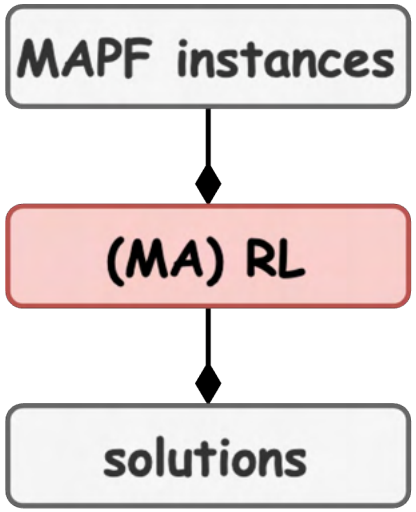


Learning

independent learning	Cruz et al. 2014, G2RL
imitation learning	PRIMAL, Yang et al. 2020, Chen et al. 2023b, ALPHA
communication	DHC, DCC, PICO, Cheng et al. 2023, CRAMP
cooperation	Davydov et al. 2021, Hu et al. 2023, SACHA
hierarchical learning	HELSA
evolutionary computing	MAPPER, AB-MAPPER
curriculum learning	CPL (Zhao et al. 2023), CACTUS (Phan et al. 2024)

Roadmap

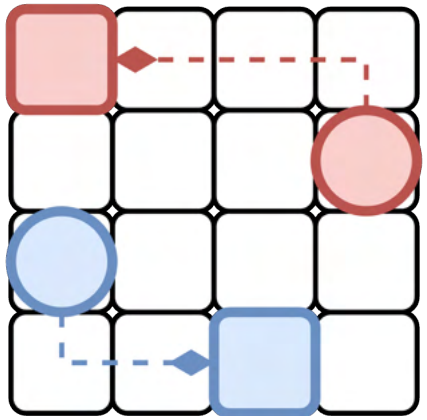


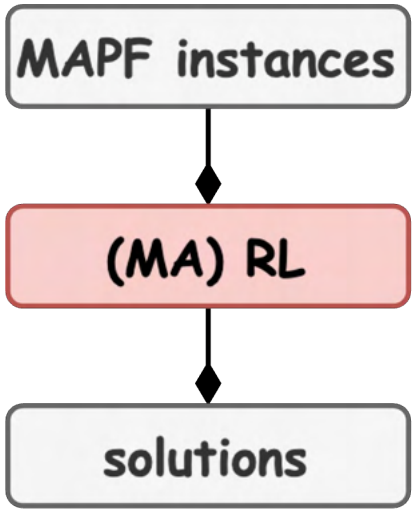


Learning

independent learning	Cruz et al. 2014, G2RL
imitation learning	PRIMAL, Yang et al. 2020, Chen et al. 2023b, ALPHA
communication	DHC, DCC, PICO, Cheng et al. 2023, CRAMP
cooperation	Davydov et al. 2021, Hu et al. 2023, SACHA
hierarchical learning	HELSA
evolutionary computing	MAPPER, AB-MAPPER
curriculum learning	CPL, CACTUS
representation learning	SCRIMP (Wang et al. 2023), Chen et al. 2023a

Roadmap



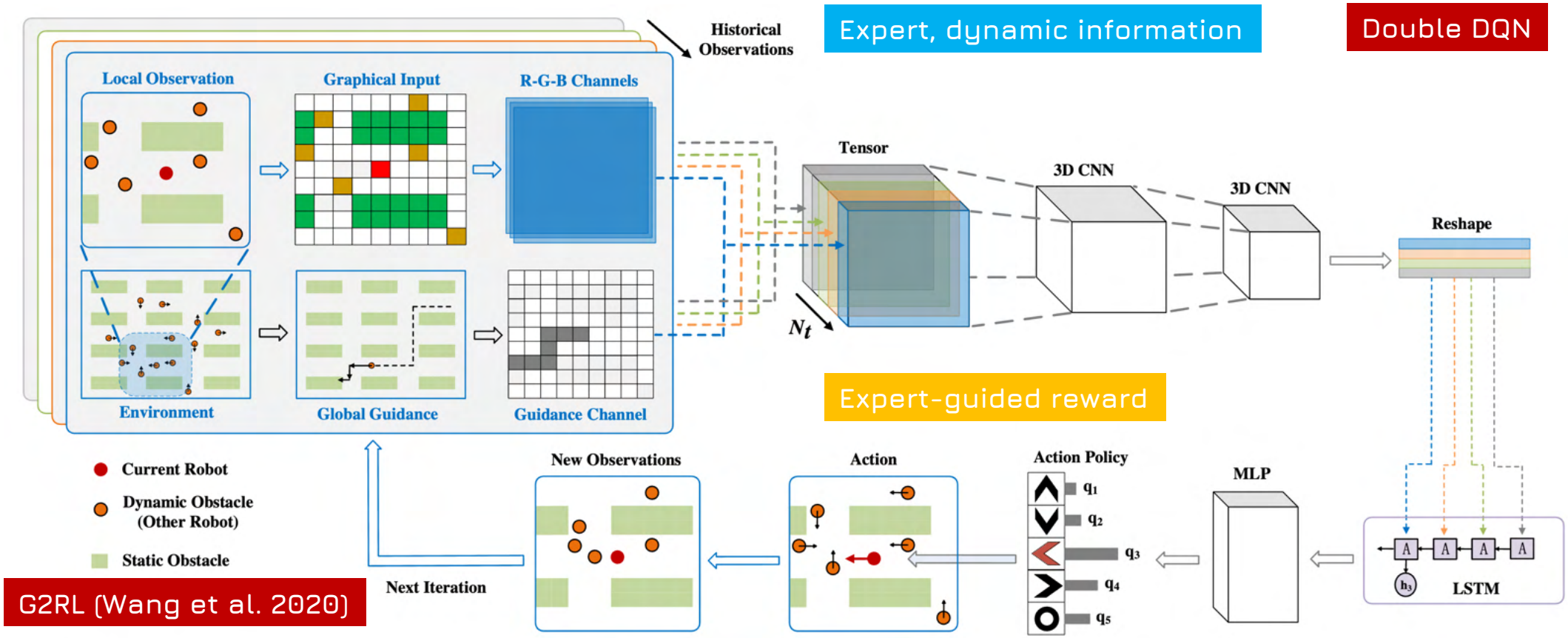


Learning

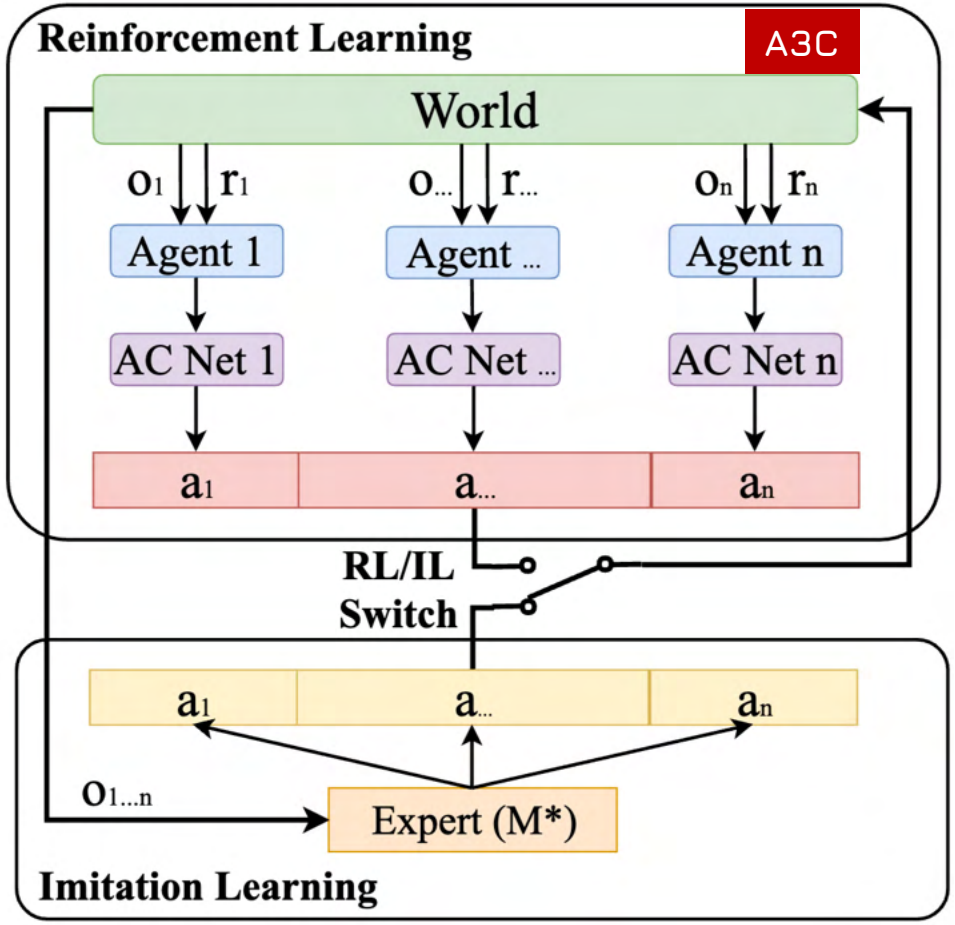
independent learning	Cruz et al. 2014, G2RL
imitation learning	PRIMAL, Yang et al. 2020, Chen et al. 2023b, ALPHA
communication	DHC, DCC, PICO, Cheng et al. 2023, CRAMP
cooperation	Davydov et al. 2021, Hu et al. 2023, SACHA
hierarchical learning	HELSA
evolutionary computing	MAPPER, AB-MAPPER
curriculum learning	CPL, CACTUS
representation learning	SCRIMP, Chen et al. 2023a
postprocessing	MAPF-ROT (Chan et al. 2022)

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- Cruz et al. 2014 use the WOLF-PHC algorithm (a classic MARL method) to solve the 2-agents MAPF problem

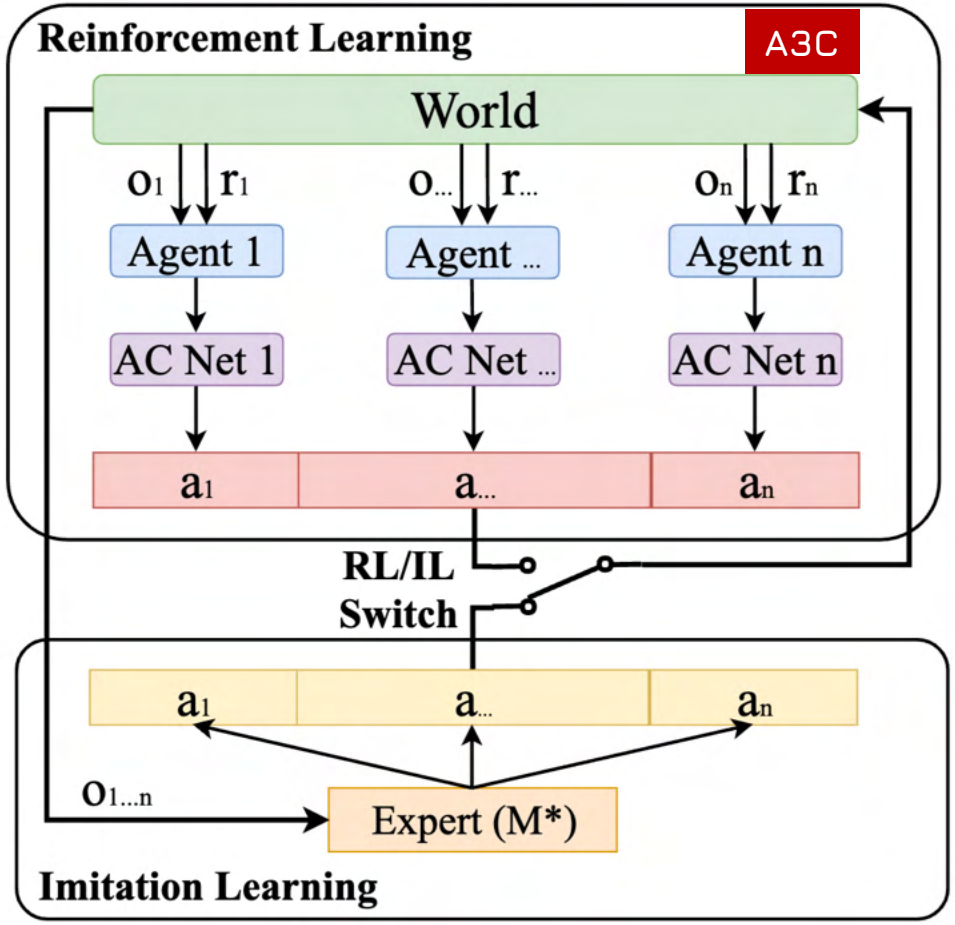


Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

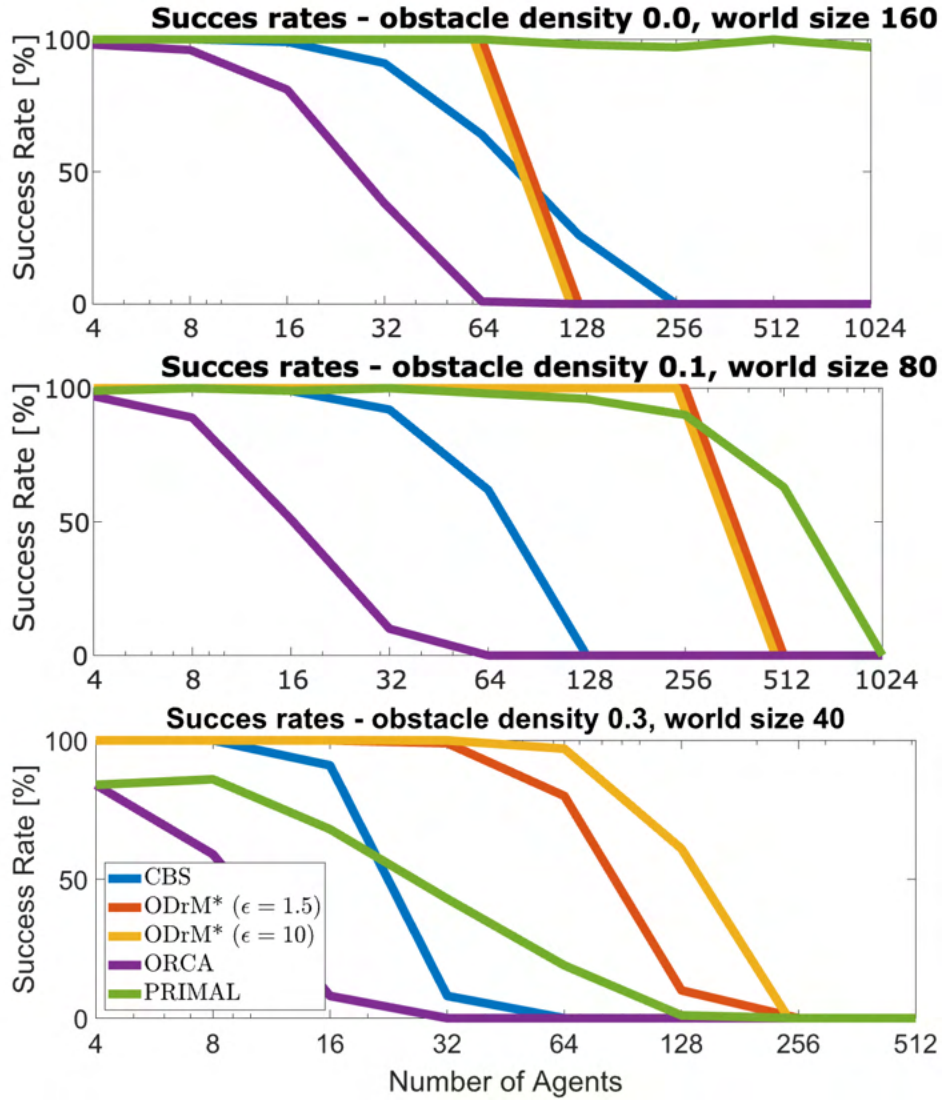


PRIMAL (Sartoretti el al. 2019)

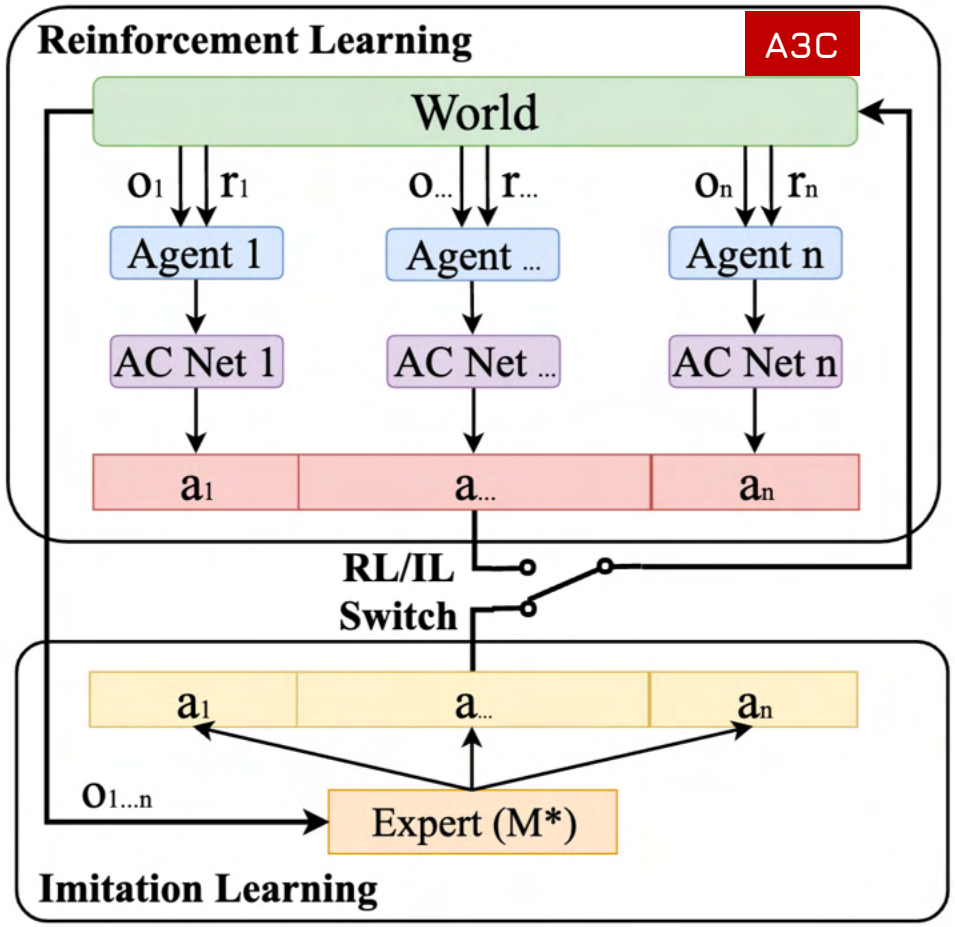
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



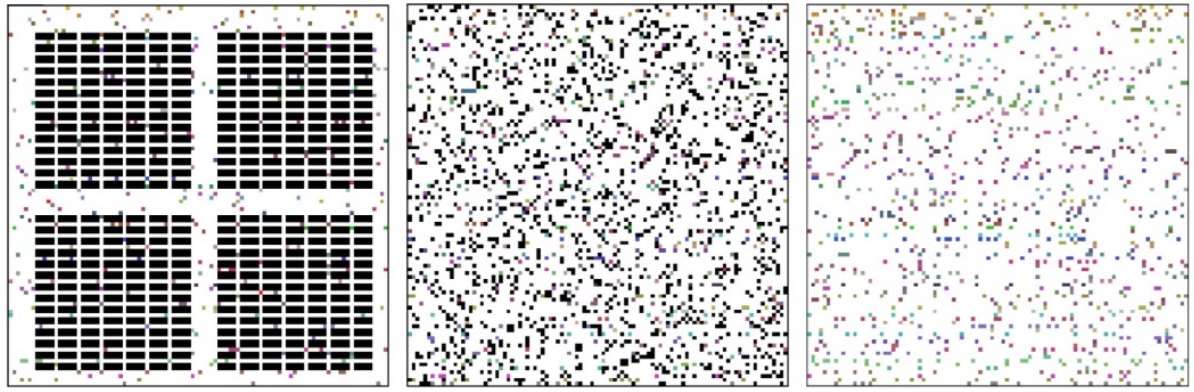
PRIMAL (Sartoretti el al. 2019)



Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



PRIMAL (Sartoretti el al. 2019)



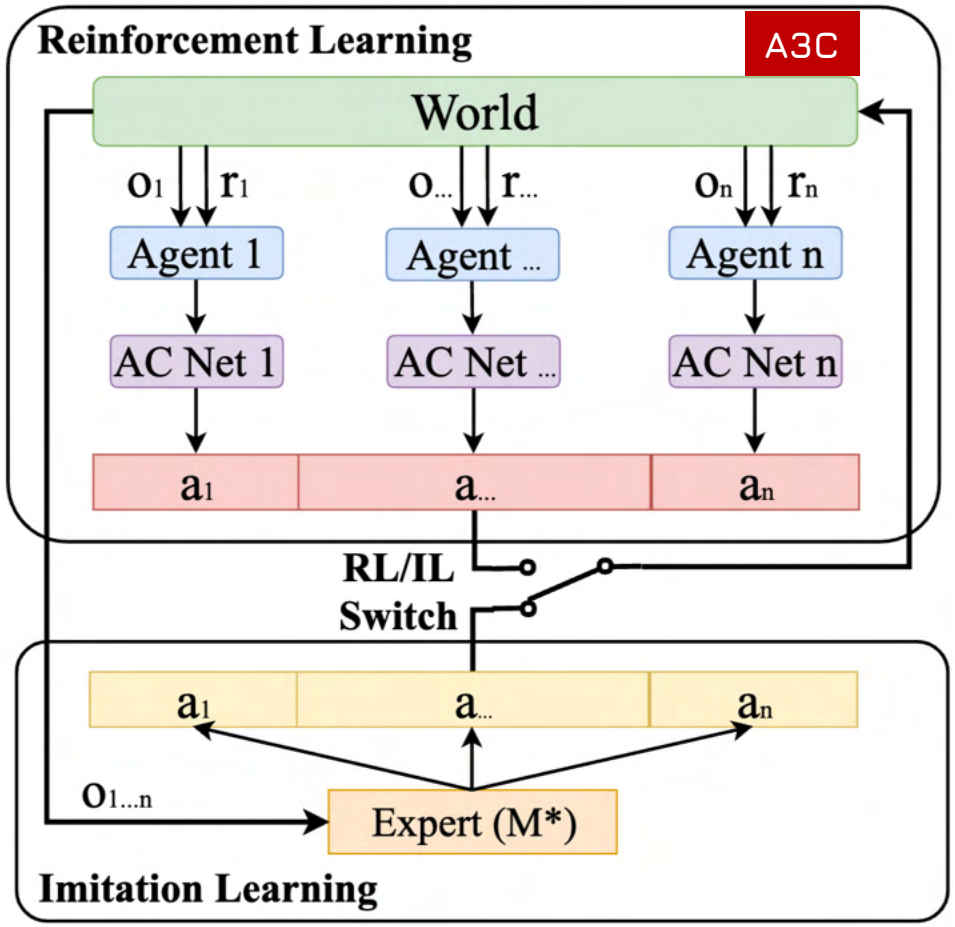
(a) Regular map (b) Random map (c) Free map

	Success Rate					
	ECBS	HCA*	Global Re-planning	Discrete-ORCA	PRIMAL	G2RL
Regular	**100%**	**100%**	95.7%	88.7%	92.3%	99.7%
Random	**100%**	**100%**	98.2%	55.0%	80.6%	99.7%
Free	**100%**	**100%**	98.8%	99.5%	75.7%	99.8%

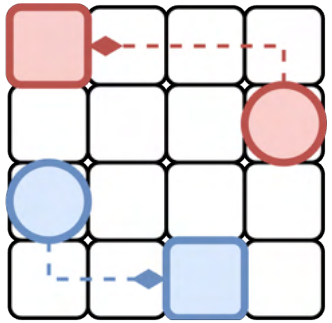
Imitation learning may not always work!

Independent learning with imitation ➡ communication ➡ cooperation

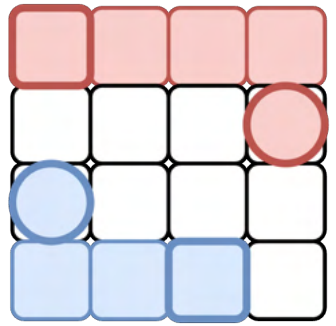
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



PRIMAL (Sartoretti el al. 2019)



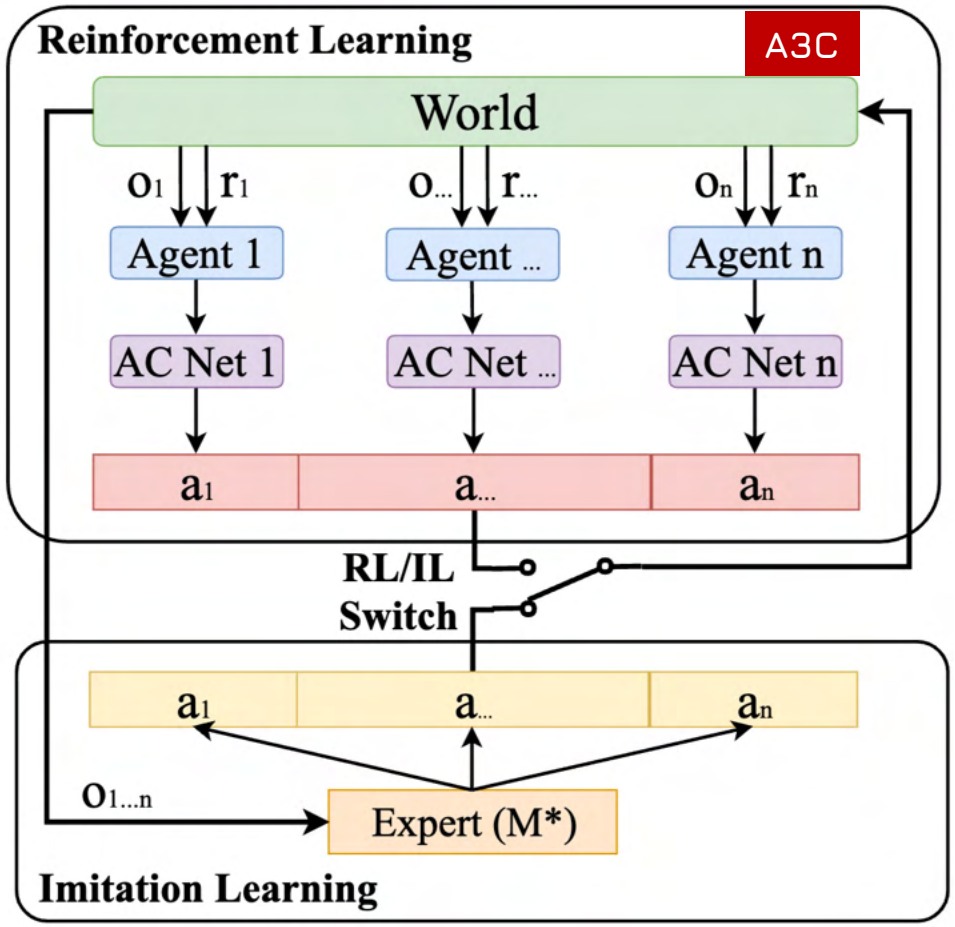
initialize



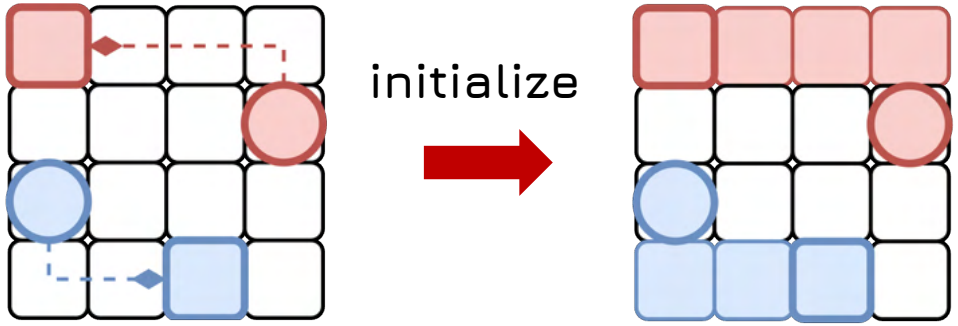
Q-Table

Yang et al. 2020

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



PRIMAL (Sartoretti el al. 2019)

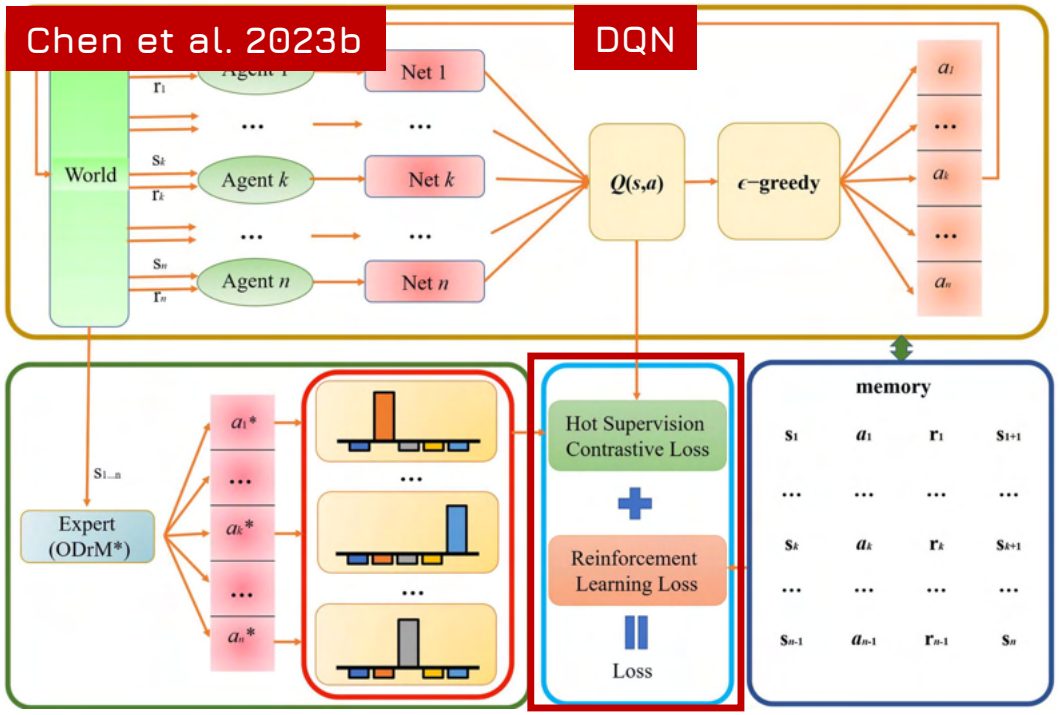


Q-Table

Yang et al. 2020

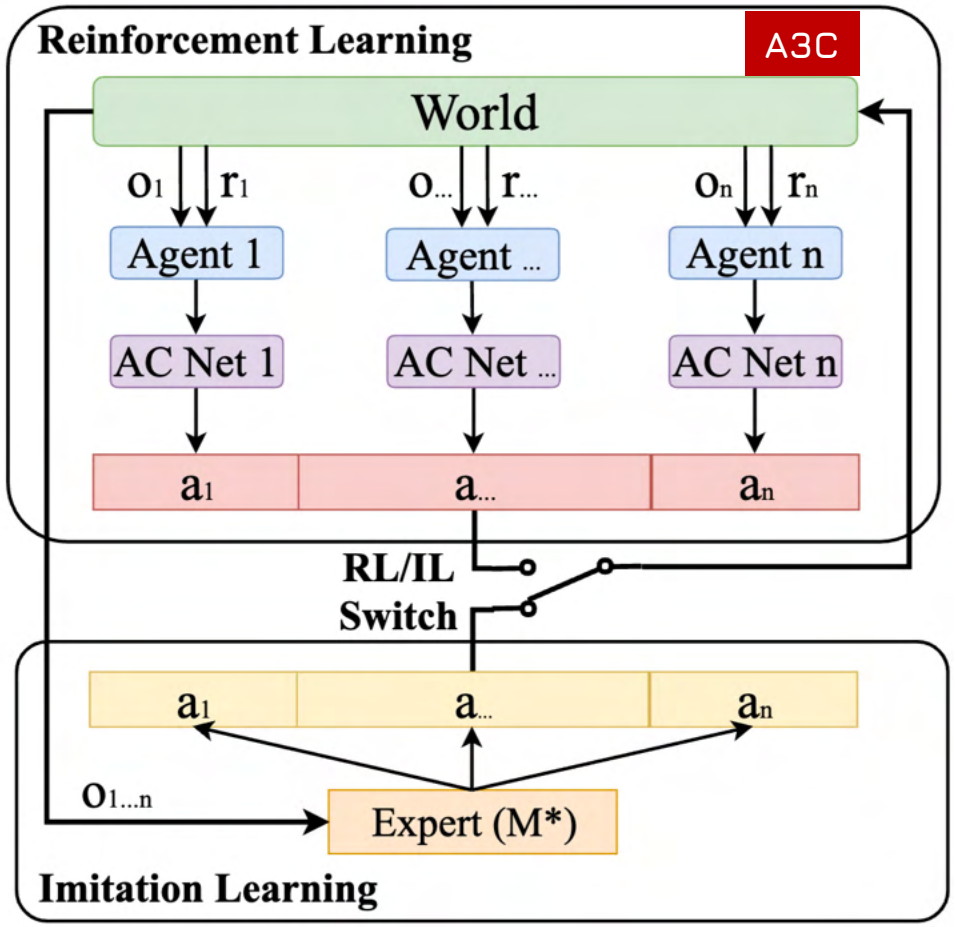
Chen et al. 2023b

DQN

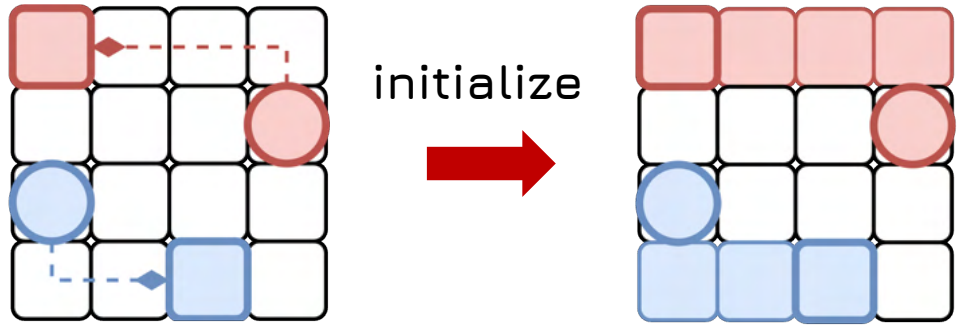


Independent learning with imitation ➡ communication ➡ cooperation

hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



PRIMAL (Sartoretti el al. 2019)

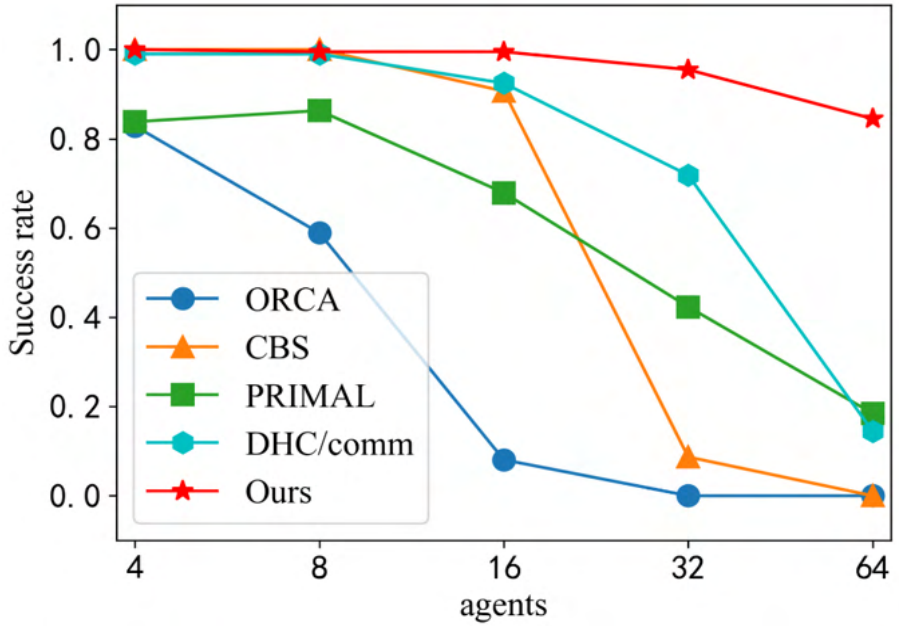


Chen et al. 2023b

DQN

Q-Table

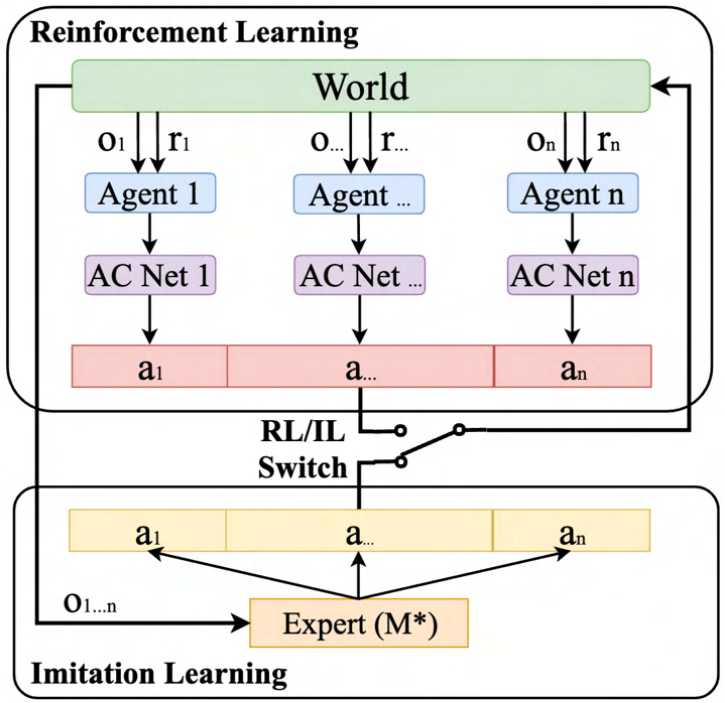
Yang et al. 2020

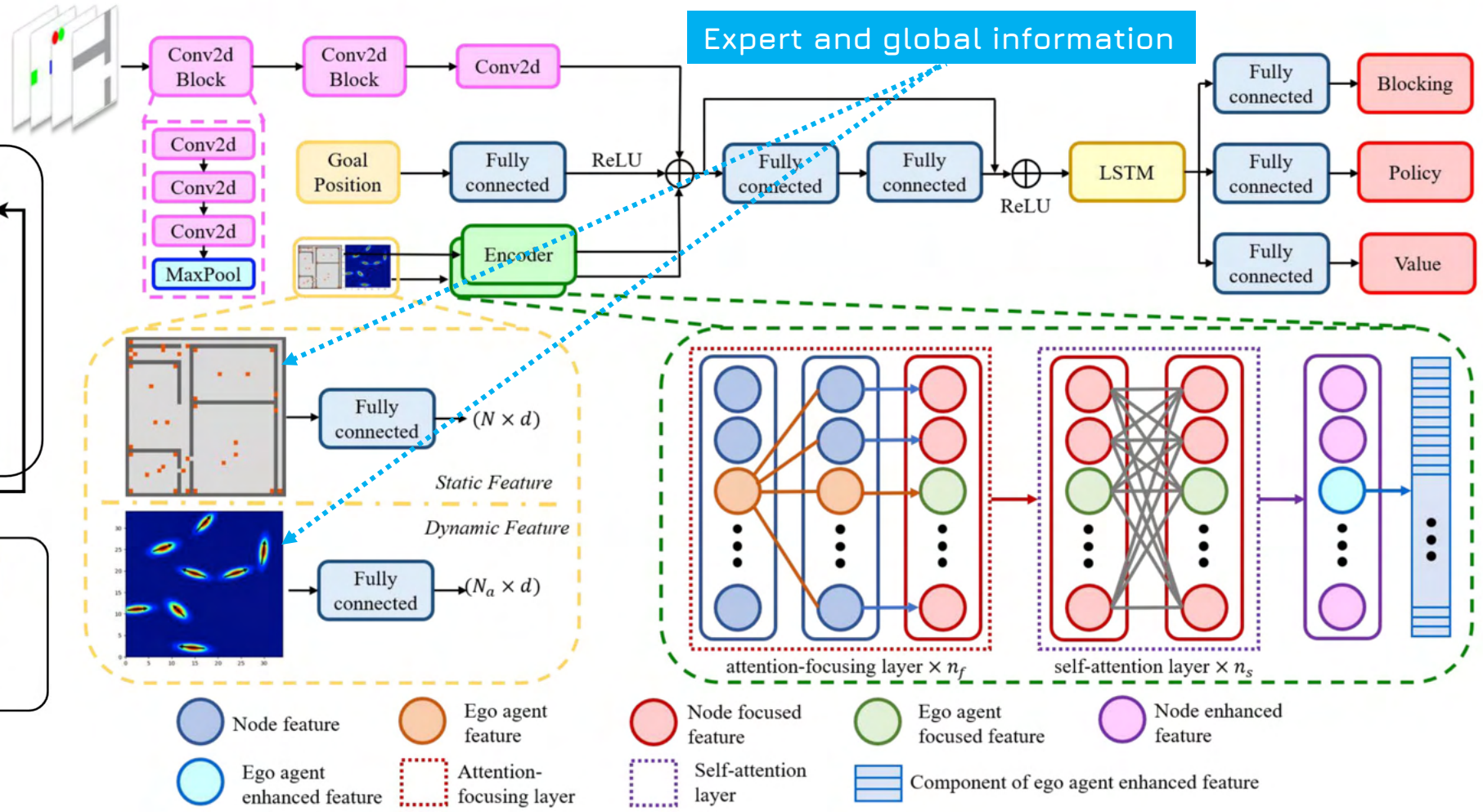


Chen et al. 2023b

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

ALPHA (He et al. 2023)





PRIMAL (Sartoretti el al. 2019)

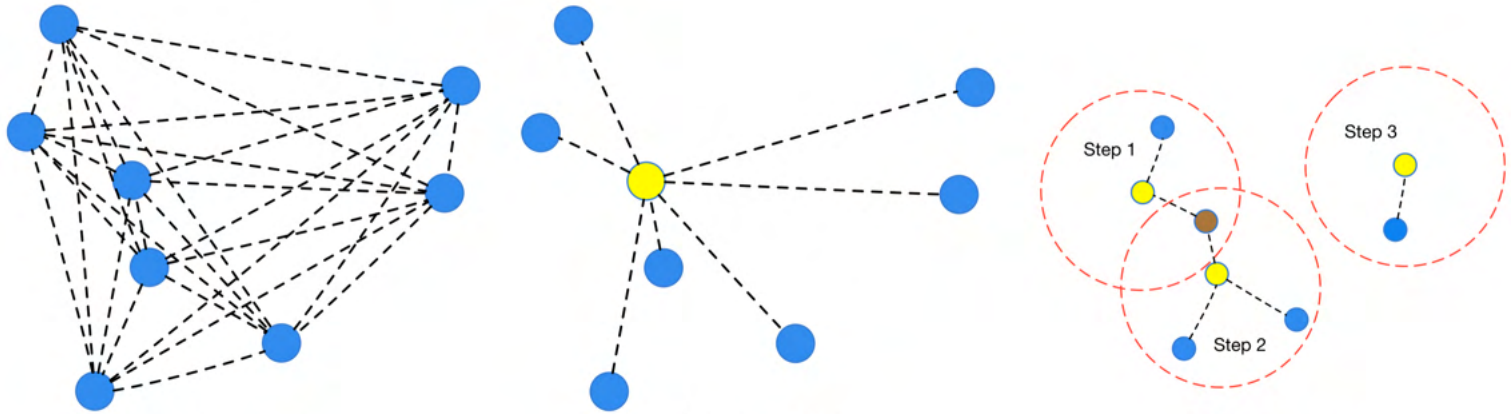
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

ALPHA (He et al. 2023)

Model	MS↓						AR↑						SR↑					
	20 × 20 room-liked environment with 4, 8, 16, 32, 64, 128 agents																	
~~ODrM*~~	~~30.58~~	~~43.19~~	~~97.25~~	~~292.93~~	~~512.00~~	~~512.00~~	~~100%~~	~~98.00%~~	~~88.00%~~	~~47.00%~~	~~0.00%~~	~~0.00%~~	~~100%~~	~~98%~~	~~88%~~	~~47%~~	~~0%~~	~~0%~~
PRIMAL	201.32	275.93	439.95	506.83	512.00	512.00	93.50%	90.88%	88.63%	81.72%	66.79%	**35.74%**	79%	67%	30%	2%	0%	0%
~~MAPPER~~	~~79.81~~	~~101.05~~	~~246.69~~	~~427.33~~	~~512.00~~	~~512.00~~	~~98.75%~~	~~97.53%~~	~~95.75%~~	~~89.71%~~	~~53.92%~~	~~6.96%~~	~~97%~~	~~97%~~	~~82%~~	~~41%~~	~~0%~~	~~0%~~
G2RL	42.22	65.46	159.12	356.94	511.45	512.00	98.00%	98.75%	98.00%	94.43%	68.50%	18.07%	99%	97%	87%	54%	1%	0%
DHC	45.50	73.86	175.22	354.43	509.69	512.00	99.00%	98.62%	96.56%	90.69%	69.70%	20.09%	98%	93%	77%	45%	1%	0%
SCRIMP	43.42	61.56	186.34	**214.32**	**488.98**	–	99.25%	99.37%	98.87%	97.53%	**82.32%**	–	98%	96%	93%	75%	**15%**	–
ALPHA	**37.39**	**52.26**	**120.65**	310.21	503.87	512.00	**100%**	**100%**	**99.75%**	**97.69%**	70.12%	25.71%	**100%**	**100%**	**96%**	**78%**	8%	0%
	40 × 40 room-like environment with 4, 8, 16, 32, 64, 128 agents																	
~~ODrM*~~	~~56.73~~	~~69.34~~	~~91.89~~	~~146.88~~	~~375.37~~	~~512.00~~	~~100%~~	~~100%~~	~~97.00%~~	~~85.00%~~	~~32.00%~~	~~0.00%~~	~~100%~~	~~100%~~	~~97%~~	~~85%~~	~~32%~~	~~0%~~
PRIMAL	285.55	384.93	463.86	492.82	511.80	512.00	91.00%	87.62%	85.56%	82.69%	73.14%	61.71%	73%	47%	23%	11%	1%	0%
~~MAPPER~~	~~104.82~~	~~157.12~~	~~218.35~~	~~348.95~~	~~491.58~~	~~512.00~~	**100%**	~~99.37%~~	~~98.00%~~	~~93.71%~~	~~76.60%~~	~~31.02%~~	**100%**	~~96%~~	~~91%~~	~~66%~~	~~16%~~	~~0%~~
G2RL	**57.60**	93.31	166.92	241.39	433.13	512.00	**100%**	99.75%	99.06%	98.65%	93.53%	70.50%	**100%**	98%	89%	81%	33%	0%
DHC	104.19	127.78	188.62	263.81	427.02	512.00	97.75%	98.00%	97.88%	95.94%	91.17%	72.53%	92%	91%	80%	65%	28%	0%
SCRIMP	58.53	91.84	**116.05**	**183.54**	396.93	**484.76**	**100%**	99.62%	99.56%	99.21%	**94.10%**	**85.09%**	**100%**	97%	95%	84%	42%	**12%**
ALPHA	64.04	**88.75**	140.96	206.85	**392.23**	506.48	**100%**	**100%**	**99.75%**	**99.34%**	93.46%	73.99%	**100%**	**100%**	**97%**	**93%**	**60%**	7%
	60 × 60 room-liked environment with 4, 8, 16, 32, 64, 128 agents																	
~~ODrM*~~	~~84.71~~	~~98.43~~	~~196.46~~	~~163.53~~	~~228.95~~	~~457.17~~	~~100%~~	~~100%~~	~~99.00%~~	~~88.00%~~	~~72.00%~~	~~14.00%~~	~~100%~~	~~100%~~	~~99%~~	~~88%~~	~~72%~~	~~14%~~
PRIMAL	363.45	465.35	495.85	508.17	512.00	512.00	84.75%	78.37%	79.75%	73.62%	71.51%	62.83%	54%	25%	11%	3%	0%	0%
~~MAPPER~~	~~177.61~~	~~241.31~~	~~280.69~~	~~388.55~~	~~490.02~~	~~512.00~~	~~**99.50%**~~	~~97.75%~~	~~98.31%~~	~~93.87%~~	~~85.96%~~	~~62.47%~~	~~97%~~	~~89%~~	~~90%~~	~~61%~~	~~17%~~	~~0%~~
G2RL	**104.40**	**140.85**	168.70	280.04	431.21	512.00	99.00%	98.62%	97.36%	95.62%	93.76%	86.01%	96%	94%	91%	68%	34%	0%
DHC	131.59	203.71	186.66	323.19	406.40	496.70	97.75%	96.75%	98.88%	95.16%	93.30%	87.79%	91%	77%	86%	54%	35%	7%
SCRIMP	106.79	166.37	**125.50**	**211.03**	421.65	498.72	**99.50%**	**99.25%**	99.61%	98.73%	96.79%	88.08%	**98%**	95%	**97%**	81%	31%	8%
ALPHA	110.82	158.59	173.02	263.74	**357.17**	**485.23**	**99.50%**	**99.25%**	**99.75%**	**99.03%**	**97.91%**	**89.16%**	**98%**	**97%**	**97%**	**86%**	**67%**	**25%**

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

Communication topologies in MARL



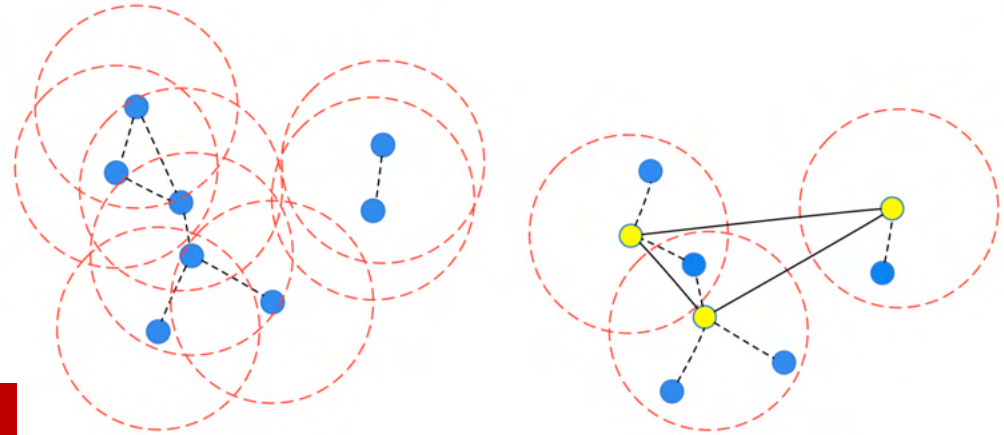
(a) FC (b) STAR (c) TREE

DHC (Ma et al. 2021)

DCC (Ma et al. 2021)

Cheng et al. 2023

CRAMP (Pham et al. 2024)



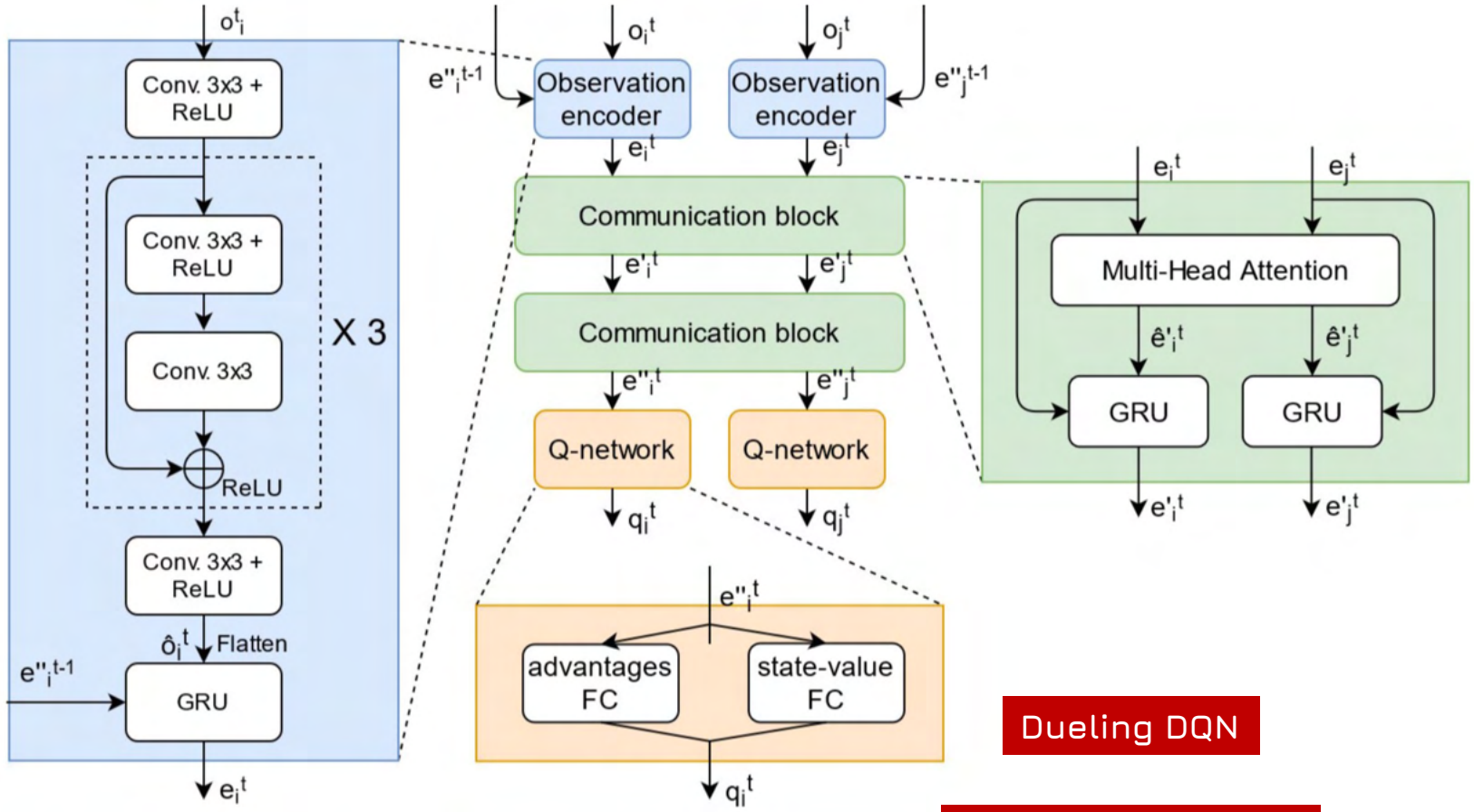
(d) NBOR (e) Hierarchical

PICO (Li et al. 2022)

Sheng, Junjie, et al. "Learning structured communication for multi-agent reinforcement learning." *Autonomous Agents and Multi-Agent Systems* 36.2 (2022): 50.

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

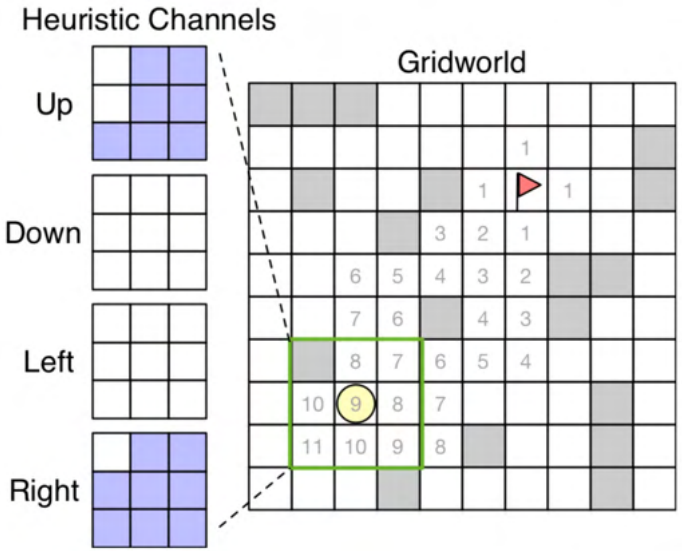
- Attention-based neighborhood communication



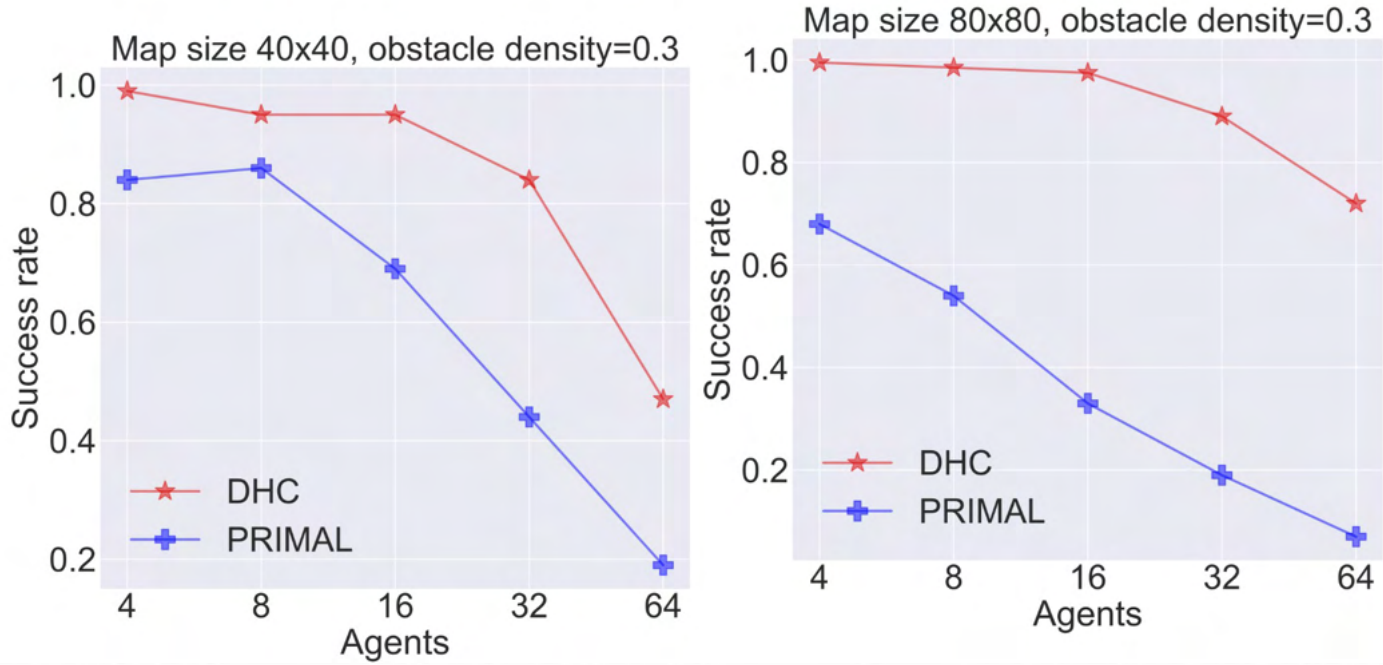
Dueling DQN

DHC (Ma et al. 2021)

Heuristic information



Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

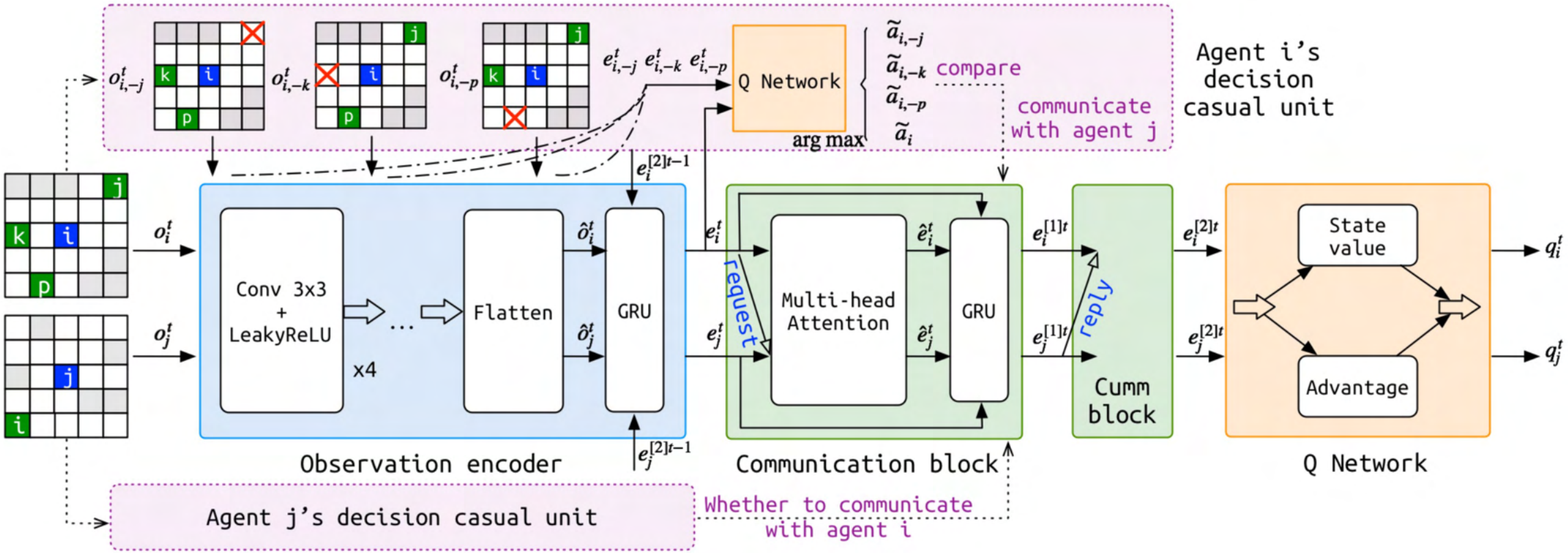


Average Step	Map size 40 × 40			Map size 80 × 80		
Agents	ODrM*	DHC	PRIMAL	ODrM*	DHC	PRIMAL
4	50.00	**52.33**	79.08	93.40	**96.72**	134.86
8	52.17	**63.90**	76.53	104.92	**109.24**	153.20
16	59.78	**79.63**	107.14	114.75	**122.54**	180.74
32	67.39	**100.10**	155.21	121.31	**138.32**	250.07
64	82.60	**147.26**	170.48	134.42	**163.50**	321.63

DHC (Ma et al. 2021)

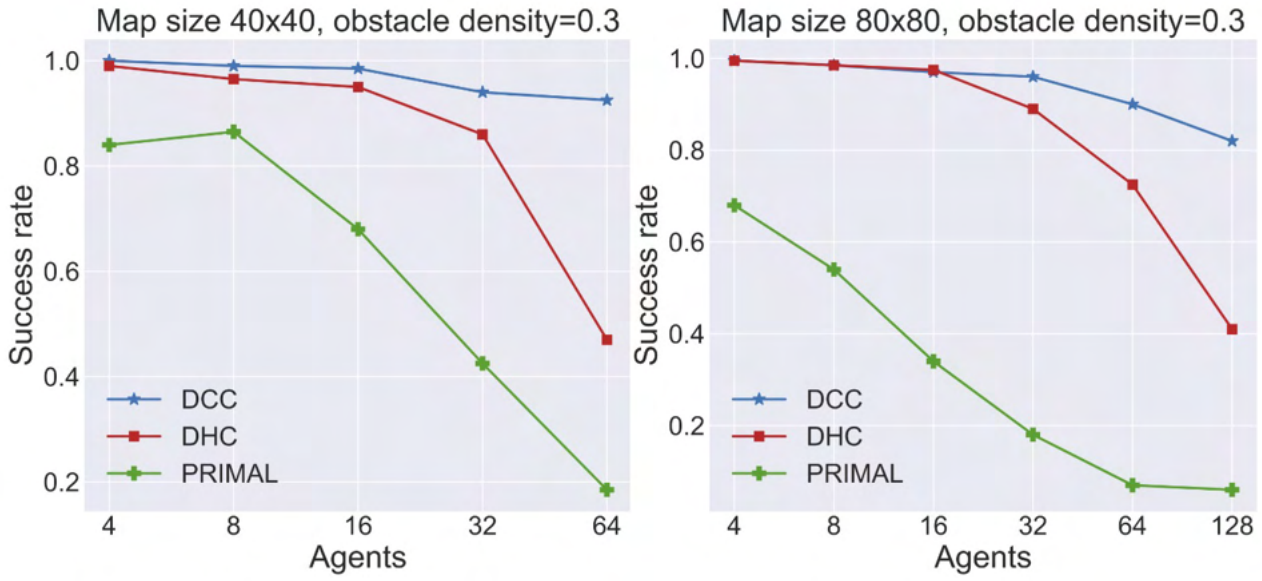
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- Selective communication



DCC (Ma et al. 2021)

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



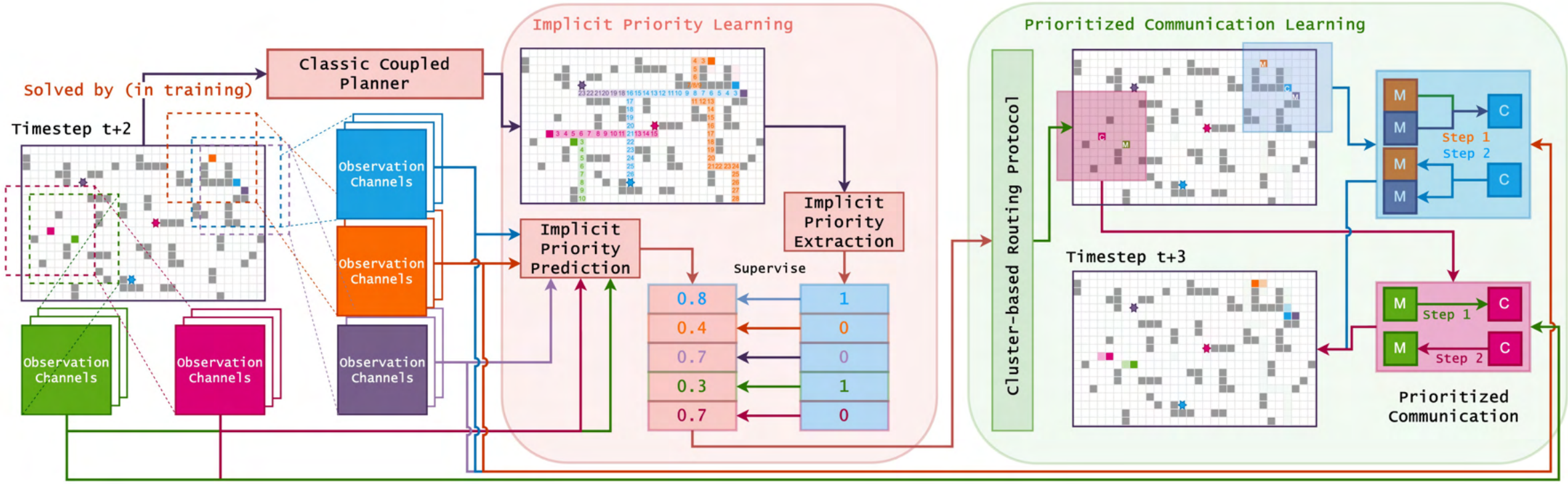
Average steps in 40 × 40 maps				
Agents	ODrM*	DCC	DHC	PRIMAL
4	50.00	**48.575**	52.33	79.08
8	52.17	**59.60**	63.90	76.53
16	59.78	**71.34**	79.63	107.14
32	67.39	**93.54**	100.10	155.21
64	82.60	**135.55**	147.26	170.48

Average steps in 80 × 80 maps				
Agents	ODrM*	DCC	DHC	PRIMAL
4	93.40	**93.89**	96.72	134.86
8	104.92	109.89	**109.24**	153.20
16	114.75	**122.24**	122.54	180.74
32	121.31	**132.99**	138.32	250.07
64	134.42	**159.67**	163.50	321.63
128	143.84	**192.90**	213.15	350.76

DCC (Ma et al. 2021)

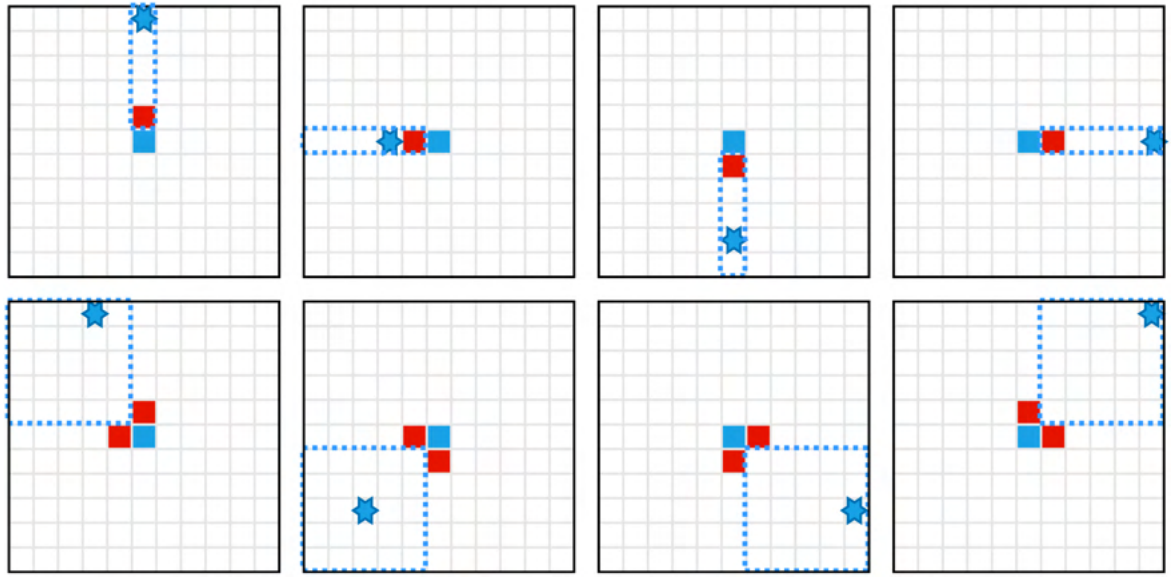
Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

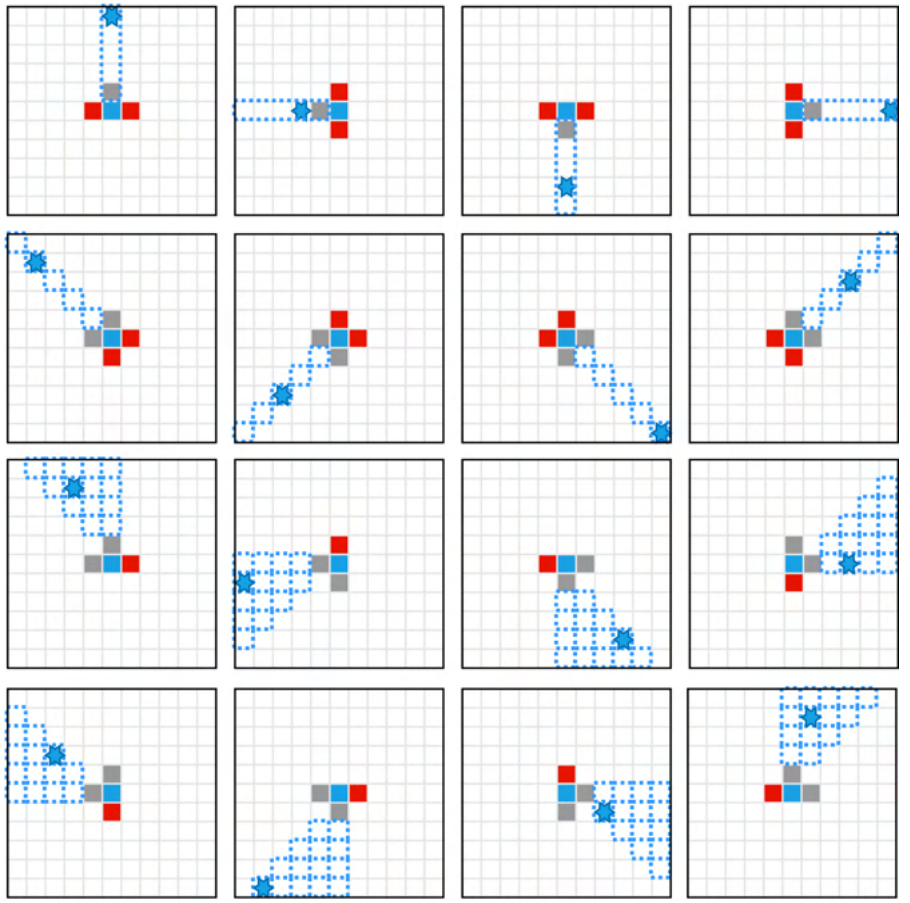
- Constructing hierarchical communication topology via the learned priorities



PICO (Li et al. 2022)

- Building an auxiliary imitation learning task to predict the priority of each agent by imitating classic coupled planner

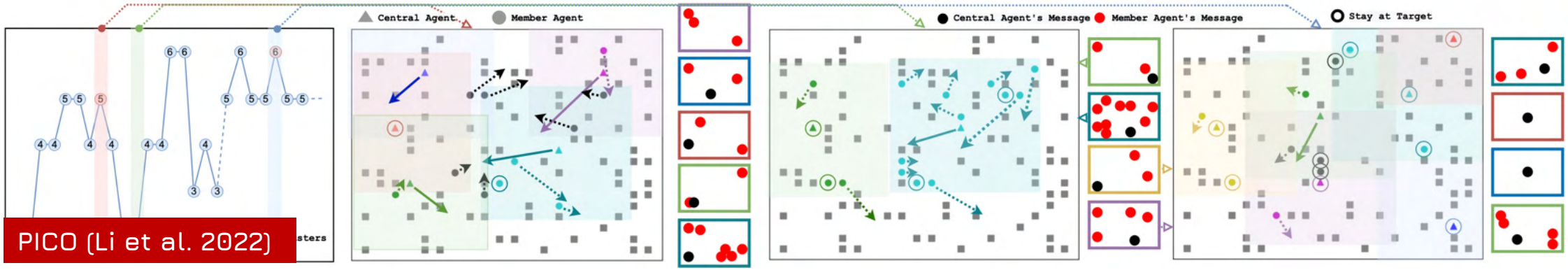




PICO (Li et al. 2022)

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

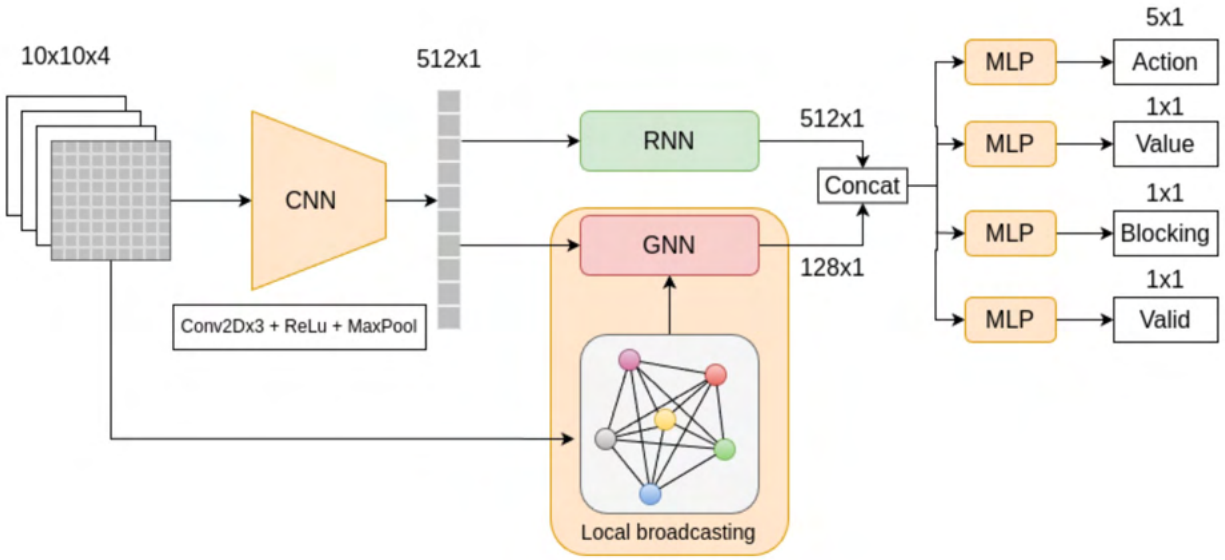
Methods	8 Agents (0%,10%,20%,30% Obstacle Densities)																							
	CA ↓				CO ↓				SR ↑				MS ↓				CR ↓				TM ↓			
ODrM* [11]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	25.55	23.95	25.86	29.45	0.0	0.0	0.0	0.0	112.75	108.83	117.49	117.38
PRIMAL [7]	1.94	3.02	3.03	5.98	0.0	0.01	0.01	0.0	93.0	90.0	48.0	15.0	34.86	62.69	148.59	234.22	0.06	0.05	0.02	0.03	221.3	223.01	345.09	565.11
DHC [20]	1.66	2.72	3.74	4.17	0.0	0.0	0.0	0.0	91.0	87.0	55.0	11.0	33.7	63.0	136.89	241.55	0.05	0.04	0.03	0.02	271.56	239.66	401.48	638.62
PICO-heuristic	2.81	2.7	3.68	3.4	0.0	0.0	0.0	0.0	90.0	86.0	50.0	10.0	33.57	59.81	141.94	236.99	0.05	0.04	0.03	0.02	248.19	252.88	451.09	650.62
PICO	0.59	0.62	1.31	2.32	0.0	0.0	0.0	0.0	100.0	96.0	55.0	25.0	27.45	41.81	134.7	204.83	0.02	0.02	0.01	0.01	123.96	142.69	290.13	463.41
Methods	16 Agents (0%,10%,20%,30% Obstacle Densities)																							
	CA ↓				CO ↓				SR ↑				MS ↓				CR ↓				TM ↓			
ODrM* [11]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	99.0	26.54	27.06	28.51	33.47	0.0	0.0	0.0	0.0	216.35	224.4	228.96	265.0
PRIMAL [7]	6.59	8.29	11.63	17.64	0.02	0.04	0.06	0.06	92.0	88.0	50.0	3.0	57.47	71.82	176.16	249.37	0.12	0.11	0.07	0.07	481.63	510.11	765.89	1396.23
DHC [20]	7.28	8.95	11.75	18.7	0.0	0.06	0.11	0.13	94.0	88.0	48.0	5.0	54.02	71.19	169.18	242.06	0.13	0.12	0.07	0.08	476.51	477.0	822.23	1503.58
PICO-heuristic	5.93	9.71	12.28	17.54	0.0	0.04	0.04	0.28	92.0	91.0	44.0	2.0	54.07	69.49	183.38	251.16	0.12	0.14	0.07	0.07	416.37	456.39	721.84	1538.51
PICO	2.98	3.98	4.96	8.0	0.0	0.02	0.02	0.03	100.0	95.0	57.0	7.0	30.83	49.06	144.67	240.15	0.1	0.08	0.03	0.03	251.39	298.55	526.4	1291.71
Methods	32 Agents (0%,10%,20%,30% Obstacle Densities)																							
	CA ↓				CO ↓				SR ↑				MS ↓				CR ↓				TM ↓			
ODrM* [11]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	91.0	29.09	29.67	32.88	38.93	0.0	0.0	0.0	0.0	441.34	462.07	505.13	530.07
PRIMAL [7]	26.23	30.48	47.27	98.33	0.0	0.36	1.56	2.07	92.0	72.0	9.0	0.0	53.91	108.02	244.76	256.0	0.47	0.28	0.19	0.38	958.39	1094.29	2226.83	3431.49
DHC [20]	27.04	34.79	49.29	95.51	0.02	0.65	4.28	4.67	92.0	62.0	3.0	0.0	48.76	135.78	242.67	256.0	0.52	0.26	0.21	0.37	957.06	1144.55	2009.0	3742.41
PICO-heuristic	27.49	31.01	49.02	93.5	0.01	0.64	3.76	5.18	91.0	63.0	0.0	0.0	49.45	121.05	256.0	256.0	0.56	0.27	0.19	0.36	1144.46	1179.09	2236.79	3723.63
PICO	14.8	20.62	36.28	83.38	0.0	0.21	1.28	1.63	100.0	75.0	19.0	0.0	38.09	96.5	224.54	256.0	0.4	0.22	0.15	0.33	550.87	774.01	1712.7	3175.57
Methods	64 Agents (0%,10%,20%,30% Obstacle Densities)																							
	CA ↓				CO ↓				SR ↑				MS ↓				CR ↓				TM ↓			
ODrM* [11]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	96.0	79.0	17.0	31.82	31.87	29.82	7.53	0.0	0.0	0.0	0.0	916.53	936.43	844.11	202.05
PRIMAL [7]	115.79	171.31	341.95	634.7	0.11	2.27	8.04	26.08	75.0	7.0	0.0	0.0	111.27	241.73	256.0	256.0	1.04	0.71	1.34	2.48	2418.58	3679.7	6611.44	9156.88
DHC [20]	106.78	146.92	312.92	622.58	0.25	12.75	38.37	145.89	72.0	0.0	0.0	0.0	108.76	256.0	256.0	256.0	0.98	0.57	1.22	2.43	2121.45	3261.68	6091.9	8407.14
PICO-heuristic	112.99	142.03	303.23	621.68	0.25	12.75	38.37	145.89	70.0	3.0	0.0	0.0	108.76	243.36	256.0	256.0	1.03	0.58	1.18	2.42	2201.45	3246.68	6162.9	8338.14
PICO	90.95	128.4	279.61	591.14	0.36	8.76	38.36	130.27	83.0	13.0	0.0	0.0	94.36	224.88	256.0	256.0	0.96	0.55	1.09	2.31	1472.92	2621.25	5342.04	7713.69



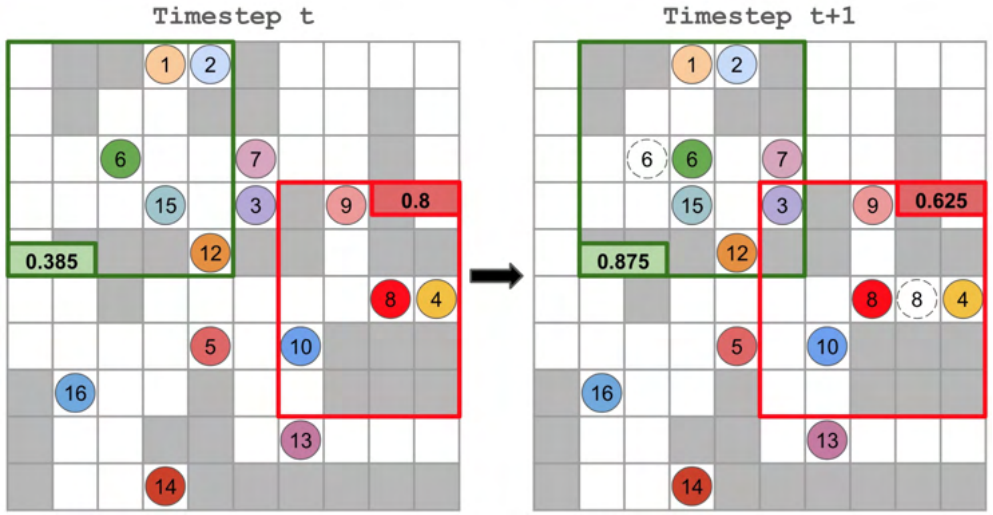
PICO (Li et al. 2022)

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

Local broadcasting with GNNs



Expert-guided reward



CRAMP (Pham et al. 2024)

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

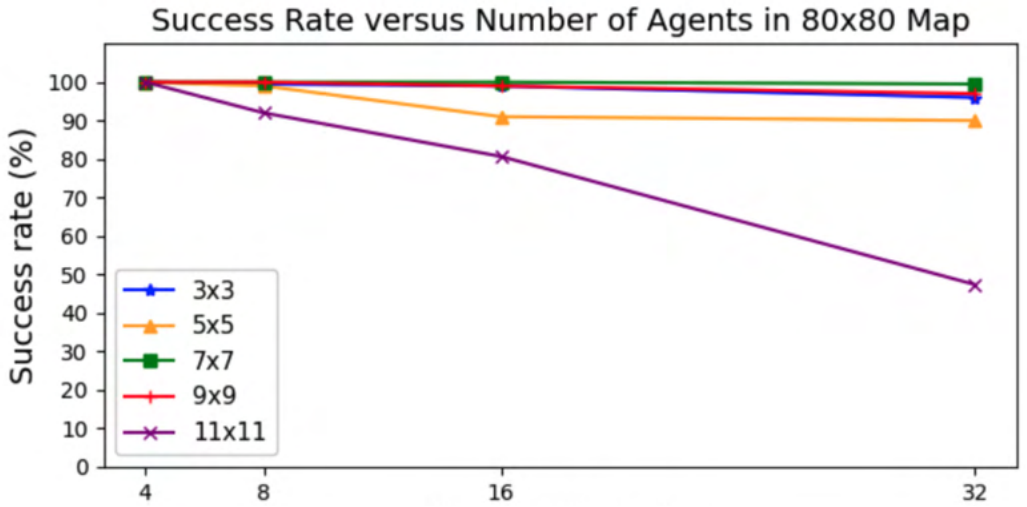
CRAMP (Pham et al. 2024)

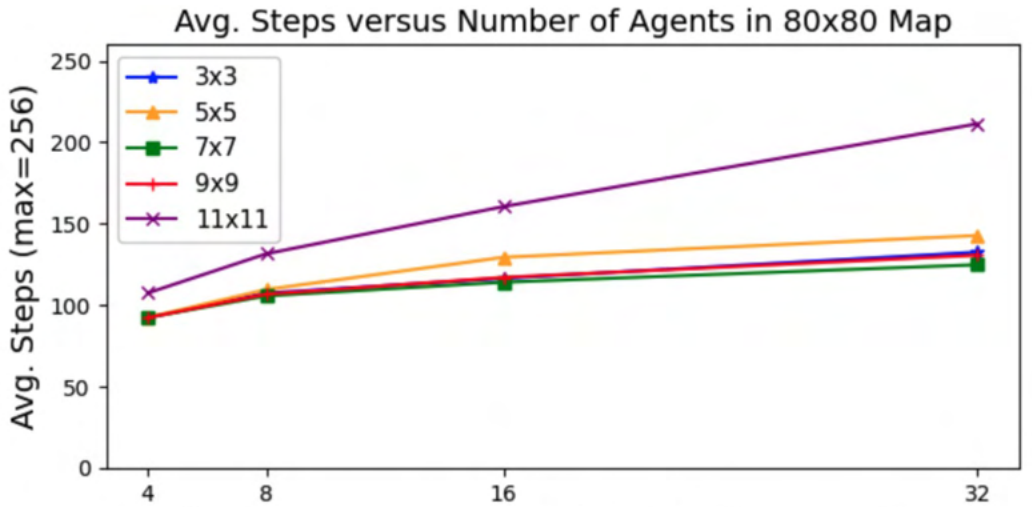
Methods	8 agents															
	Success rate				Makespan				Total moves				Collision count			
	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
PRIMAL	93	90	48	15	35	63	149	234	221	223	345	565	1.94	3.02	3.03	5.98
DHC	91	87	55	11	34	63	137	242	272	240	401	639	1.66	2.72	3.74	4.17
PICO	100	96	55	25	27	42	135	205	124	143	290	463	0.59	0.62	1.31	2.32
CPL	100	99	94	65	25	31	45	124	111	121	150	282	0.30	0.60	1.20	3.20
CRAMP	100	100	95	64	27	27	40	55	117	117	144	156	0.48	0.60	0.90	1.27
Methods	16 agents															
	Success rate				Makespan				Total moves				Collision count			
	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
PRIMAL	92	88	50	3	57	72	176	249	482	510	766	1396	6.59	8.29	11.63	17.64
DHC	94	88	48	5	54	71	169	242	477	477	822	1504	7.28	8.95	11.75	18.70
PICO	100	95	57	7	31	49	145	240	251	299	526	1292	2.98	3.98	4.96	8.00
CPL	100	95	81	22	27	41	84	213	221	249	374	780	2.80	3.70	5.30	16.10
CRAMP	100	98	83	23	30	33	45	88	236	253	275	388	2.60	3.20	3.70	5.73
Methods	32 agents															
	Success rate				Makespan				Total moves				Collision count			
	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
PRIMAL	92	72	9	0	54	108	245	-	958	1094	2227	-	26.23	30.48	47.27	-
DHC	92	62	3	0	49	136	243	-	957	1145	2009	-	27.04	34.79	49.29	-
PICO	100	75	19	0	38	97	225	-	551	774	1713	-	14.80	20.62	36.28	-
CPL	100	92	50	0	32	58	159	-	471	564	1032	-	11.90	17.40	30.30	-
CRAMP	100	82	34	4	37	47	85	170	489	552	754	1130	11.09	16.52	23.37	26.50
Methods	64 agents															
	Success rate				Makespan				Total moves				Collision count			
	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
PRIMAL	75	7	0	0	111	242	-	-	2419	3680	-	-	115.79	171.31	-	-
DHC	72	0	0	0	109	-	-	-	2121	-	-	-	106.78	-	-	-
PICO	83	13	0	0	94	225	-	-	1473	2621	-	-	90.95	128.40	-	-
CPL	80	20	0	0	92	128	-	-	1230	2204	-	-	84.00	109.00	-	-
CRAMP	78	27	2	0	90	125	278	-	1383	1877	2762	-	80.92	104.25	159.00	-

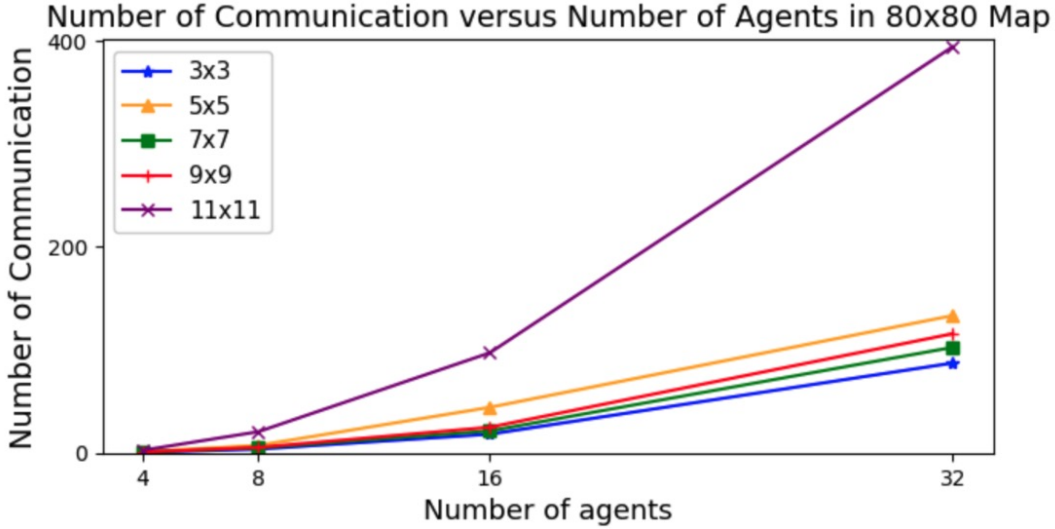
Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

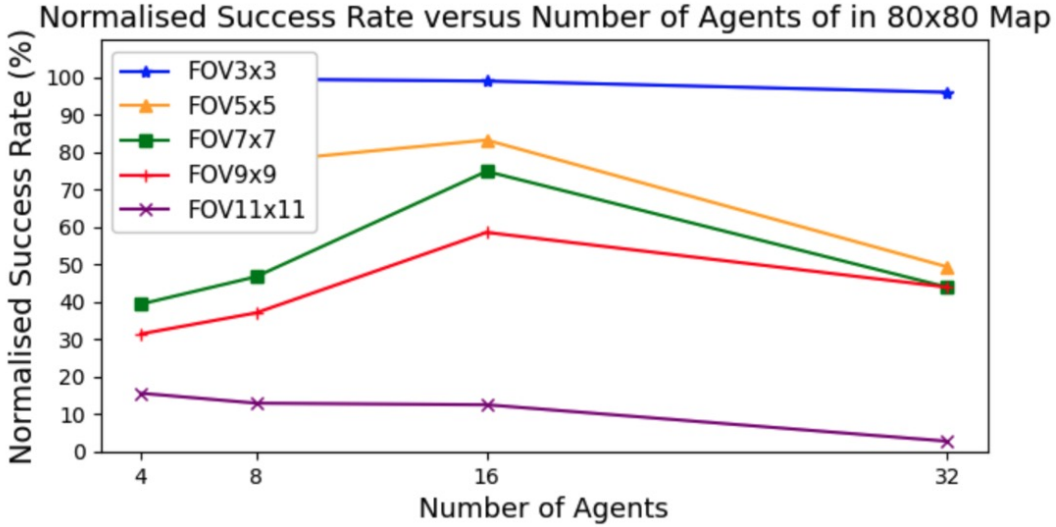
- Small FOV is enough!

Cheng et al. 2023





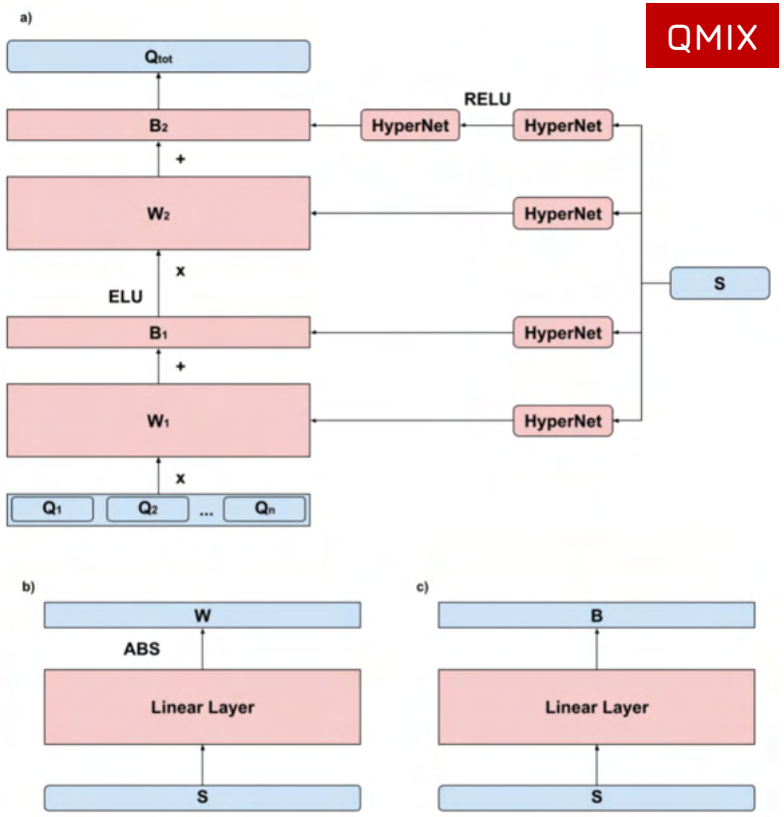




Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

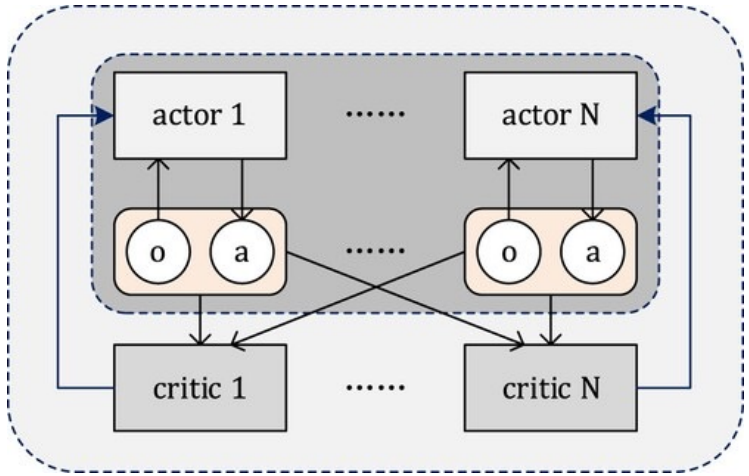
- Promoting cooperation with the centralized critic

QMIX



Davydov et al. 2021

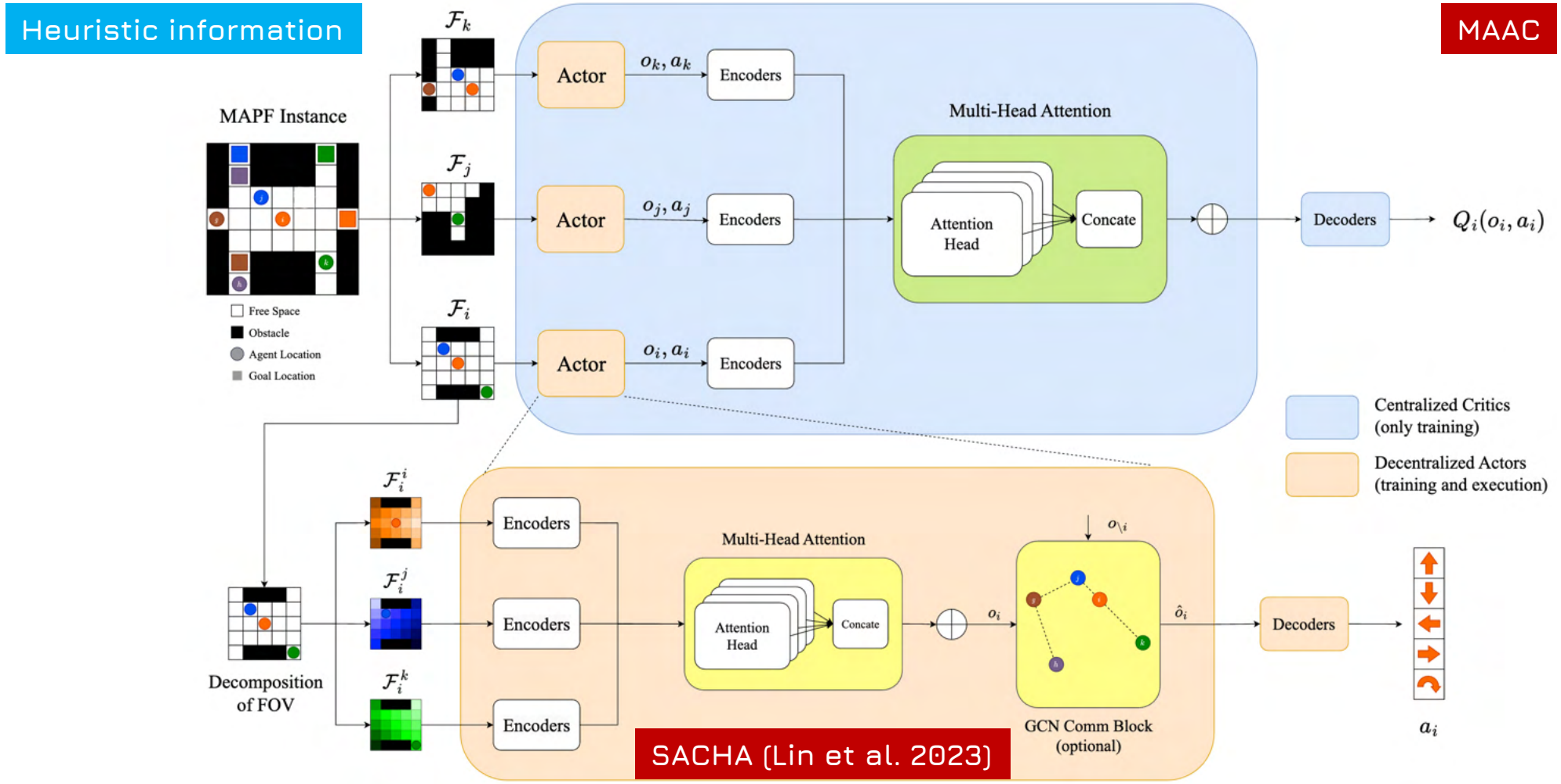
MADDPG



Hu et al. 2023

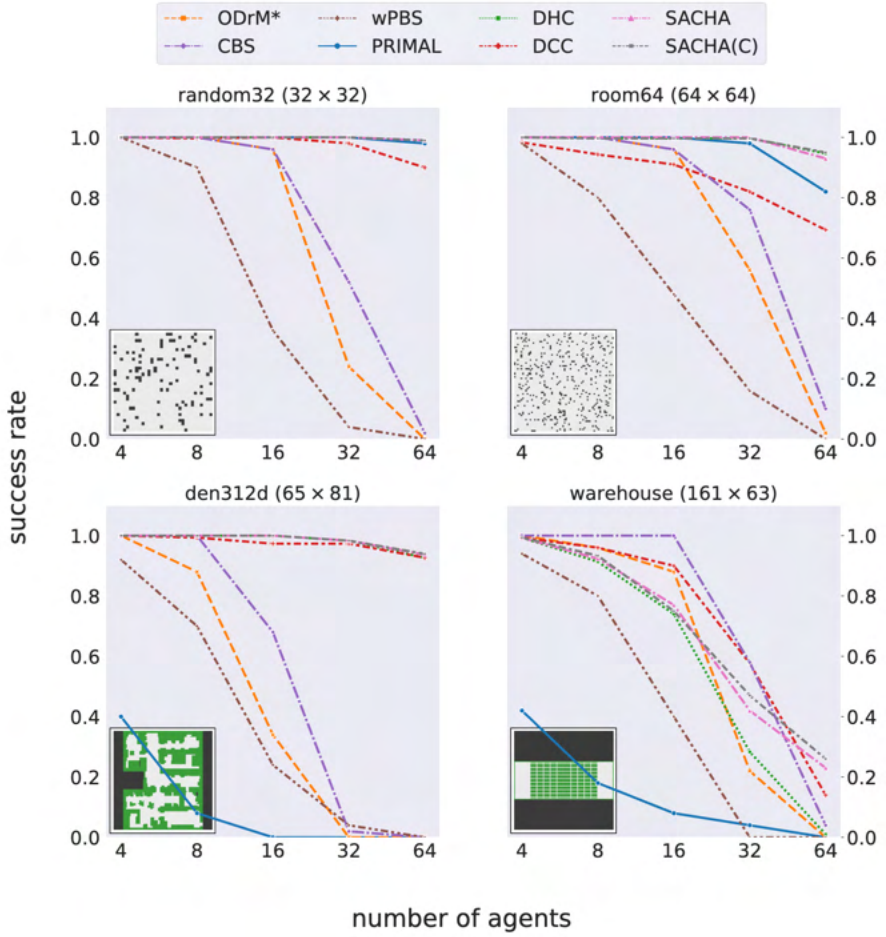
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- Promoting cooperation with the centralized critic



Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- Promoting cooperation with the centralized critic

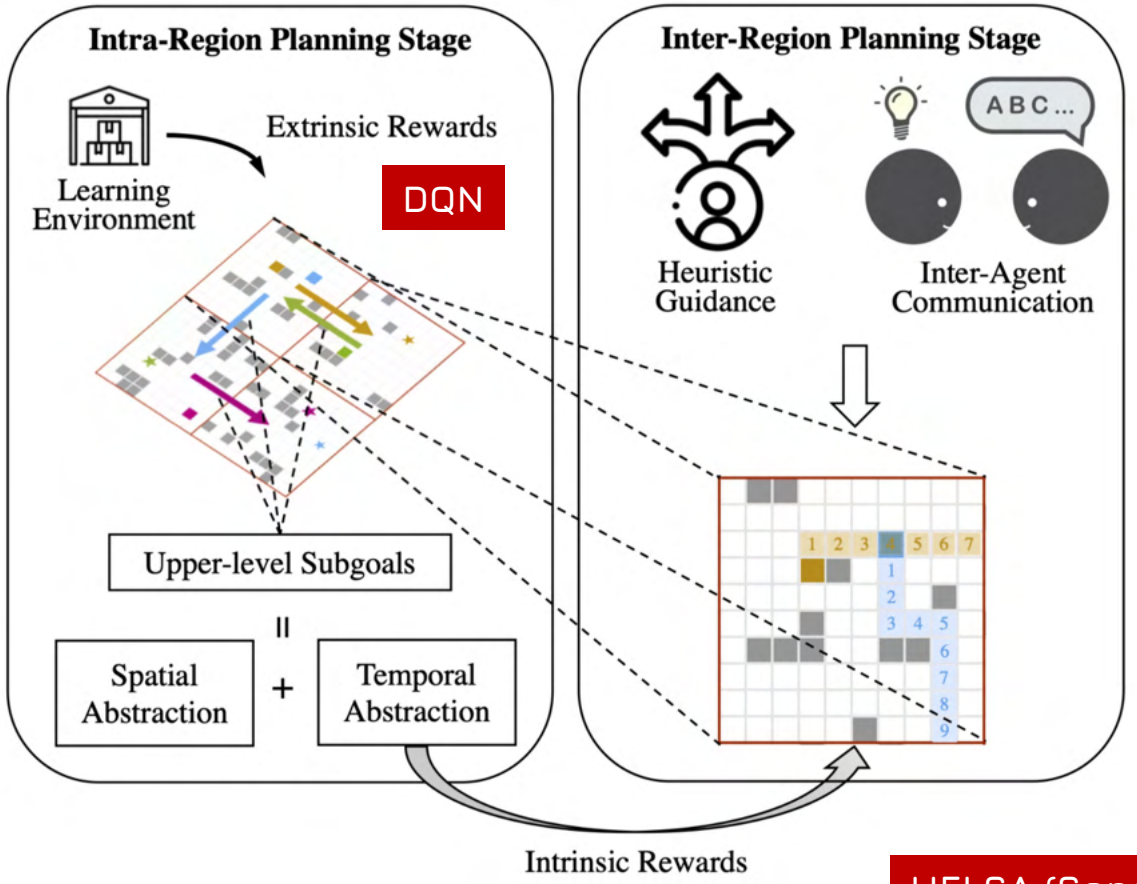


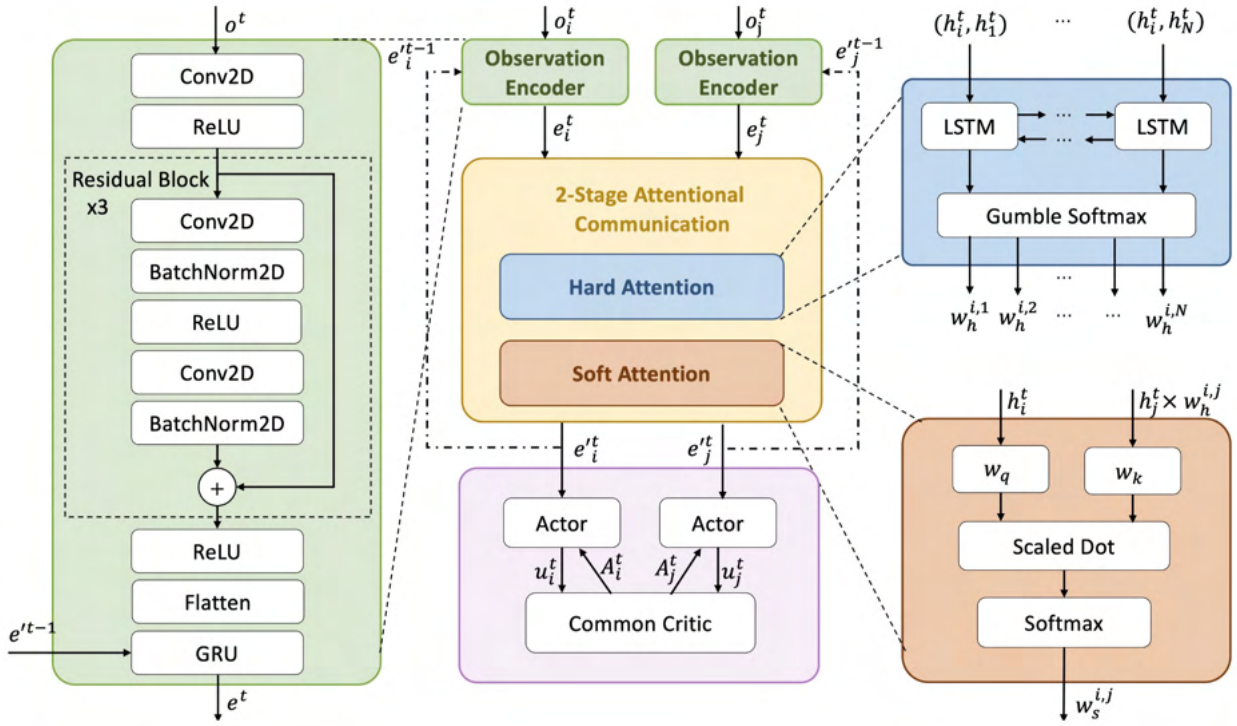
Map	Agents	Average Step per Agent							
		CBS (120s)	ODrM* (20s)	wPBS (120s)	PRI MAL	DHC	DCC	SAC HA	SAC HA(C)
random32	4	21.82	21.82	22.90	32.96	35.70	32.83	**29.93**	31.03
	8	21.38	21.37	46.06	38.62	42.64	39.56	**36.34**	38.30
	16	31.16	31.26	172.12	45.12	48.67	43.56	41.71	**41.30**
	32	133.86	199.47	246.61	50.34	52.17	56.11	50.26	**47.72**
	64	251.30	~~256.00~~	~~256.00~~	69.40	**66.05**	88.79	76.47	74.48
random64	4	42.94	42.95	48.14	67.82	71.04	70.80	**65.47**	67.10
	8	42.74	42.80	84.52	74.68	82.43	88.94	**70.49**	72.38
	16	51.51	51.52	154.47	89.22	94.22	102.27	83.74	**82.17**
	32	94.36	136.67	222.08	98.02	103.05	126.71	95.67	**93.08**
	64	234.66	251.65	~~256.00~~	105.12	120.68	154.72	99.02	**96.42**
den312d	4	51.74	51.76	69.32	196.54	86.56	82.99	**78.33**	81.43
	8	55.50	78.74	116.32	245.02	100.70	97.95	**84.24**	89.73
	16	118.97	186.44	208.28	~~256.00~~	109.24	108.29	97.86	**96.74**
	32	251.86	~~256.00~~	248.06	~~256.00~~	124.38	119.15	111.28	**104.30**
	64	~~256.00~~	~~256.00~~	~~256.00~~	~~256.00~~	153.17	145.21	**140.79**	142.97
warehouse	4	77.79	77.79	104.41	355.80	146.12	135.89	**131.43**	134.59
	8	83.48	100.37	170.46	451.82	198.82	169.50	**164.83**	166.72
	16	81.64	133.59	340.18	492.04	281.37	208.72	**192.30**	198.72
	32	262.15	417.22	~~512.00~~	505.58	432.28	**335.81**	370.65	354.33
	64	494.93	~~512.00~~	~~512.00~~	~~512.00~~	~~512.00~~	473.92	449.83	**437.29**

SACHA (Lin et al. 2023)

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- High-level agent: generating the subgoal
- Low-level agent: generating the path





HELSA (Song et al. 2023)

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

- High-level agent: generating the subgoal
- Low-level agent: generating the path

Model	8 agents, 40-sized map, 0.2 density					32 agents, 80-sized map, 0.2 density					128 agents, 160-sized map, 0.2 density				
	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓
PRIMAL [4]	**1.0**	56.49	98.90	0.42	**0.0**	0.88	164.39	305.73	4.12	**0.0**	0.07	356.51	1007.08	113.06	4.27
DHC [6]	**1.0**	31.40	55.77	0.38	**0.0**	0.98	69.18	139.77	3.20	**0.0**	0.87	132.31	399.19	29.38	0.06
DCC [7]	**1.0**	**28.84**	**50.49**	0.40	**0.0**	0.98	**64.47**	**134.34**	5.91	0.01	0.67	149.50	567.41	37.48	**0.0**
HELSA	**1.0**	29.71	52.29	**0.21**	**0.0**	**1.0**	65.85	136.17	**0.54**	**0.0**	**0.97**	**126.51**	**296.14**	**3.69**	**0.0**
Model	288 agents, 240-sized map, 0.2 density					512 agents, 320-sized map, 0.2 density					800 agents, 320-sized map, 0.2 density				
	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓	SR ↑	AS ↓	MS ↓	CA ↓	CO ↓
PRIMAL [4]	0.0	530.06	1536.0	593.59	34.48	0.0	736.50	2048.0	1498.20	173.49	-	-	-	-	-
DHC [6]	0.70	193.13	804.55	99.52	**0.01**	0.53	252.62	1304.48	236.22	0.30	0.40	315.08	1906.36	468.61	0.71
DCC [7]	0.19	235.32	1375.04	151.88	12.97	0.04	300.78	2020.76	423.40	57.41	-	-	-	-	-
HELSA	**0.93**	**175.56**	**629.58**	**49.41**	0.03	**0.87**	**221.17**	**935.99**	**101.78**	**0.04**	**0.74**	**268.83**	**1211.15**	**269.67**	**0.37**

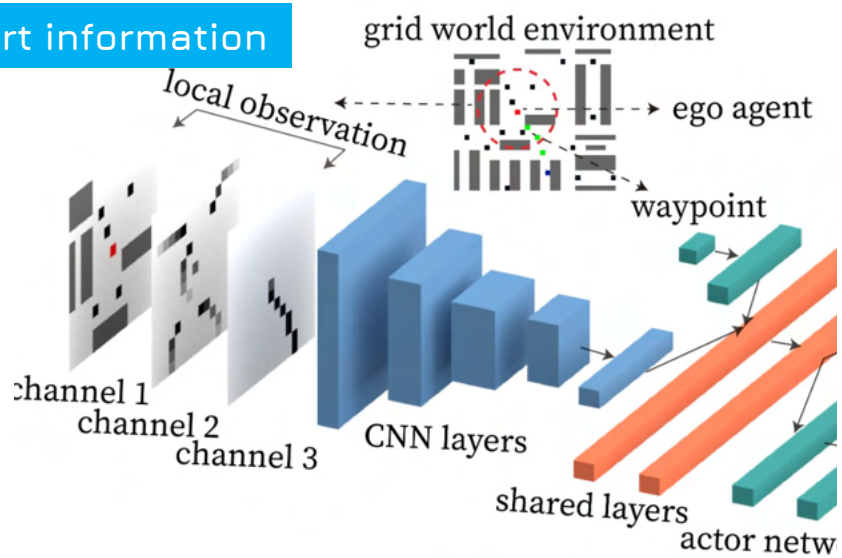
HELSA (Song et al. 2023)

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

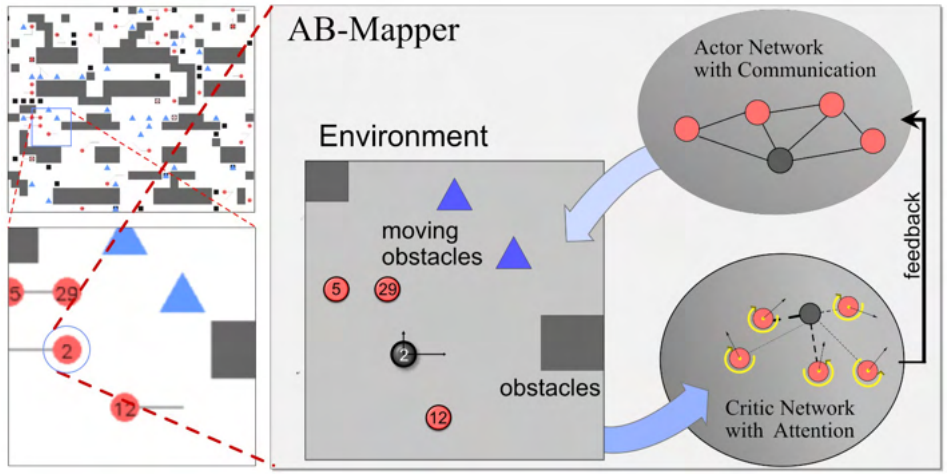
Initialize agents' model weights $\Theta = \{\Theta_1, ..., \Theta_N\}$
repeat
  Set accumulated reward $R_1^{(k)}, ..., R_N^{(k)} = 0$
  // update model parameters via A2C algorithm
  for $k = 1, ..., K$ **do**
    for each agent i **do**
      Executing the current policy π_{Θ_i} for T timesteps, collecting action, observation and reward $\{a_i^t, \boldsymbol{o}_i^t, r_i^t\}$, where $t \in [0, T]$
      Compute return $R_i = \sum_{t=0}^{T} \gamma^t r_i^t$
      Estimate advantage $\hat{A}_i = R - V^{\pi_{\Theta_i}}(\boldsymbol{o}_i)$
      Compute gradients $\nabla_{\Theta_i} J = \mathbb{E}[\nabla_{\Theta_i} \log \pi_{\Theta_i} \hat{A}_i]$
      Update Θ_i based on gradients $\nabla_{\Theta_i} J$
    end for
    $R_i^{(k)} = R_i^{(k)} + R_i$
  end for
  Normalize accumulated reward to get $\bar{R}_1^{(k)}, ..., \bar{R}_N^{(k)}$
  Find maximum reward $\bar{R}_j^{(k)}$ with agent index j
  // Evolutionary selection
  for each agent i **do**
    Sample m from uniform distribution between $[0, 1]$
    Compute evolution probability $p_i = 1 - \frac{\exp(\eta \bar{R}_i^{(k)})}{\exp(\eta \bar{R}_j^{(k)})}$
    if $m < p_i$ **then**
      $\Theta_i \leftarrow \Theta_j$

MAPPER (Liu et al. 2020)

Expert information

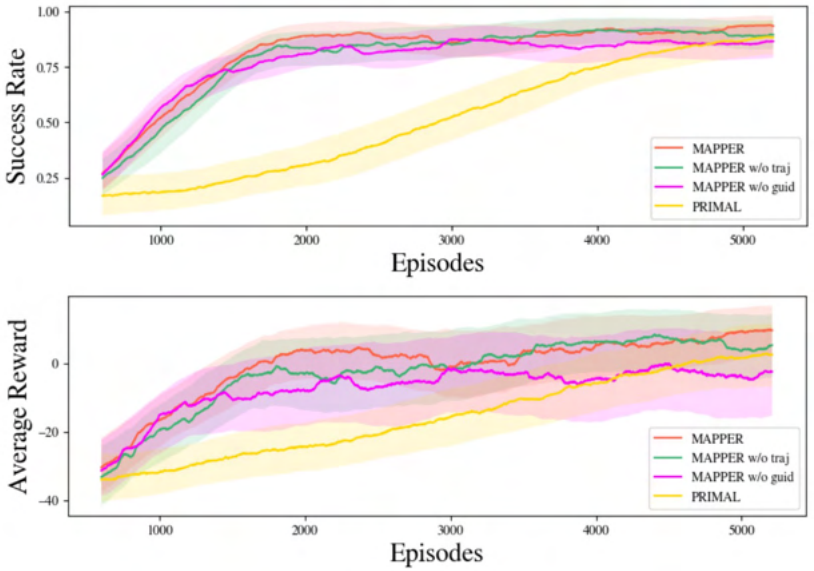


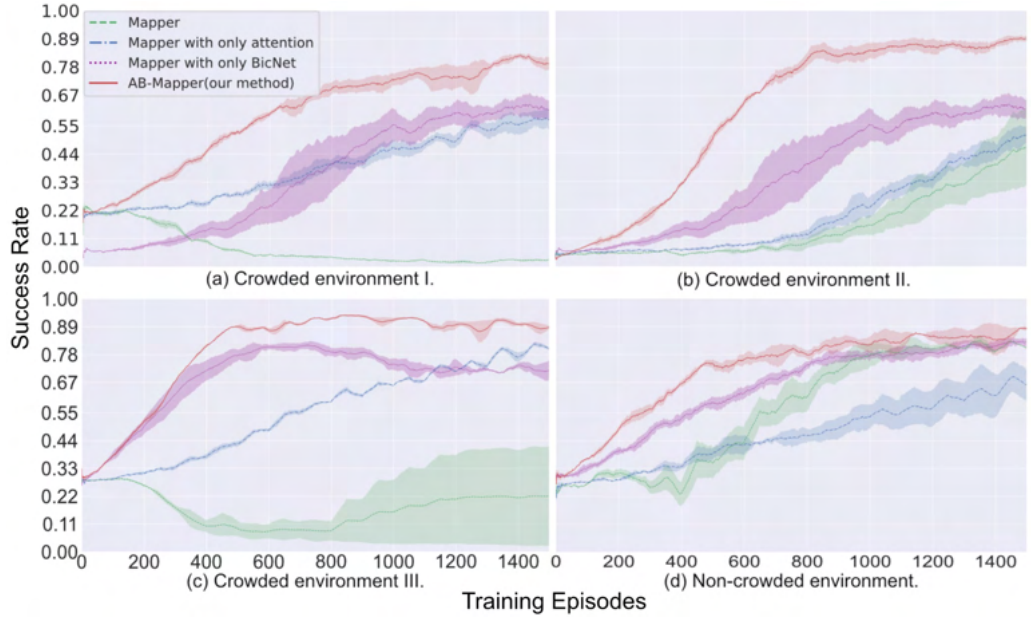
AB-MAPPER (Guan et al. 2021)



Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

Environment Setting			Success Rate				
map size	agent	dynamic obstacle	MAPPER	MAPPER w/o traj	MAPPER w/o guid	PRIMAL*	LRA*
20x20	15	10	**1.0**	0.971	0.877	0.964	0.996
20x20	35	30	**1.0**	0.961	0.836	0.980	0.999
20x20	45	30	**0.999**	0.854	0.607	0.971	0.997
60x65	70	100	**1.0**	0.256	0.516	0.352	**1.0**
60x65	130	140	**1.0**	0.473	0.221	0.404	0.992
120x130	150	40	**0.997**	0.324	0.211	0.389	0.994

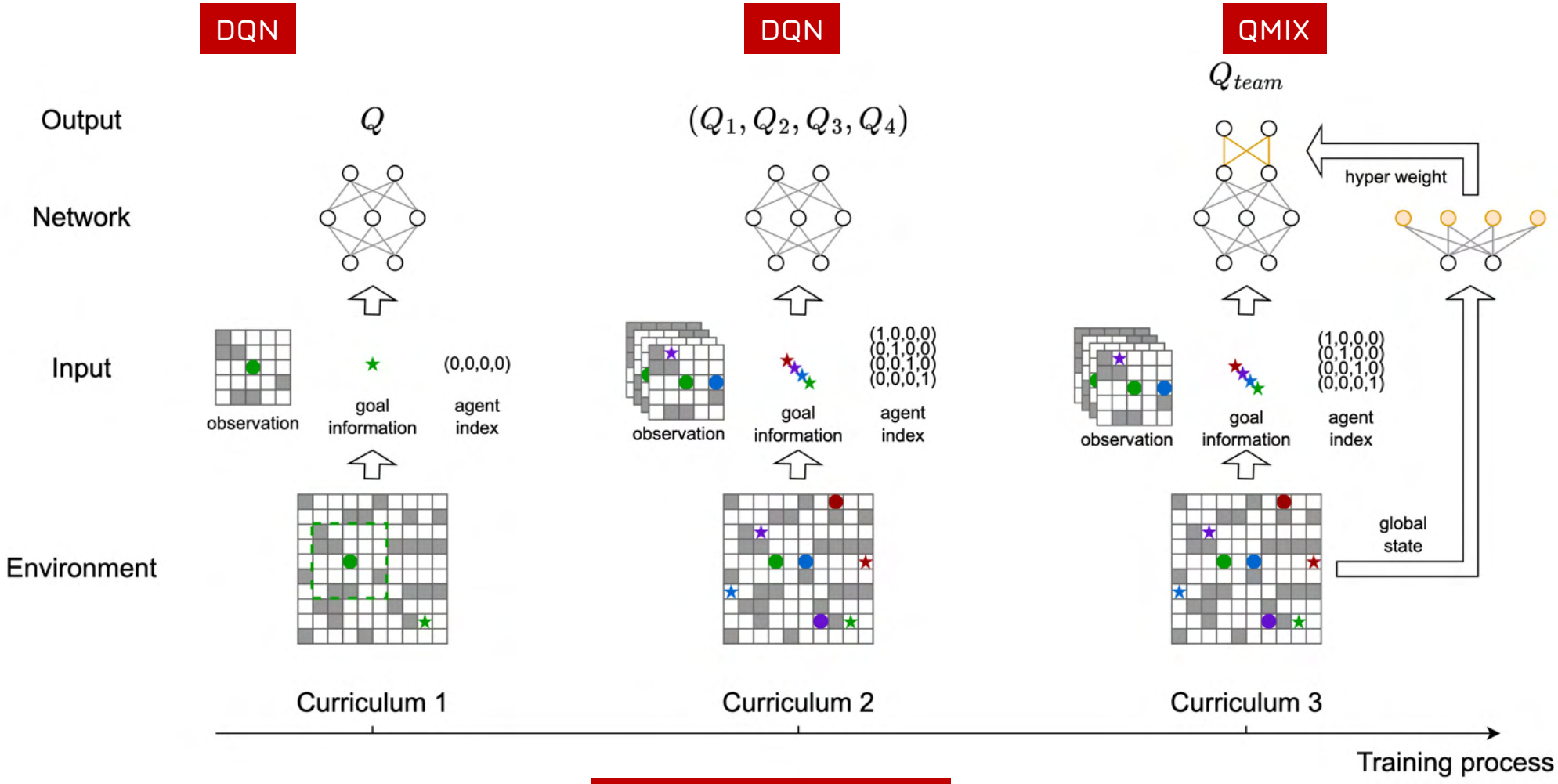




MAPPER (Liu et al. 2020)

AB-MAPPER (Guan et al. 2021)

Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



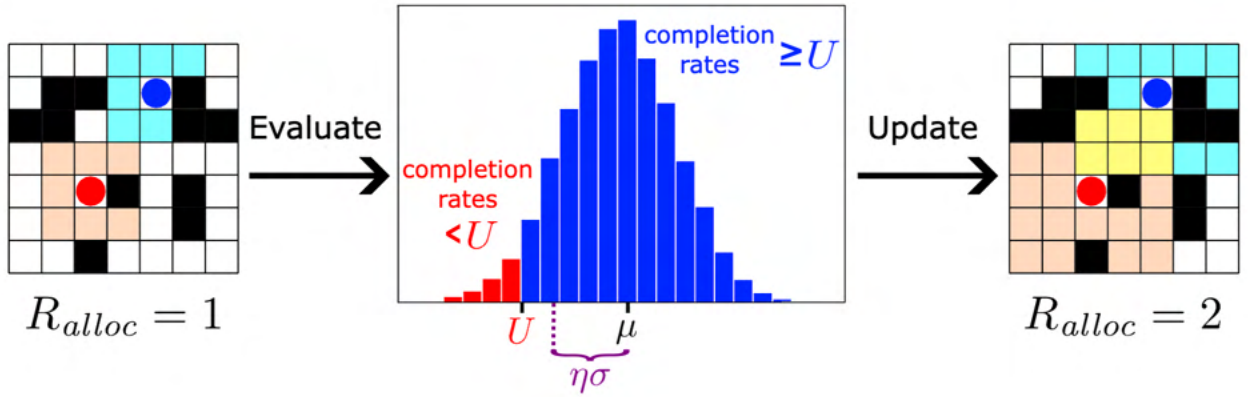
CPL (Zhao et al. 2023)

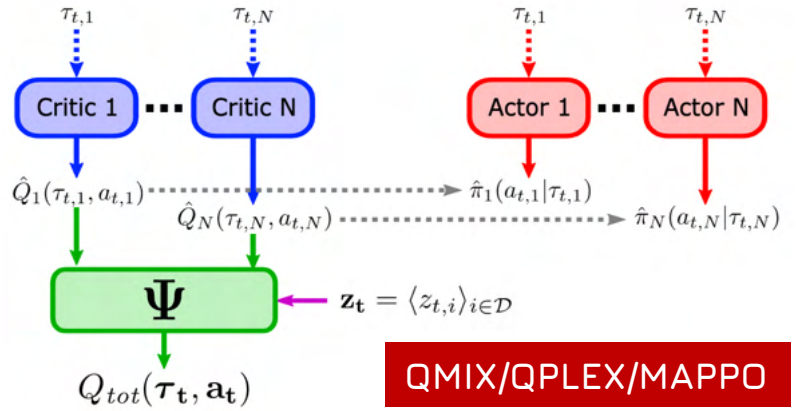
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

Methods	8 Agent (0%, 10%, 20%, 30% Obstacle Densities)																							
	CA↓				CO↓				SR↑				MS↓				CR↓				TM↓			
CPL	**0.3**	**0.6**	**1.2**	3.2	**0.0**	**0.0**	**0.0**	**0.0**	**100**	**99**	**94**	**65**	**25**	**31**	**45**	**124**	**0.02**	0.02	0.03	0.03	**111**	**121**	**150**	**282**
PRIMAL	1.9	3.0	3.0	6.0	**0.0**	**0.0**	**0.0**	**0.0**	93	90	48	15	35	63	149	234	0.06	0.05	0.02	0.03	221	233	345	565
PICO	0.6	**0.6**	1.3	**2.3**	**0.0**	**0.0**	**0.0**	**0.0**	100	96	55	25	27	42	135	205	**0.02**	**0.01**	**0.01**	**0.01**	124	143	290	463
Methods	**16 Agent (0%, 10%, 20%, 30% Obstacle Densities)**																							
	CA↓				CO↓				SR↑				MS↓				CR↓				TM↓			
CPL	**2.8**	**3.7**	5.3	16.1	**0.0**	**0.0**	**0.0**	**0.0**	**100**	**95**	**81**	**22**	**27**	**41**	**84**	**213**	**0.10**	0.09	0.06	0.08	**221**	**249**	**374**	**780**
PRIMAL	6.6	8.3	11.6	17.6	**0.0**	**0.0**	0.1	0.1	92	88	50	3	57	72	176	249	0.11	0.12	0.07	0.07	482	510	766	1396
PICO	3.0	3.9	**5.0**	**8.0**	**0.0**	**0.0**	**0.0**	**0.0**	**100**	**95**	57	7	31	49	145	240	**0.10**	**0.08**	**0.03**	**0.03**	251	299	526	1292
Methods	**32 Agent (0%, 10%, 20%, 30% Obstacle Densities)**																							
	CA↓				CO↓				SR↑				MS↓				CR↓				TM↓			
CPL	**11.9**	**17.4**	**30.3**	**45.6**	**0.0**	**0.0**	**0.0**	**0.0**	**100**	**92**	**50**	0	**32**	**58**	**159**	256	**0.38**	0.30	0.19	**0.18**	**471**	**564**	**1032**	3603
PRIMAL	26.2	30.5	47.3	98.3	**0.0**	0.4	1.6	2.1	92	72	9	0	54	108	245	256	0.49	0.28	0.19	0.38	958	1094	2227	3431
PICO	14.8	20.6	36.3	83.4	**0.0**	0.2	1.3	1.6	**100**	75	19	0	38	97	225	256	0.39	**0.21**	**0.16**	0.33	551	774	1713	**3176**
Methods	**64 Agent (0%, 10%, 20%, 30% Obstacle Densities)**																							
	CA↓				CO↓				SR↑				MS↓				CR↓				TM↓			
CPL	**84**	**109**	**101**	**109**	**0.0**	**0.0**	**0.0**	**0.0**	80	**20**	0	0	**92**	**218**	256	256	**0.91**	**0.50**	**0.39**	**0.43**	**1230**	**2204**	7630	8566
PRIMAL	116	171	342	635	0.1	2.3	8.0	36.1	75	7	0	0	111	242	256	256	1.04	0.71	1.34	2.48	2419	3680	6611	9157
PICO	91	128	280	591	0.4	8.8	38.4	130.3	**83**	13	0	0	94	225	256	256	0.96	0.57	1.09	2.31	1473	2621	**5342**	**7714**

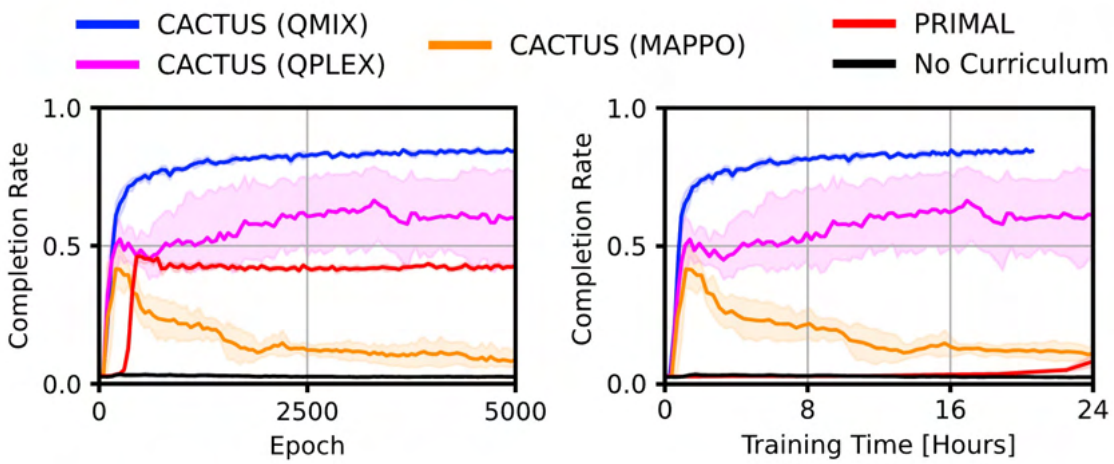
CPL (Zhao et al. 2023)

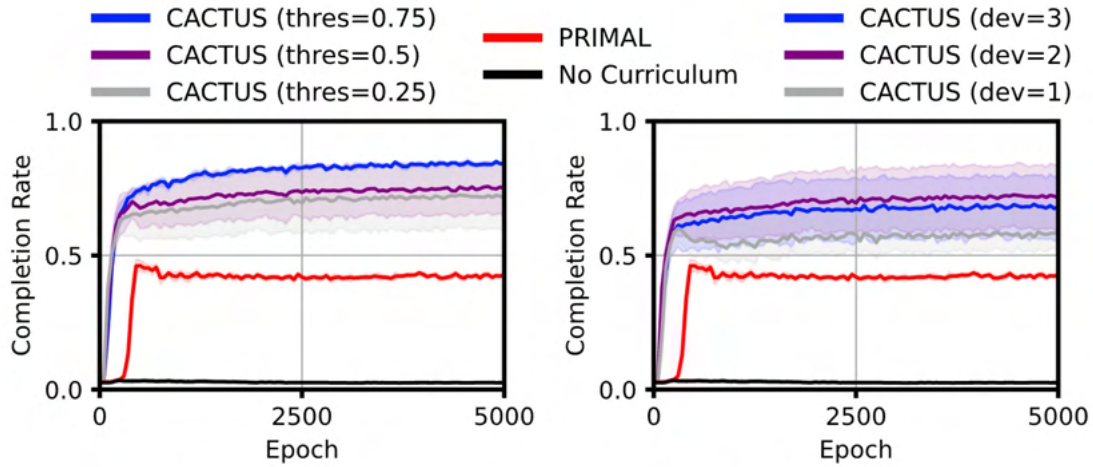
Independent learning with imitation ➡ communication ➡ cooperation
hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning





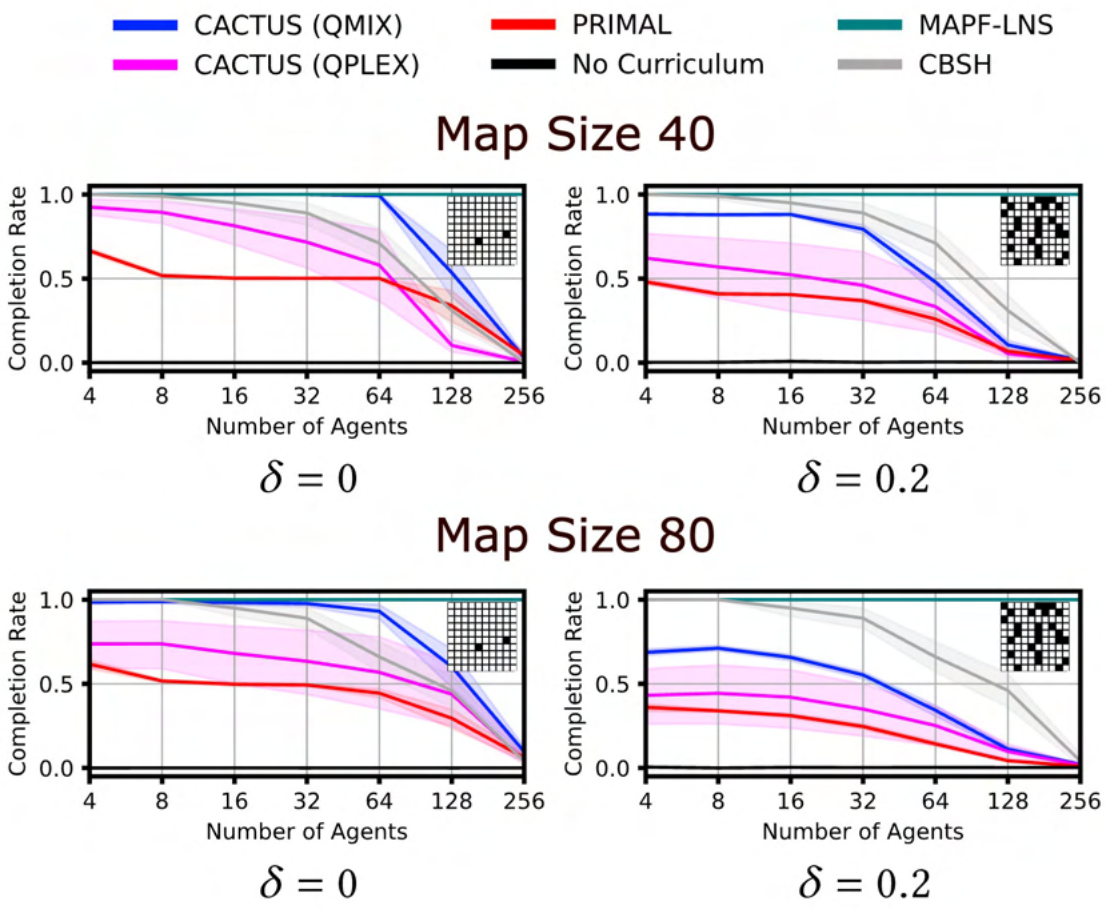
QMIX/QPLEX/MAPPO

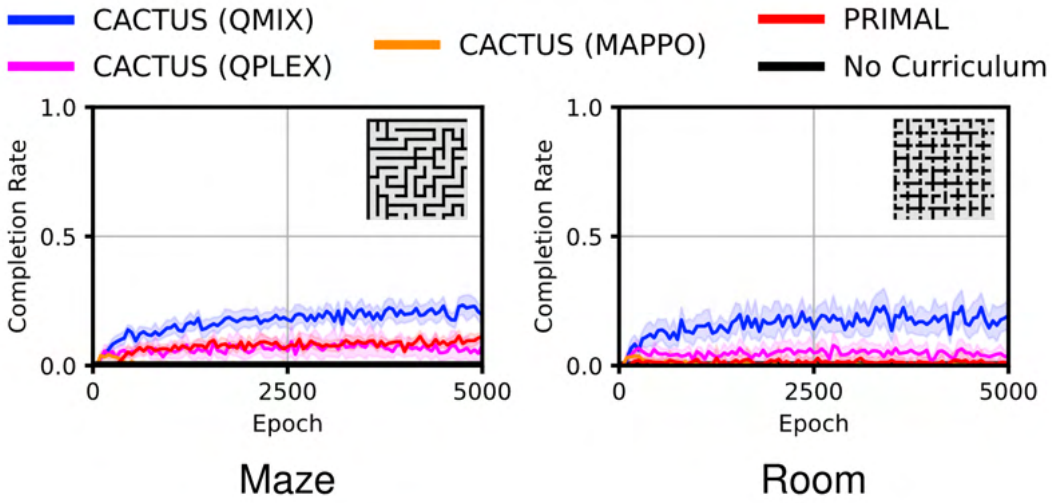




CACTUS (Phan et al. 2024)

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

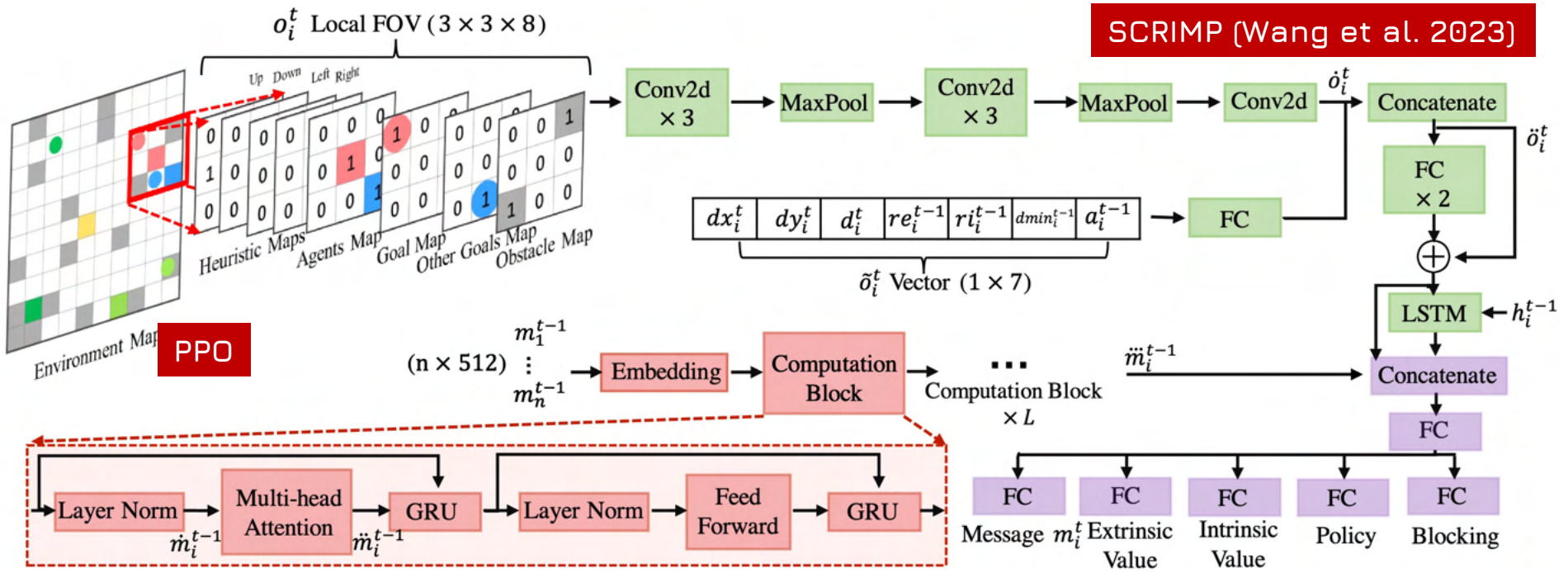




CACTUS (Phan et al. 2024)

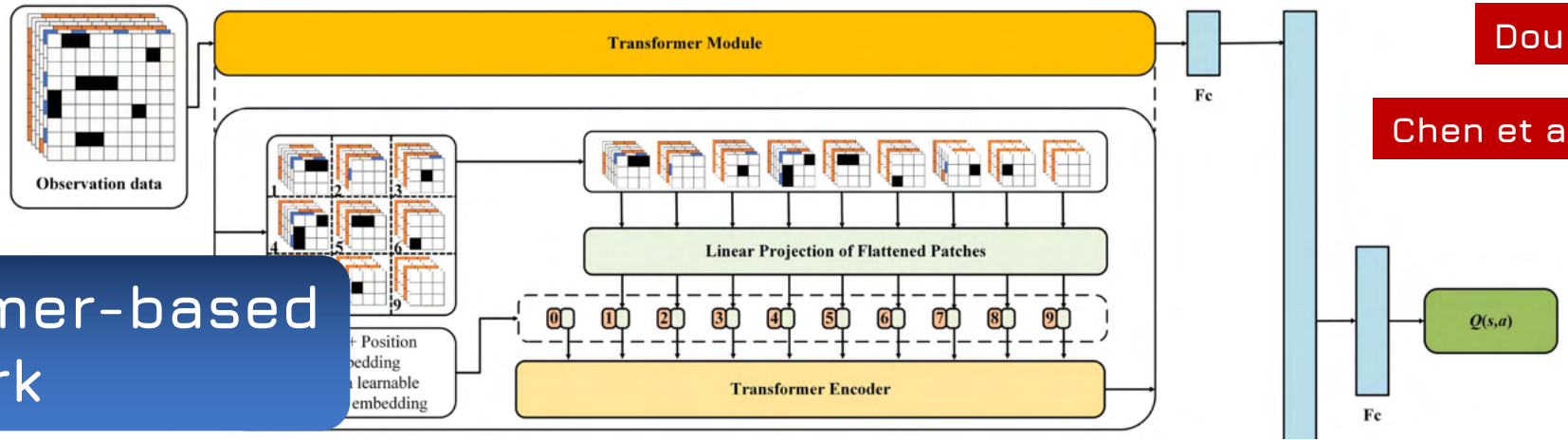
Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning

SCRIMP (Wang et al. 2023)



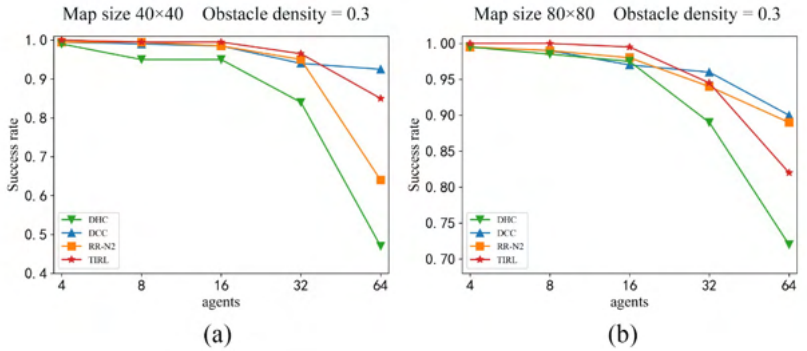
Double DQN

Chen et al. 2023a



Transformer-based RL network

Independent learning with imitation ➡ communication ➡ cooperation hierarchical ➡ evolutionary ➡ curriculum ➡ representation learning



Average Steps in 40×40 Map size					
Agents	*ODrM*(ϵ=10)*	*TIRL*	*DCC*	*DHC*	*PRIMAL*
4	47.86	**47.86**	48.575	52.33	79.08
8	55.47	**56.42**	59.60	63.9	76.53
16	61.89	**62.85**	71.34	79.63	107.14
32	66.94	**73.46**	93.54	100.1	155.21
64	85.27	**103.70**	135.55	147.26	170.48

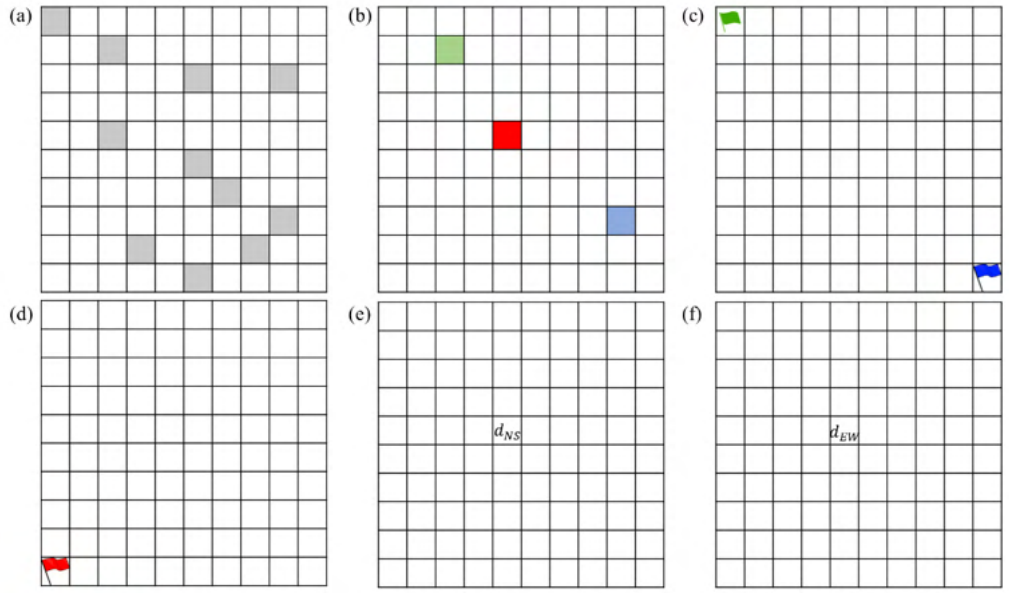
Average Steps in 80×80 Map size					
Agents	*ODrM*(ϵ=10)*	*TIRL*	*DCC*	*DHC*	*PRIMAL*
4	91.87	**91.87**	93.89	96.72	134.86
8	105.72	**105.72**	109.89	109.24	153.20
16	112.64	**114.06**	122.24	122.54	180.74
32	122.28	136.54	**132.99**	138.32	250.07
64	131.19	176.25	**159.67**	163.50	321.63

Chen et al. 2023a

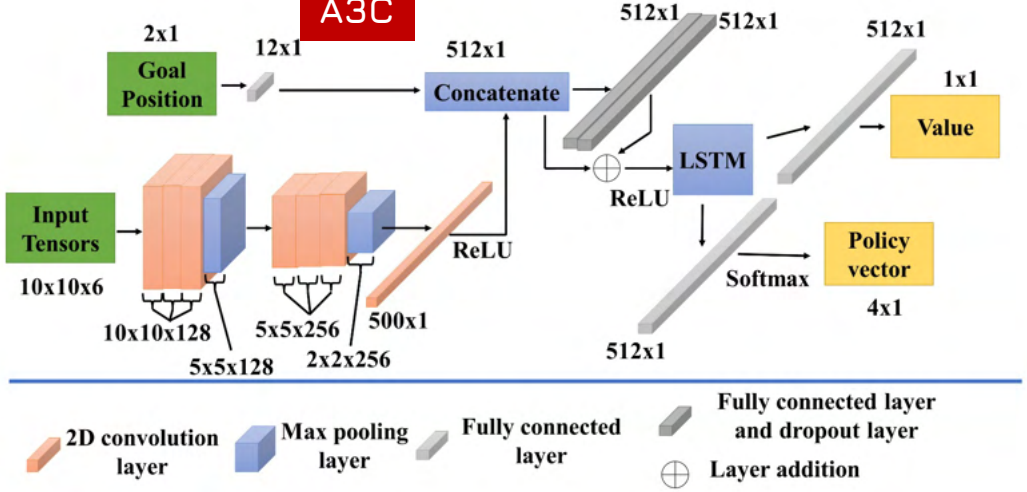
Methods	8 agents in 10 × 10 world with 0%, 15%, 30% static obstacle rate											
	EL ↓			MR ↑			CO ↓			SR ↑		
ODrM*	12.64(2.13)	13.72(2.46)	16.16(4.03)	-	-	-	0.00%(0.00%)	0.00%(0.00%)	0.00%(0.00%)	100%	100%	100%
DHC	14.34(5.08)	17.23(5.96)	29.82(20.93)	**8.00(0.00)**	7.98(0.14)	7.88(0.50)	**0.00%(0.00%)**	**0.00%(0.00%)**	**0.00%(0.00%)**	**100%**	98%	92%
PICO	17.35(12.39)	29.23(28.71)	35.04(26.93)	**8.00(0.00)**	7.51(0.72)	6.68(1.76)	**0.00%(0.00%)**	**0.00%(0.00%)**	**0.00%(0.00%)**	**100%**	63%	29%
SCRIMP	**12.05(2.53)**	**14.17(3.52)**	**18.56(7.80)**	**8.00(0.00)**	**8.00(0.00)**	**7.93(0.45)**	**0.00%(0.00%)**	0.04%(0.18%)	0.07%(0.26%)	**100%**	**100%**	**97%**
	32 agents in 30 × 30 world with 0%, 15%, 30% static obstacle rate											
ODrM*	42.97(4.62)	43.39(4.97)	53.95(10.43)	-	-	-	0.00%(0.00%)	0.00%(0.00%)	0.00%(0.00%)	100%	100%	97%
DHC	48.64(17.54)	52.48(7.10)	91.29(29.25)	**32.00(0.00)**	31.97(0.22)	31.61(1.10)	**0.00%(0.00%)**	**0.00%(0.00%)**	**0.00%(0.00%)**	**100%**	98%	78%
PICO	65.14(10.30)	-	-	30.26(1.18)	22.70(2.40)	12.20(2.53)	**0.00%(0.00%)**	**0.00%(0.00%)**	**0.00%(0.00%)**	14%	0%	0%
SCRIMP	**42.59(4.87)**	**43.90(5.11)**	**62.31(17.86)**	**32.00(0.00)**	**32.00(0.00)**	**31.94(0.34)**	**0.00%(0.00%)**	0.00%(0.02%)	0.32%(0.44%)	**100%**	**100%**	**96%**
	128 agents in 40 × 40 world with 0%, 15%, 30% static obstacle rate											
ODrM*	64.55(4.97)	67.77(6.64)	76.90(7.12)	-	-	-	0.00%(0.00%)	0.00%(0.00%)	0.00%(0.00%)	100%	100%	20%
DHC	93.00(39.24)	137.76(47.69)	-	127.92(0.31)	127.45(1.06)	97.83(16.08)	**0.00%(0.00%)**	**0.00%(0.00%)**	**0.00%(0.00%)**	93%	70%	0%
PICO	135.00(84.85)	-	-	126.59(3.88)	95.01(6.97)	58.11(8.79)	**0.00%(0.00%)**	0.024%(0.022%)	7.80%(8.52%)	2%	0%	0%
SCRIMP	**66.86(5.83)**	**73.15(8.53)**	**135.43(42.32)**	**128.00(0.00)**	**128.00(0.00)**	**121.95(13.16)**	0.01%(0.01%)	0.14%(0.13%)	3.72%(4.40%)	**100%**	**100%**	**58%**

SCRIMP (Wang et al. 2023)

Independent learning with postprocessing



A3C



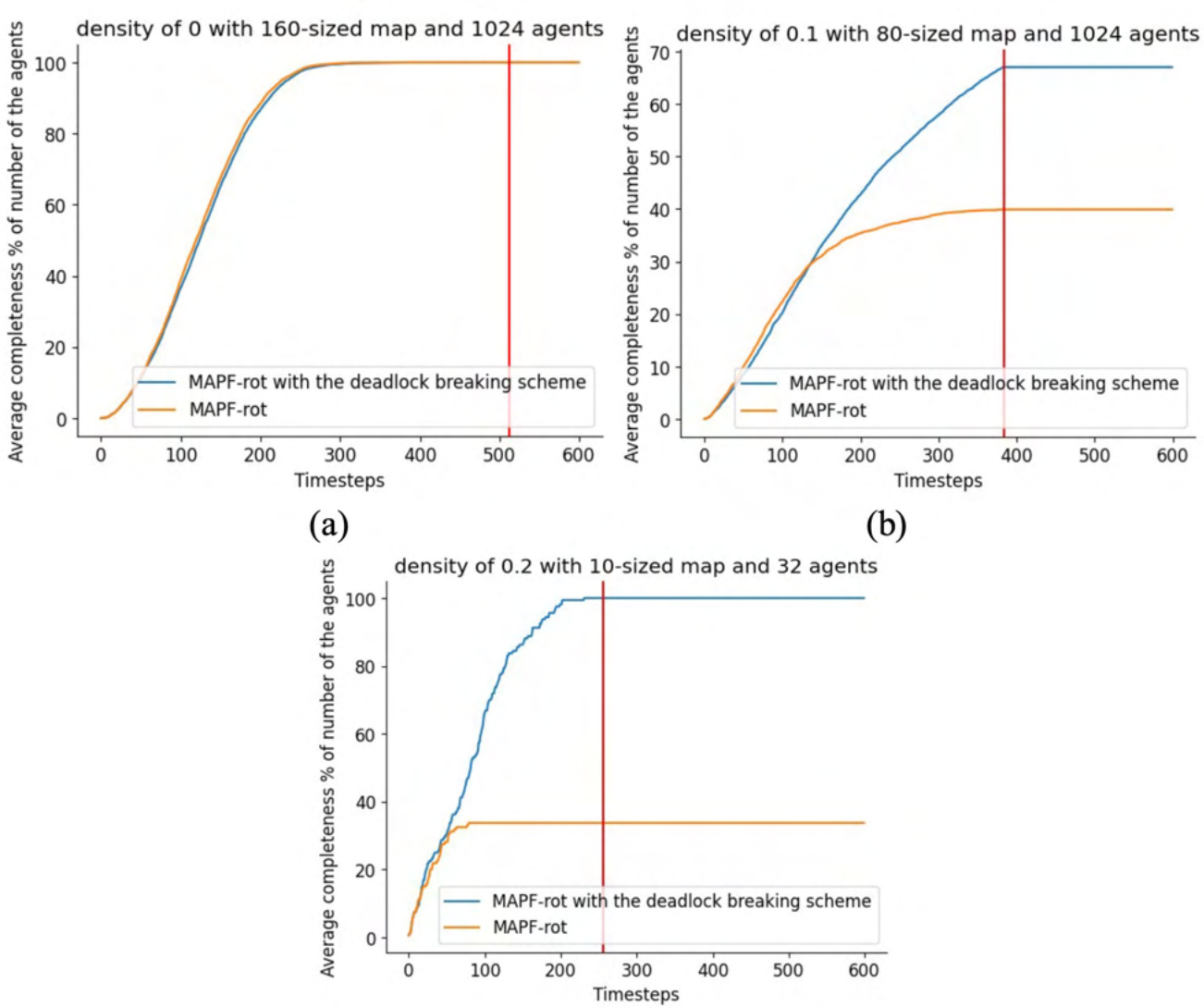
Scheme 2 Deadlock breaking scheme

function DeadlockBreaking
for each episode do
  initialize row vector $\boldsymbol{c_i} = (0\ 0\ 0\ 0)$ for action counting of each agent i
  for each agent i do
    get a policy vector P_i and action A_i from the proposed algorithm
    if agent i is staying at the same location:
      c_{ij} += 1 # j is the index of action A_i
    set $M_i = \sum_{j=0}^{3} c_{ij}$
    if $M_i = N$: # N is a preset threshold for action generation
      set $m_j = e^{\frac{P_{ij}}{1+c_{ij}/M_i}}$
      the probability of each action $p = \frac{m_j}{\sum_{j=0}^{3} m_j}$
      generate a random action A_i' based on the distribution p
      reset $\boldsymbol{c_i} = (0\ 0\ 0\ 0)$
      return A_i'
    else:
      return A_i

	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	95.60	95.60	102.80	116.40
32	127.00	129.60	143.20	141.80
128	NIL*	144.60	167.80	158.40
256	NIL*	179.80	216.25	194.20
1024	NIL*	NIL*	NIL*	NIL*

MAPF-ROT (Chan et al. 2022)

Independent learning with postprocessing



(A)

Number of agents	Success rate % on density of 0 with 160-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	100.00	100.00	100.00	100.00
32	100.00	100.00	100.00	100.00
128	100.00	100.00	100.00	100.00
256	20.00	100.00	100.00	100.00
1024	0.00	80.0	100.00	100.00

(B)

Number of agents	Success rate % on density of 0.1 with 80-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	100.00	100.00	100.00	100.00
32	100.00	100.00	100.00	100.00
128	0.00	100.00	100.00	100.00
256	0.00	100.00	100.00	100.00
1024	0.00	0.00	39.92	67.01

(C)

Number of agents	Success rate % on density of 0.2 with 10-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	100.00	100.00	100.00	100.00
32	20.00	0.00	33.75	100.00

(D)

Number of agents	Average timesteps on density of 0 with 160-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	163.80	163.80	165.20	165.20
32	231.60	231.80	329.00	330.00
128	260.60	261.60	309.20	309.40
256	NIL*	272.40	324.20	328.60
1024	NIL*	389.75	334.80	362.20

(E)

Number of agents	Average timesteps on density of 0.1 with 80-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	95.60	95.60	102.80	116.40
32	127.00	129.60	143.20	141.80
128	NIL*	144.60	167.80	158.40
256	NIL*	179.80	216.25	194.20
1024	NIL*	NIL*	NIL*	NIL*

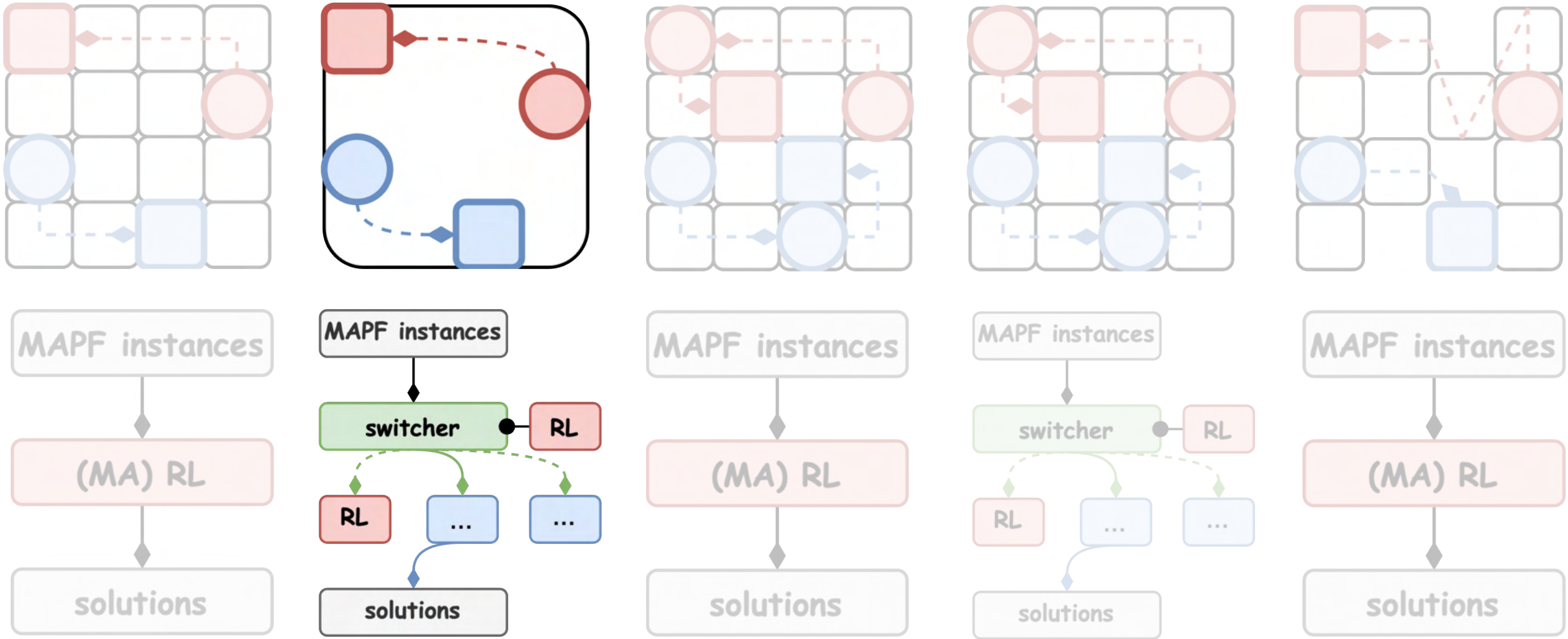
(F)

Number of agents	Average timesteps on density of 0.2 with 10-sized map			
	CBS	ODrM*	MAPF-rot	Full MAPF-rot^
4	14.80	15.80	17.80	18.60
32	NIL*	NIL*	NIL*	157.20

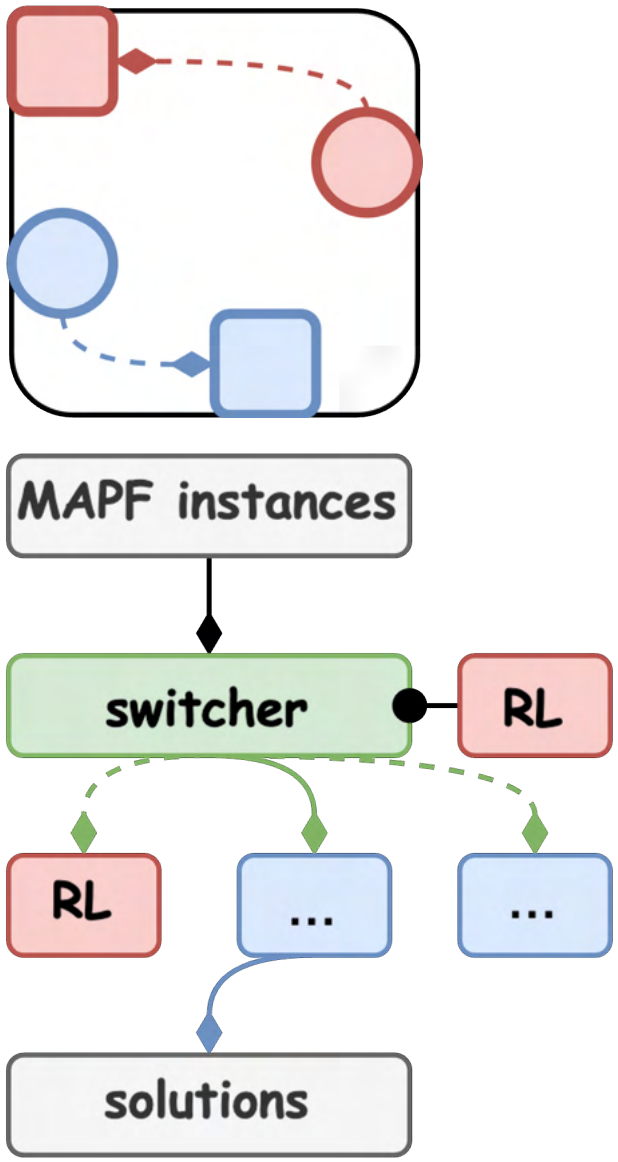
* The results are not shown when more than half of the trials are not completed
^ Full MAPF-rot means the MAPF-rot with the proposed deadlock scheme

MAPF-ROT (Chan et al. 2022)

Reinforcement Learning as the Continuous MAPF Candidate



Reinforcement Learning as the Continuous MAPF Candidate



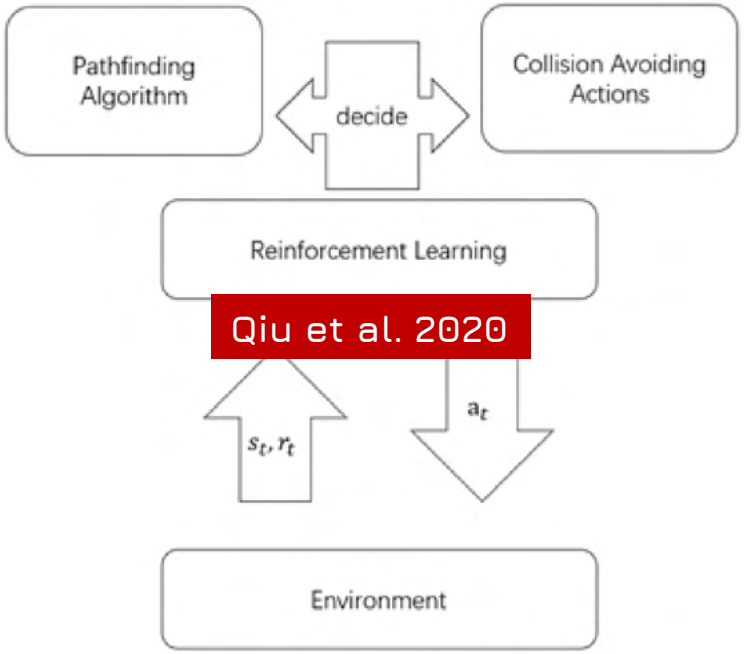
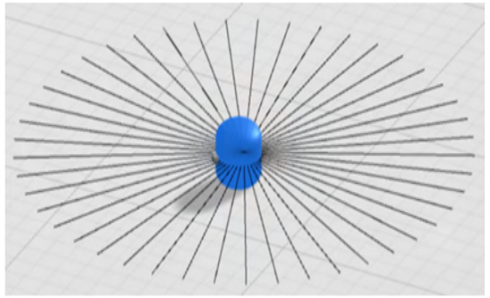


Table 2: List of Actions

Notation	Definition
a_0	take action accroding to method $\mathcal{M}$
a_1	stay at the current point
a_2	move forward: $v = v_0$
a_3	turn left: $v = 0,\ \Delta\omega = -\omega_0$
a_4	turn right: $v = 0,\ \Delta\omega = \omega_0$
a_5	move backward: $v = v_0$

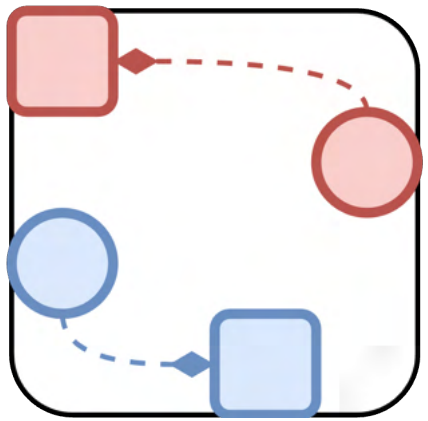
Expert-guided reward

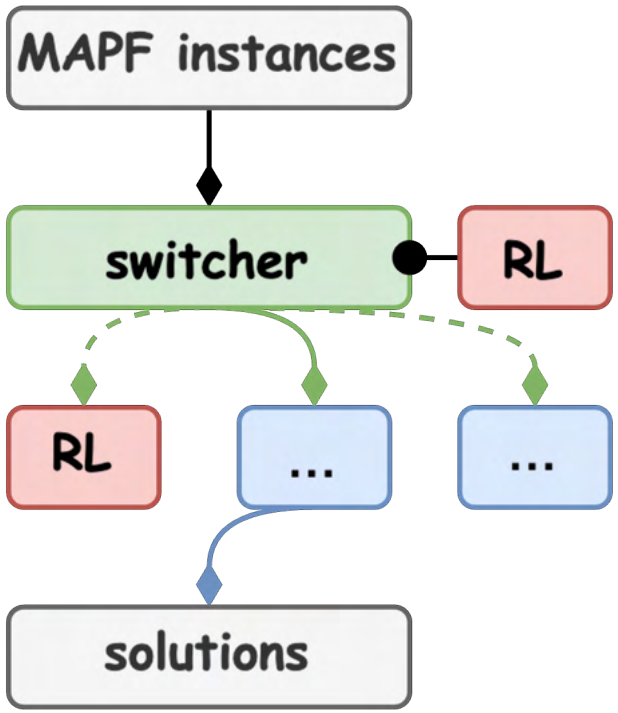


Static, local information

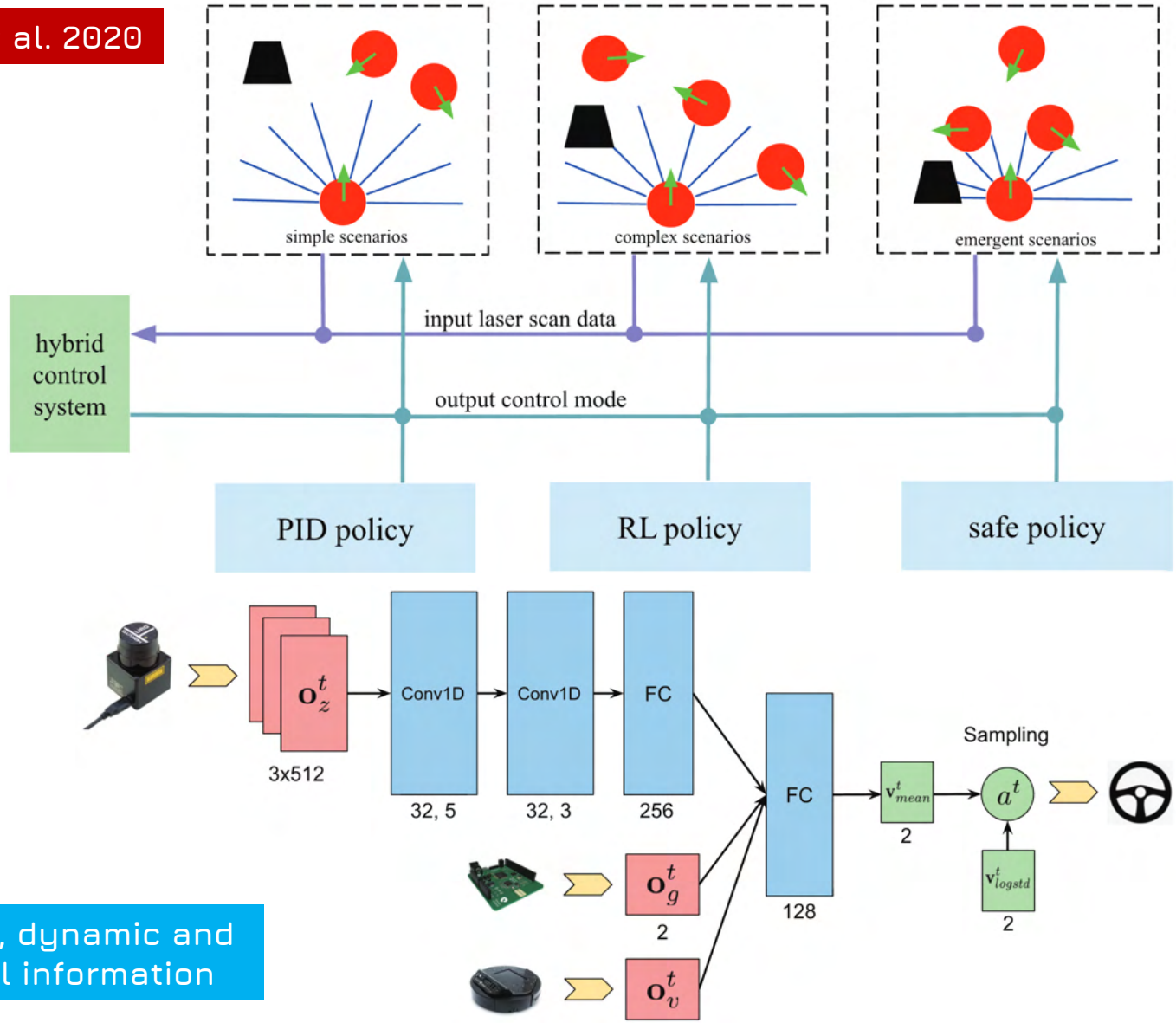
Scenario	Method	Success Rate	EDP	Collision Rate
FW	Pure RL Method	0.329	2.313	0.360
	Our Method(4 FC layers)	**0.969**	**1.177**	**0.030**
RO N=80	Pure RL Method	0.623	1.951	0.268
	Our Method(4 FC layers)	**0.999**	**0.833**	**<0.001**
RO N=120	Pure RL Method	0.432	2.003	0.377
	Our Method(4 FC layers)	**0.960**	**1.051**	**0.039**
CT	Pure RL Method	0.999	1.310	<0.001
	Our Method(4 FC layers)	**0.999**	**0.893**	**<0.001**

Reinforcement Learning as the Continuous MAPF Candidate



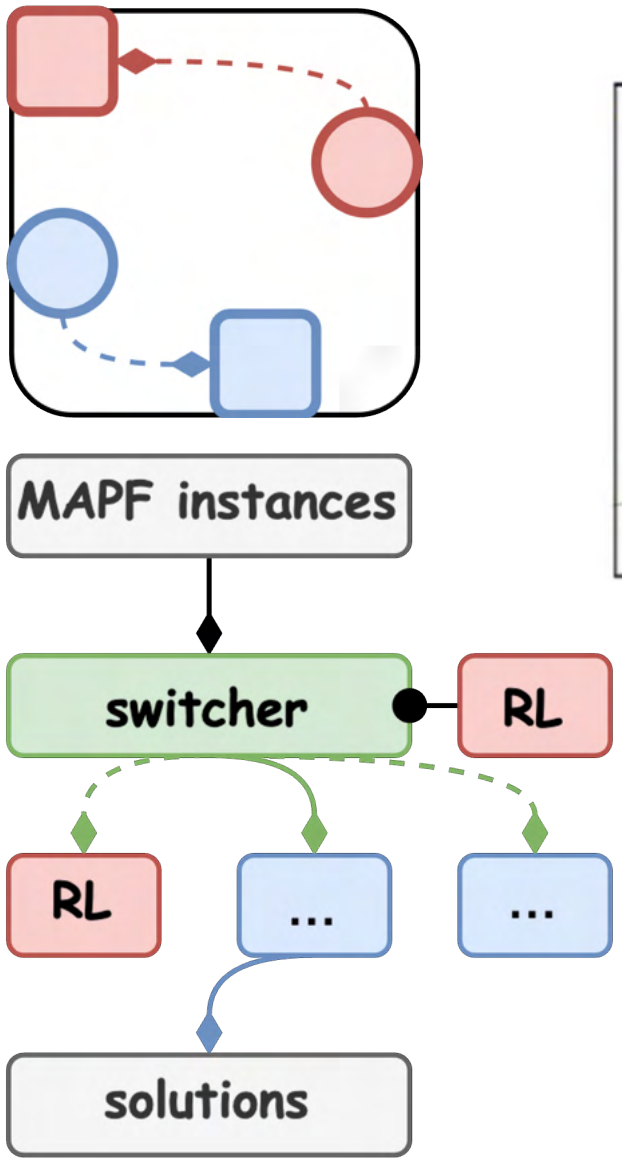


Fan et al. 2020

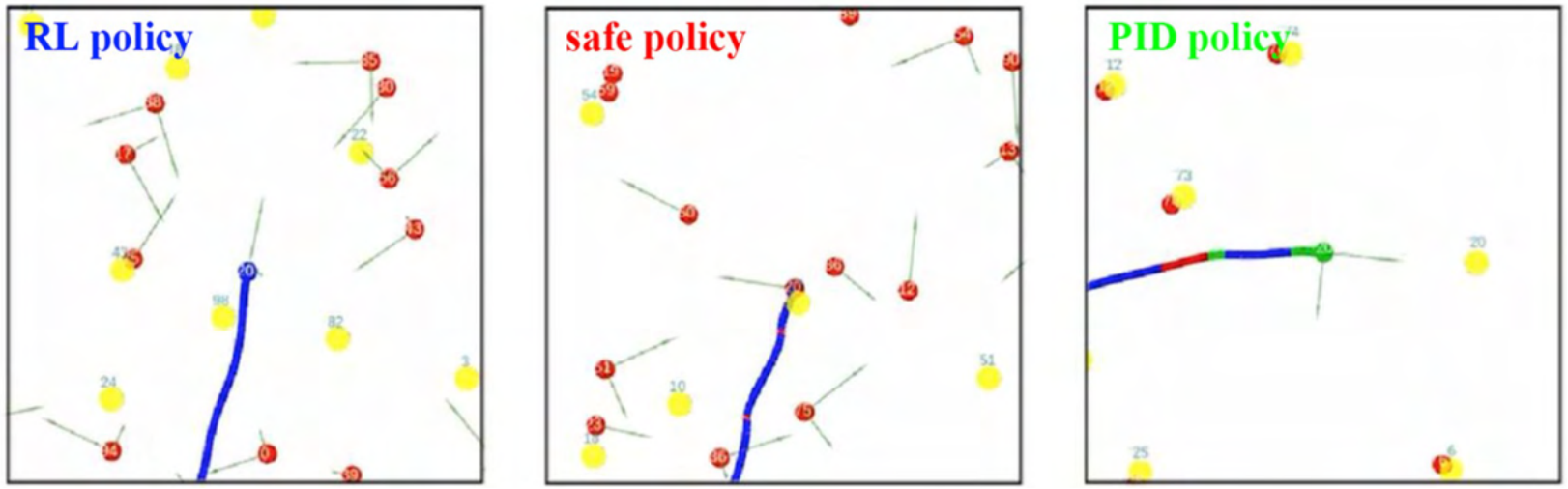


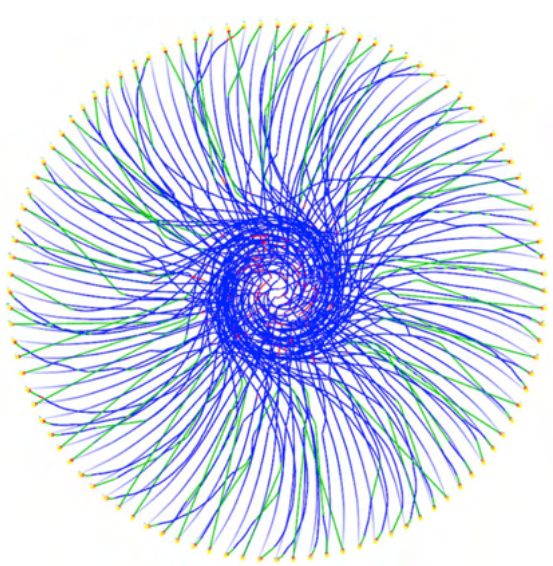
Static, dynamic and local information

Reinforcement Learning as the Continuous MAPF Candidate

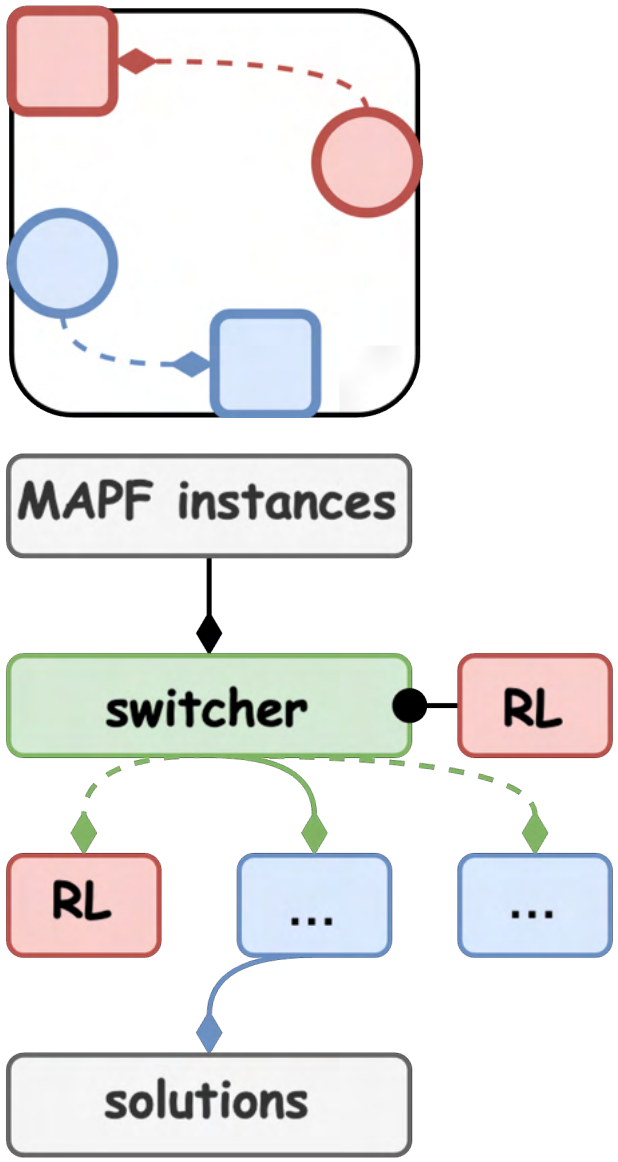


Fan et al. 2020





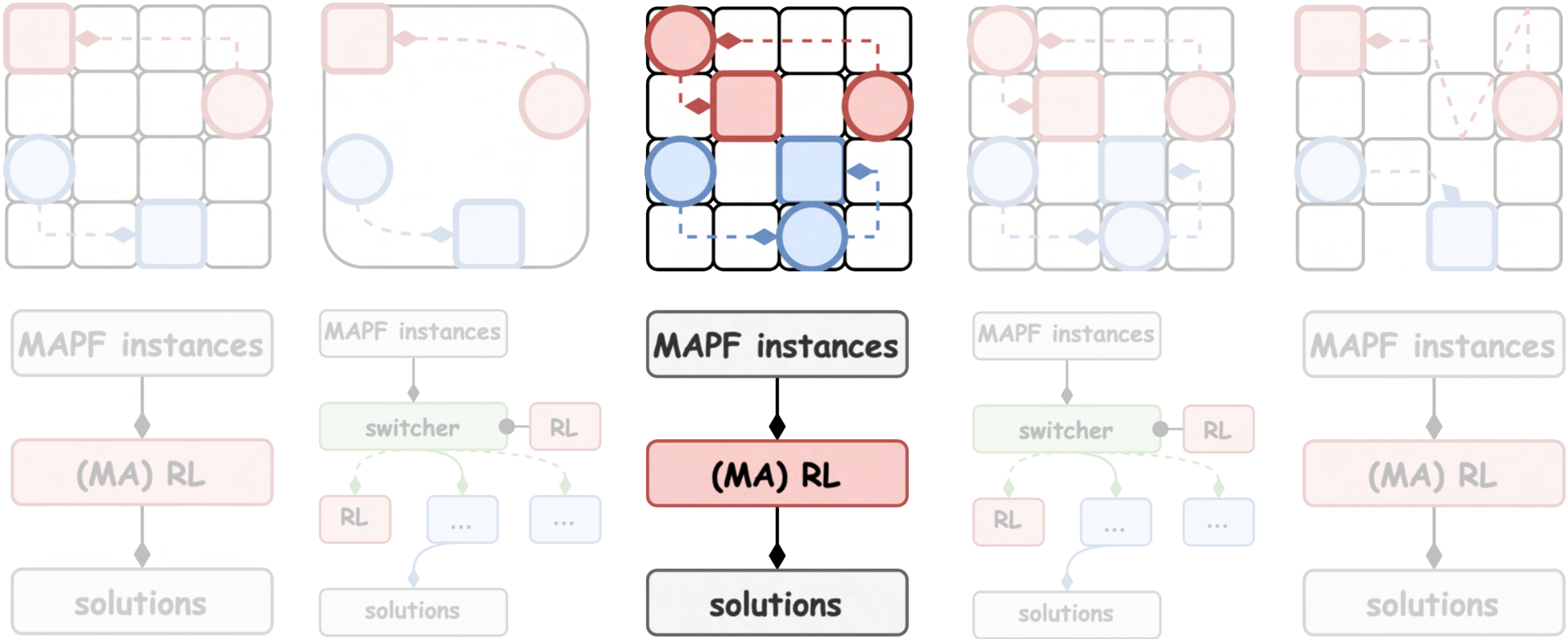
Reinforcement Learning as the Continuous MAPF Candidate



Fan et al. 2020

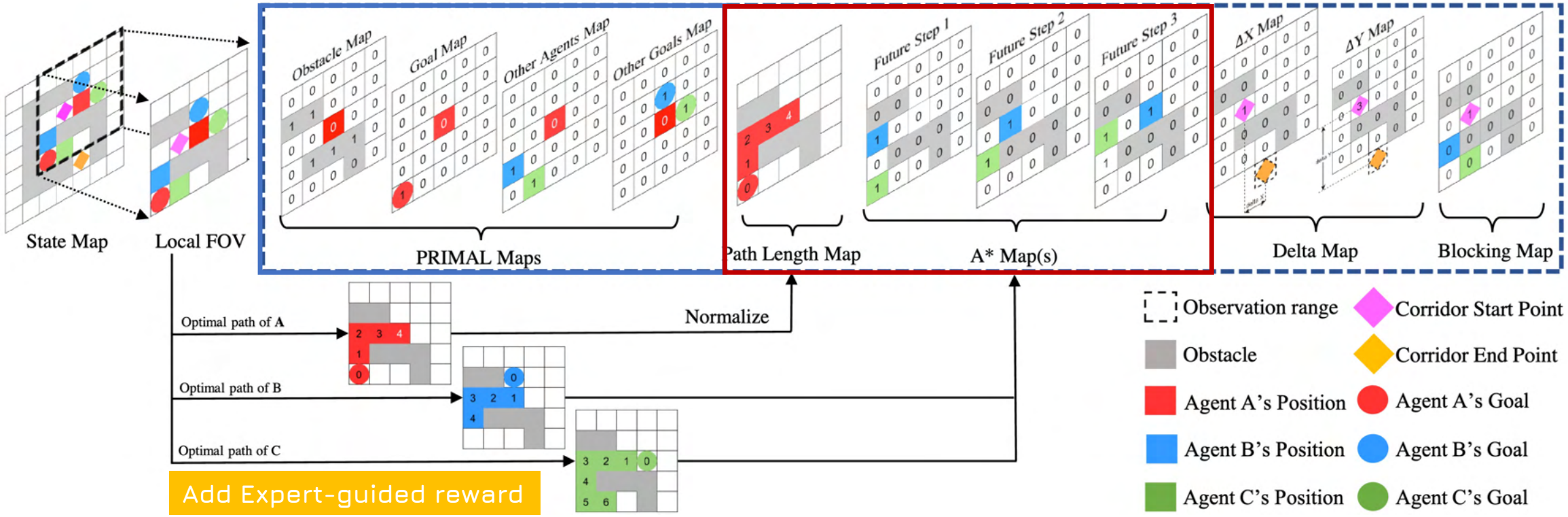
Metrics	Method	#agents (radius of the scene)						
		4 (2.5 m)	6 (3.0 m)	8 (3.5 m)	10 (4.0 m)	12 (4.5 m)	15 (5.0 m)	20 (6.0 m)
Success rate	SL	0.6000	0.7167	0.6125	0.71	0.6333	—	—
	NH-ORCA-A	**1.0**	0.9667	0.9250	0.8900	0.9000	0.8067	0.7800
	NH-ORCA-N	0.9900	0.9933	0.9975	0.9860	0.9767	0.992	0.988
	RL (Ape-X)	0.9050	0.9567	0.9975	0.9920	0.9517	0.9613	0.9430
	RL (PPO)	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9827
	Hybrid-RL	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Extra time	SL	9.254/2.592	9.566/3.559	12.085/2.007	13.588/1.206	19.157/2.657	—	—
	NH-ORCA-A	0.622/0.080	0.773/0.207	1.067/0.215	0.877/0.434	0.771/0.606	1.750/0.654	1.800/0.647
	NH-ORCA-N	2.276/1.138	3.746/2.251	5.862/2.013	6.298/1.531	7.774/2.335	9.110/2.281	13.152/3.810
	RL (Ape-X)	0.316/0.149	1.120/0.443	0.529/0.166	**0.567/0.134**	1.458/0.386	1.681/0.417	3.11/0.526
	RL (PPO)	0.323/0.025	**0.408/0.009**	0.510/0.005	0.631/0.011	0.619/0.020	0.490/0.046	0.778/0.032
	Hybrid-RL	**0.251/0.007**	**0.408/0.008**	**0.494/0.006**	**0.629/0.009**	**0.518/0.005**	**0.332/0.007**	**0.702/0.013**
Extra distance	SL	0.358/0.205	0.181/0.146	0.138/0.079	0.127/0.047	0.141/0.027	—	—
	NH-ORCA-A	0.017/0.004	**0.025/0.005**	0.041/0.034	0.034/0.009	0.062/0.024	0.049/0.019	**0.056/0.018**
	NH-ORCA-N	0.380/0.022	0.452/0.055	0.494/0.049	0.518/0.041	0.542/0.036	0.571/0.036	0.608/0.042
	RL (Ape-X)	**0.004/0.002**	0.056/0.015	**0.022/0.007**	**0.029/0.007**	0.041/0.009	0.036/0.006	0.066/0.006
	RL (PPO)	0.028/0.006	0.028/0.001	0.033/0.001	0.036/0.001	**0.038/0.002**	0.049/0.005	0.065/0.002
	Hybrid-RL	0.013/0.001	0.028/0.001	0.031/0.001	0.036/0.001	0.039/0.001	**0.033/0.001**	0.058/0.001
Average speed	SL	0.326/0.072	0.381/0.087	0.354/0.042	0.355/0.022	0.308/0.028	—	—
	NH-ORCA-A	0.859/0.012	0.867/0.026	0.839/0.032	0.876/0.045	0.875/0.054	0.820/0.052	0.831/0.042
	NH-ORCA-N	0.702/0.095	0.645/0.132	0.558/0.092	0.566/0.065	0.547/0.079	0.530/0.058	0.487/0.070
	RL (Ape-X)	0.945/0.061	0.846/0.053	0.930/0.021	0.934/0.015	0.862/0.033	0.857/0.030	0.795/0.028
	RL (PPO)	0.939/0.004	**0.936/0.001**	0.932/0.001	0.927/0.001	0.936/0.002	0.953/0.004	0.939/0.002
	Hybrid-RL	**0.952/0.001**	**0.936/0.001**	**0.934/0.001**	**0.927/0.001**	**0.946/0.001**	**0.968/0.001**	**0.945/0.001**

Reinforcement Learning as the Lifelong MAPF Solver



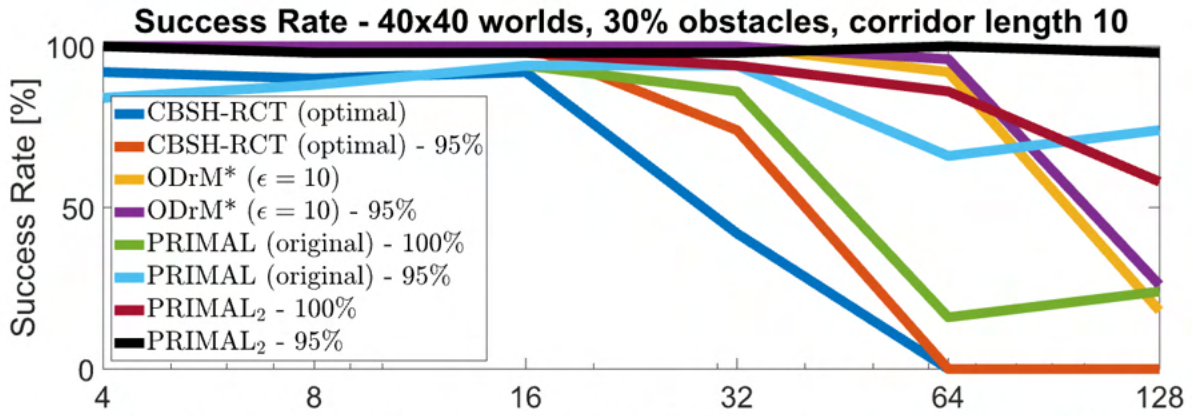
Reinforcement Learning as the Lifelong MAPF Solver

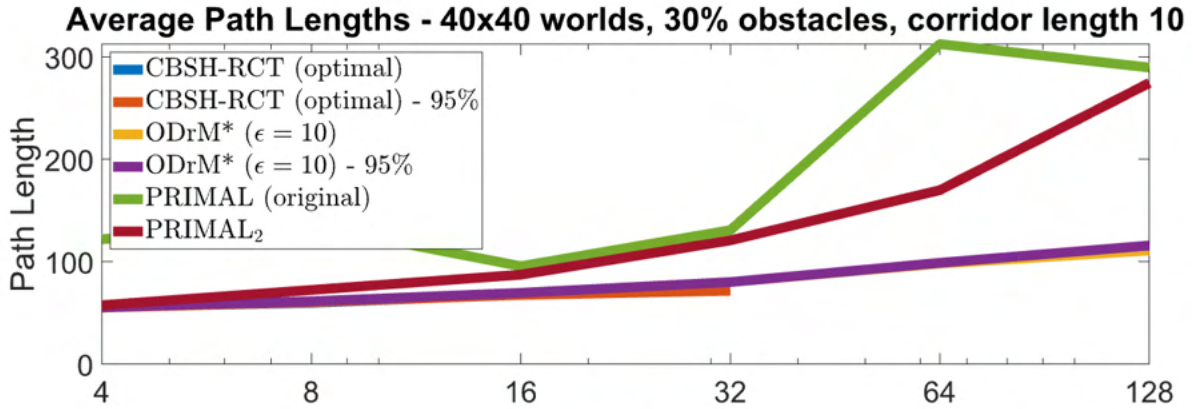
Add expert-guided information

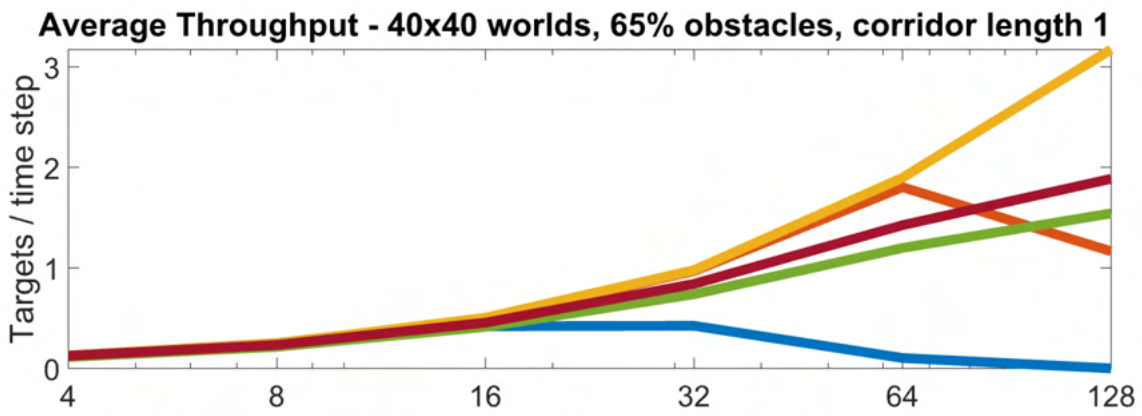


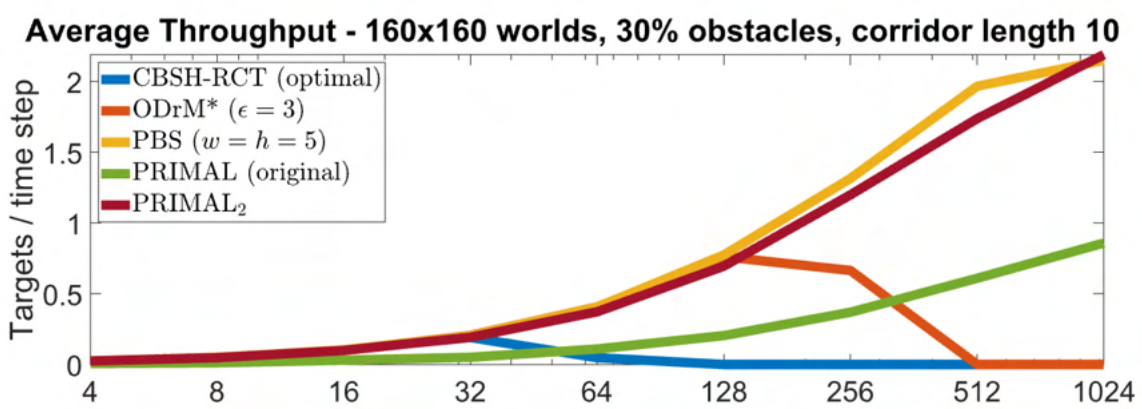
PRIMAL2 (Damani et al. 2021)

Reinforcement Learning as the Lifelong MAPF Solver



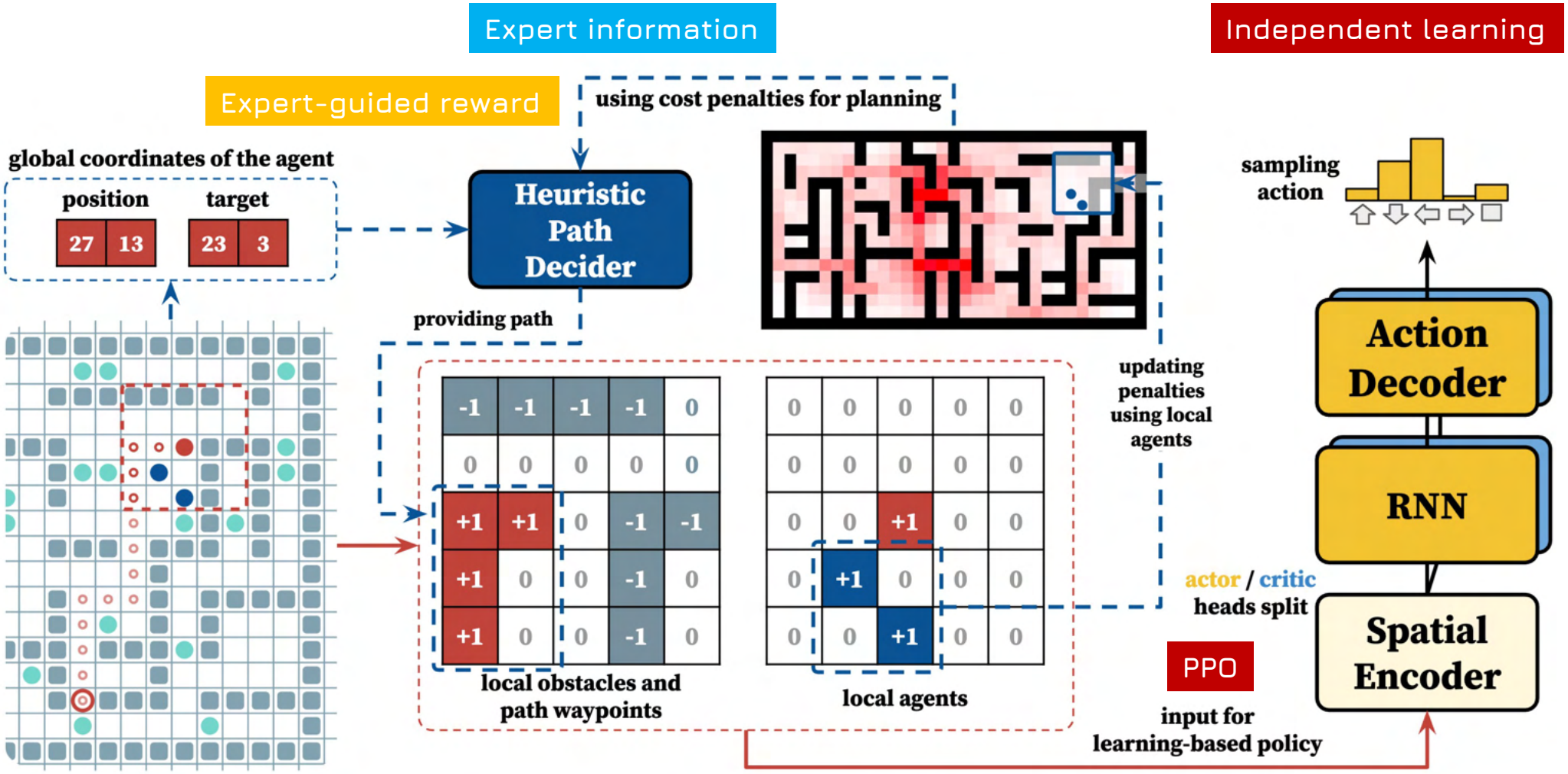






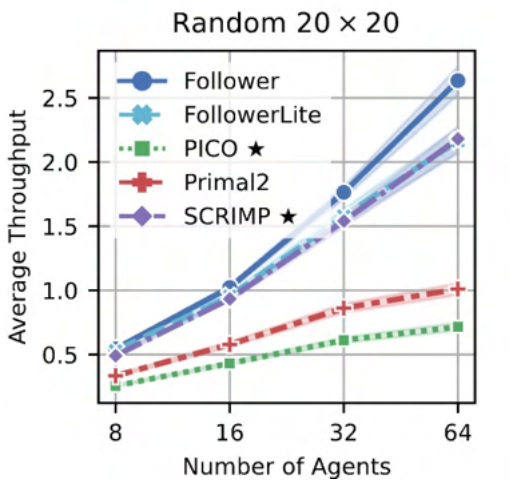
PRIMAL2 (Damani et al. 2021)

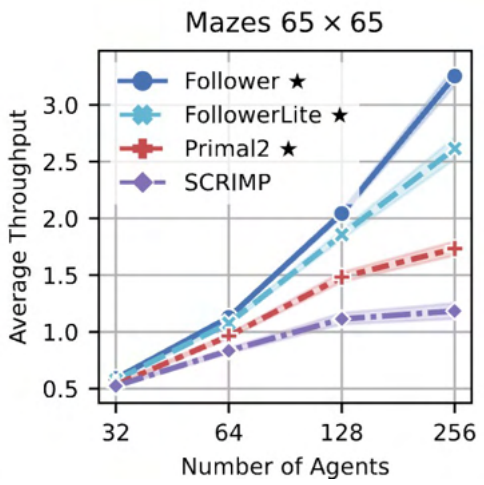
Reinforcement Learning as the Lifelong MAPF Solver

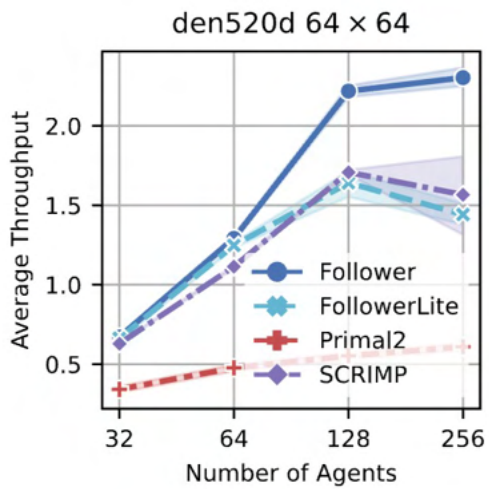


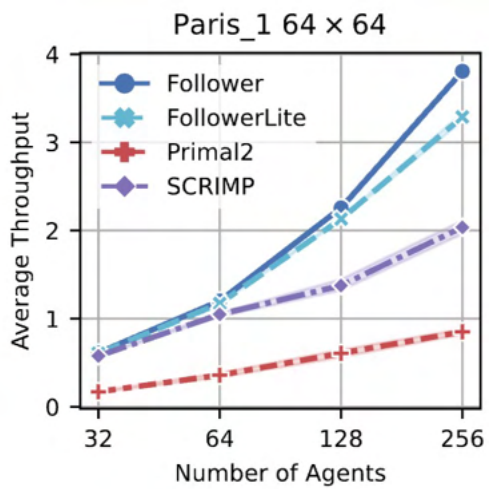
Follower (Skrynnik et al. 2023)

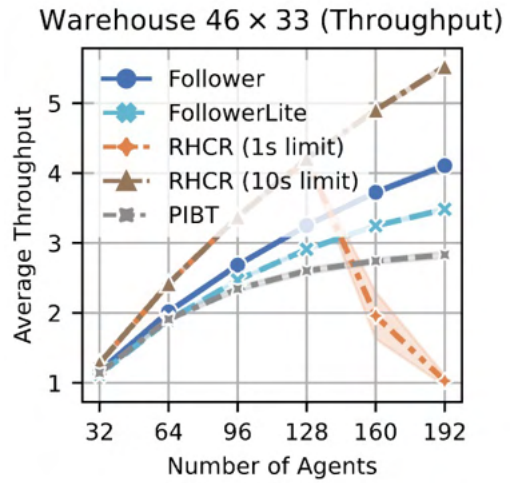
Reinforcement Learning as the Lifelong MAPF Solver

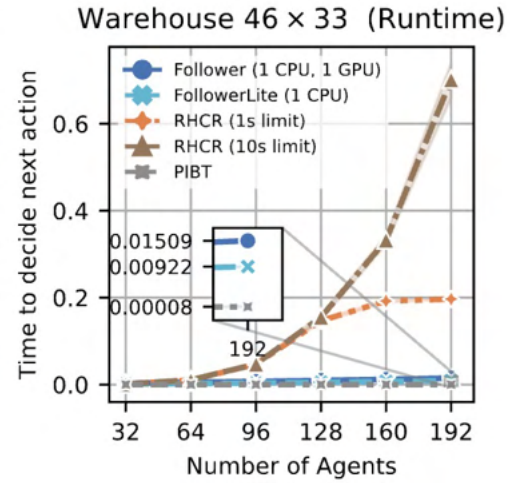


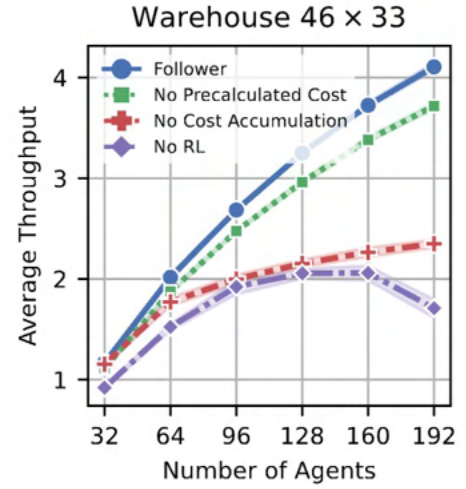


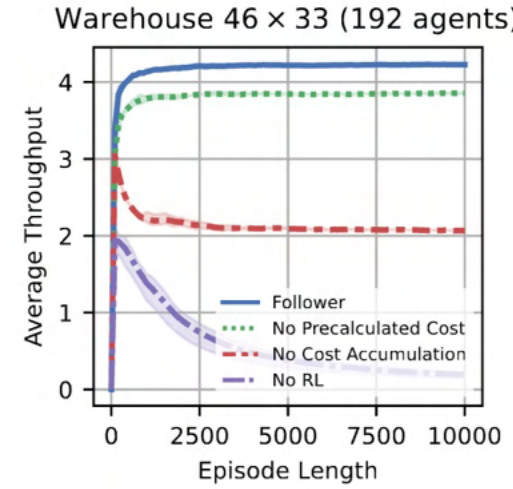






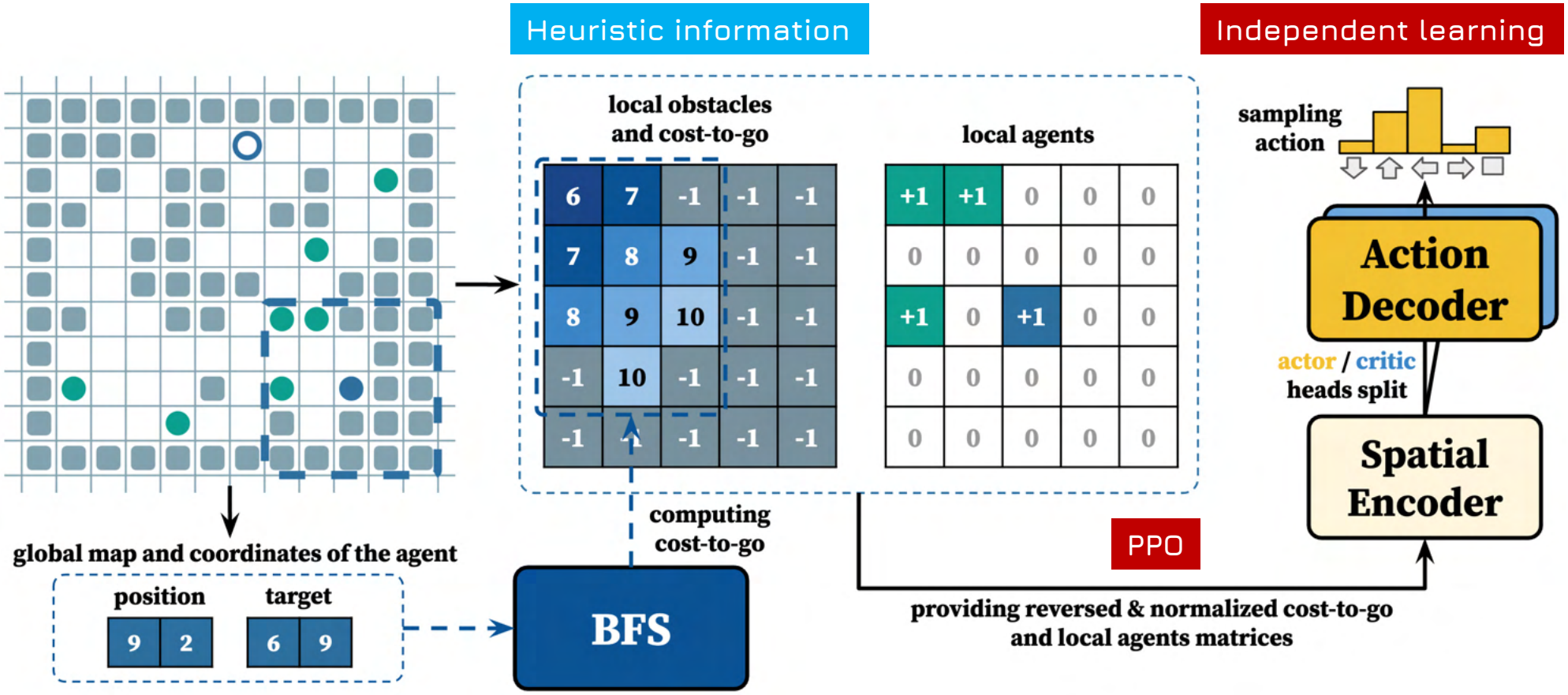






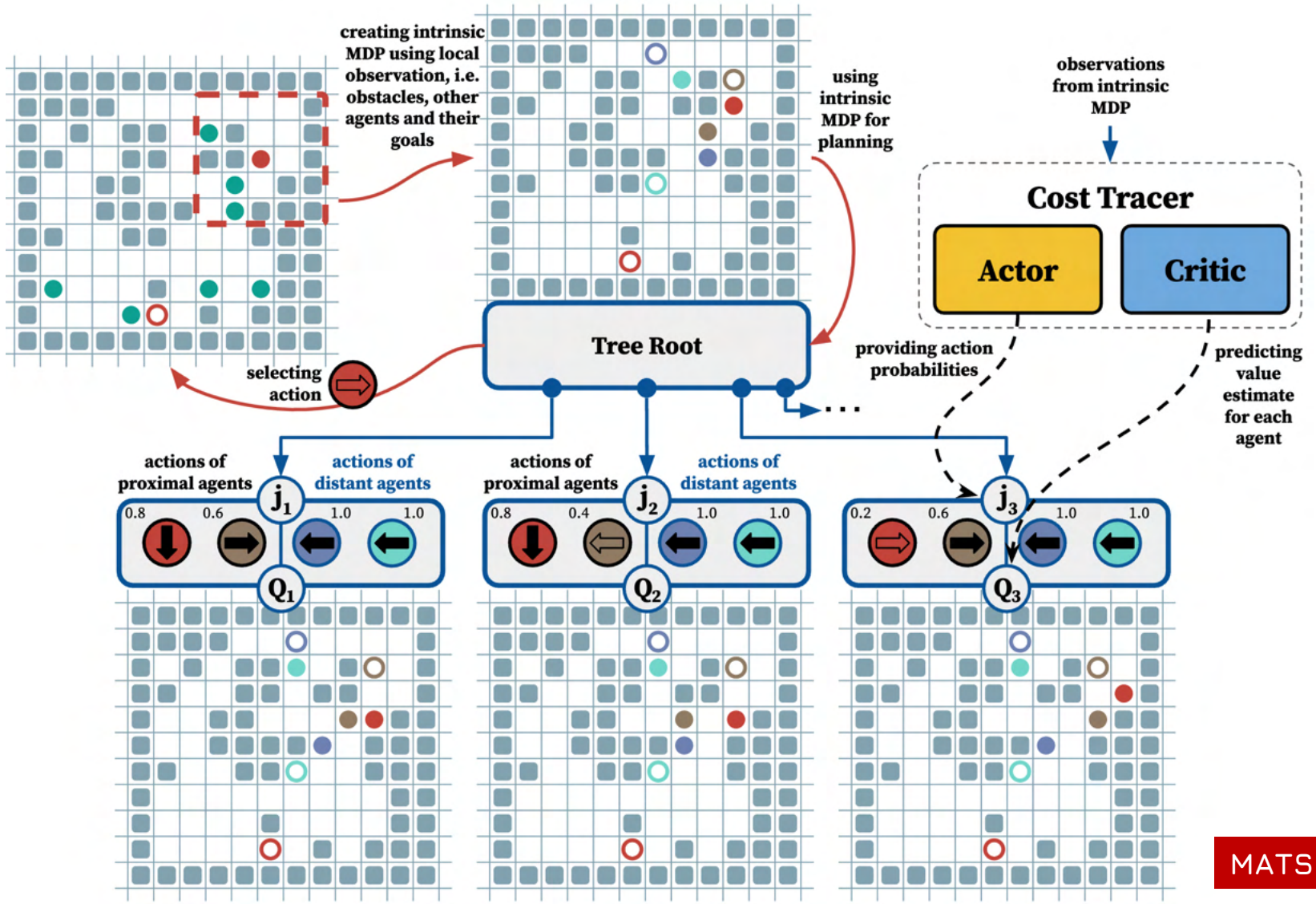
Follower (Skrynnik et al. 2023)

Reinforcement Learning as the Lifelong MAPF Solver



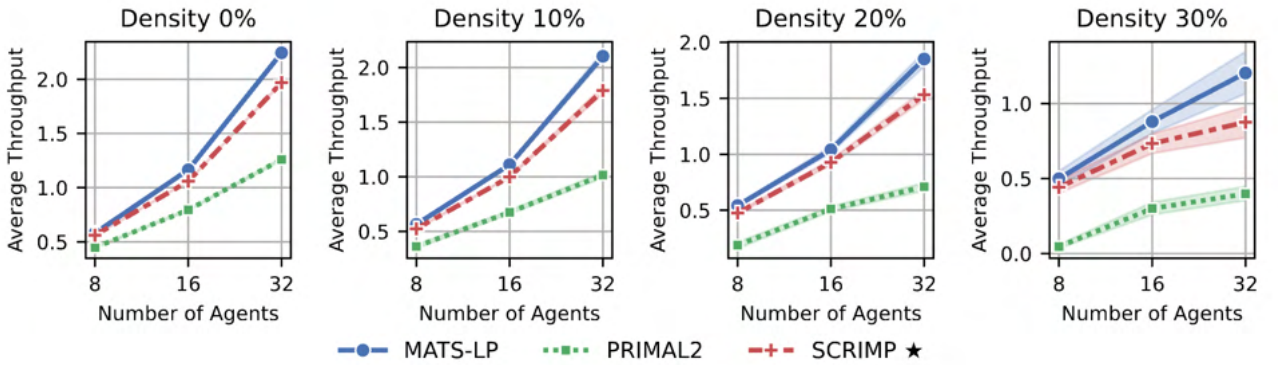
MATS-LP (Skrynnik et al. 2023)

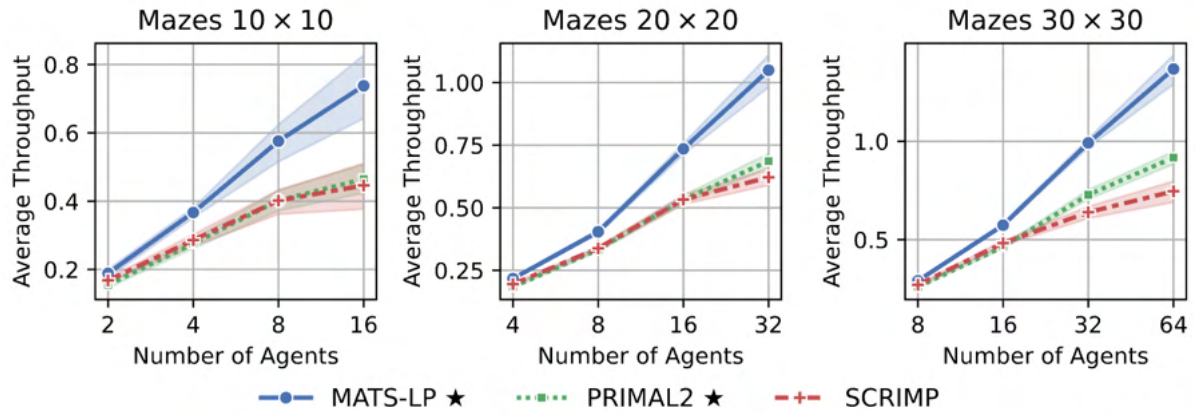
Reinforcement Learning as the Lifelong MAPF Solver

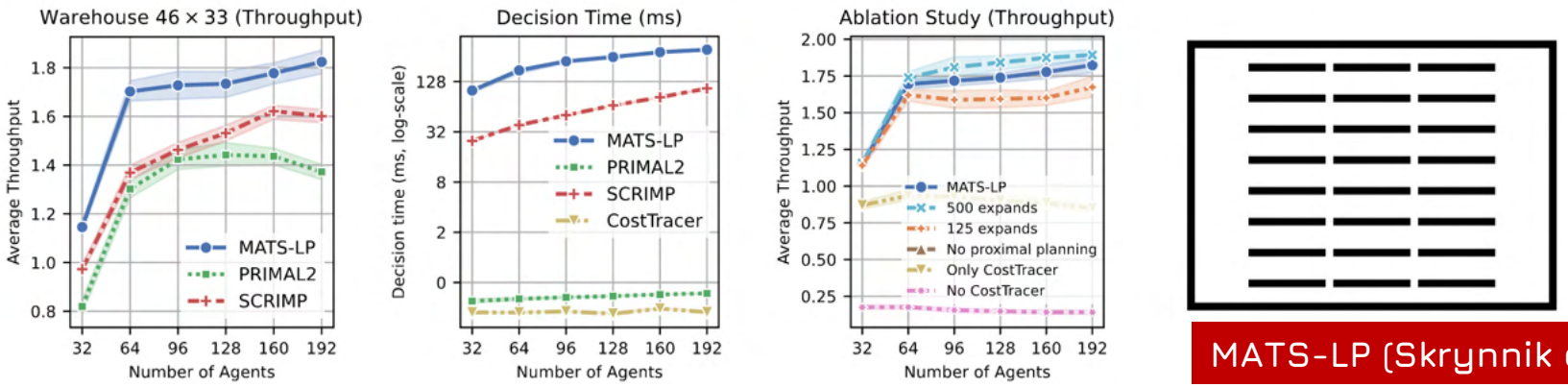


MATS-LP (Skrynnik et al. 2023)

Reinforcement Learning as the Lifelong MAPF Solver







MATS-LP (Skrynnik et al. 2023)

Reinforcement Learning as the Lifelong MAPF Solver

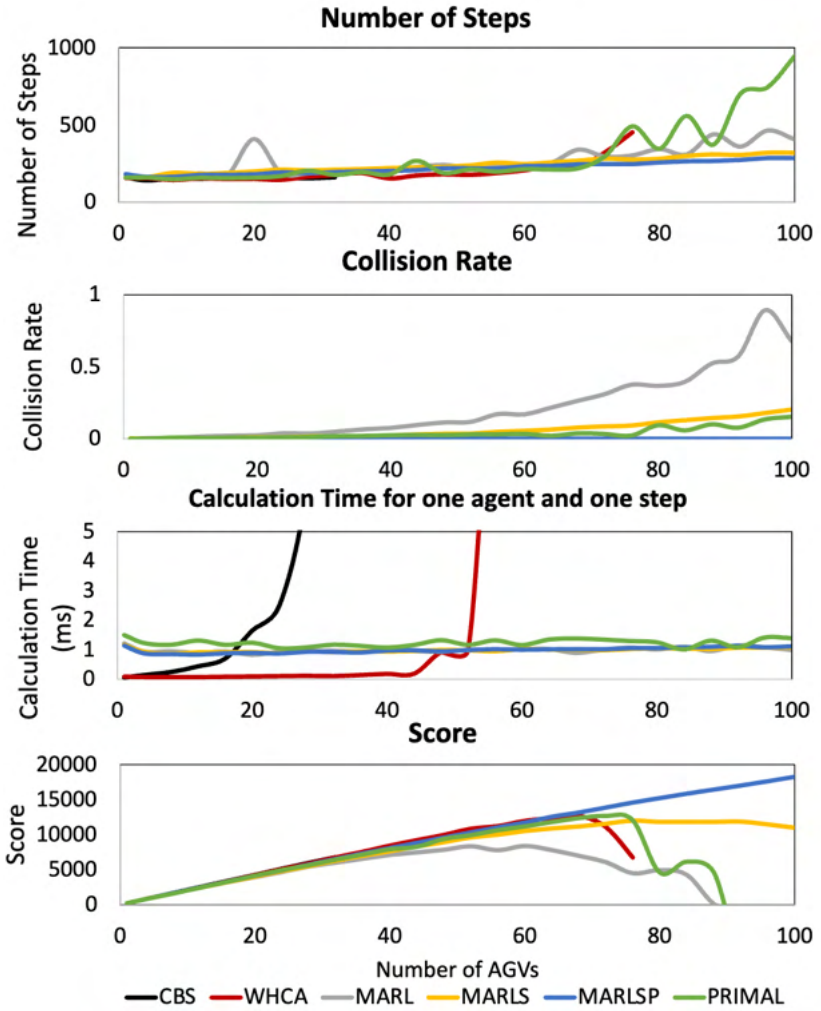
Algorithm 1: Multi-agent Training with Searching

for $m = 1$ to $N_{episode}$ **do**
  Reset environment and get initial state
  Stage 1: Sampling
  while $t < T$ **do**
    for k=1 to K **do**
      Calculate probability weights $\pi(a_t^k|s_t^k)$
      Create root node v_0 with $Q_0 = 0, N_0 = 0$
      while $n < N$ **do**
        if any node $v_l \in V$ is not fully-expanded **then**
          Expand $v_0 \rightarrow v_l$ by choosing an action sequen $\{a_t, a_{t+1}, \ldots a_{t+\tau}\}$ according to policy π
          Estimate $s_{t+\tau}$ and $r_{t+\tau}$ by simulating actions
          Add new child v_l to V with
        else
          Sample a node v_l with the maximum UCB
          Backward propagation of parent and ancestor nodes by $Q_l = Q_l + q_l, N_l = N_l + 1$
        end if
      end while
      Choose v_0 with the largest Q_l/N_l and its action a_t^k
      Execute $\mathbf{a}_t^k$ and observe reward $\mathbf{r}_t^k$, next state $\mathbf{s}_{t+1}^k$
    end for
    Store the transitions $(s_t^k, a_t^k, r_t^k, s_{t+1}^k)$ into M
  end while
  Stage 2: Learning
  Sample a batch of experience: s_t^k, $V_{target}(s_{t+1}^k; \pi)$
  Update by minimizing the value loss Eq.(1) over the batc
  Update as $\theta \leftarrow \theta + \nabla_\theta L_\theta$ according to Eq.(3)
end for

Independent learning | MCTS

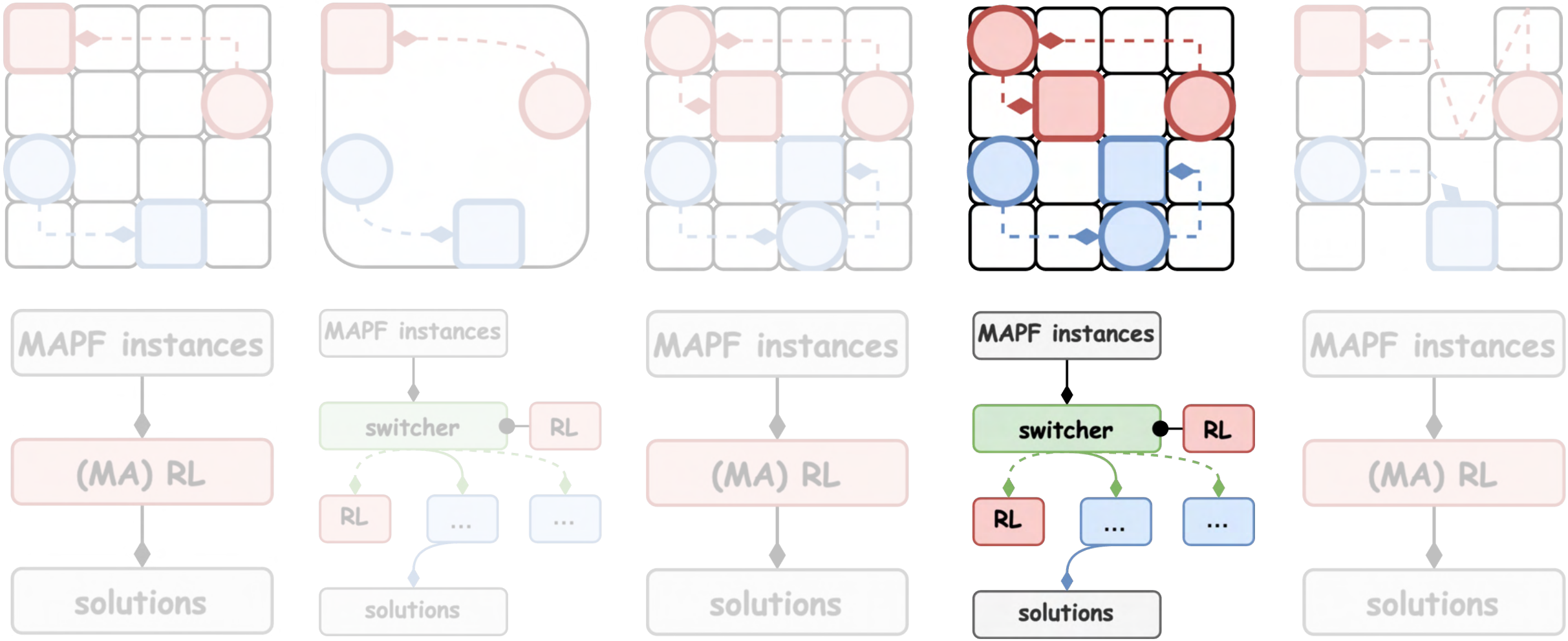
post-processing:

1) **Return** if the given action is **conflict-free**;
2) **Sort** the other four actions in **decreasing** order according to the **probability weights from the policy**;
3) **Choose the first action** and go to step 1.



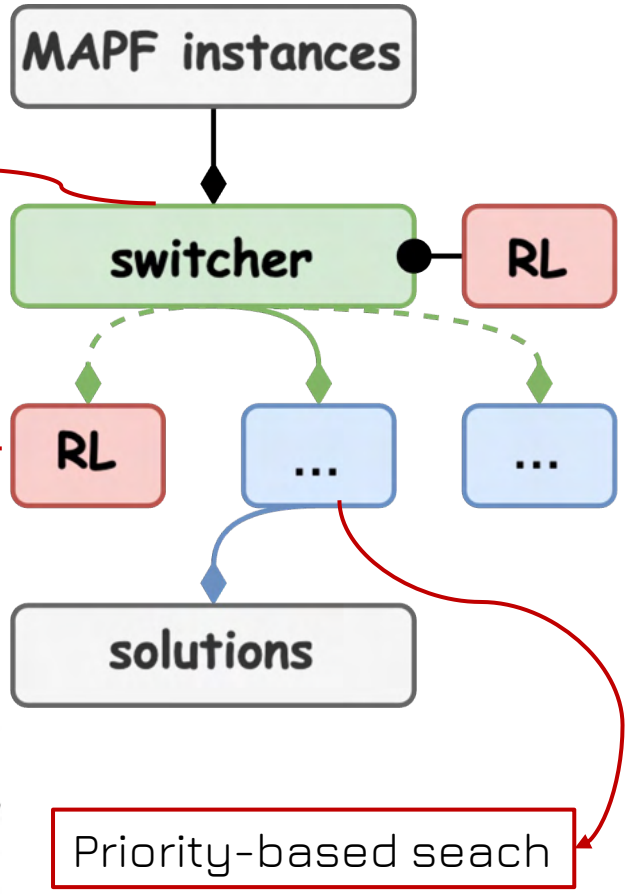
Zhang et al. 2020

Reinforcement Learning as the Lifelong MAPF Candidate

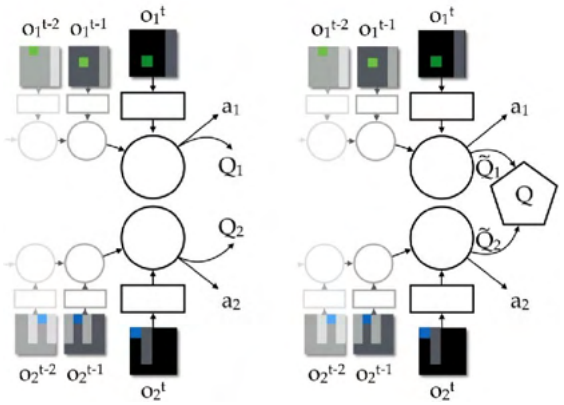


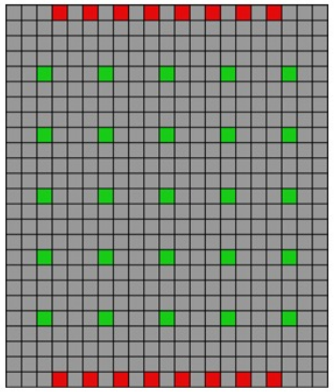
Reinforcement Learning as the Lifelong MAPF Candidate

Pre-train a collision detector, which takes local observations of two agents as input and predicts whether they will collide in the future

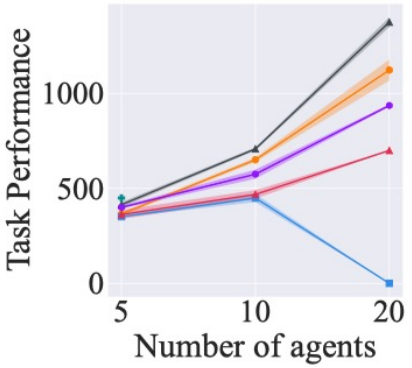


VDN

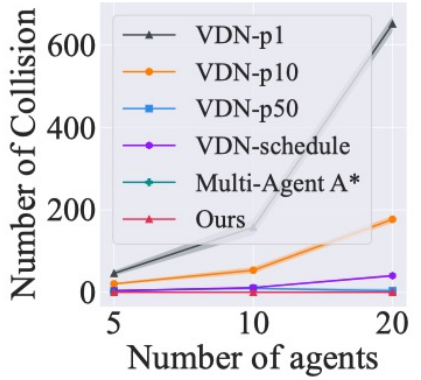




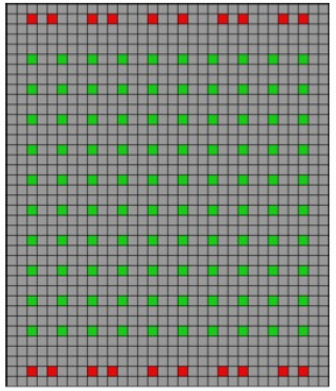
(a) map_one



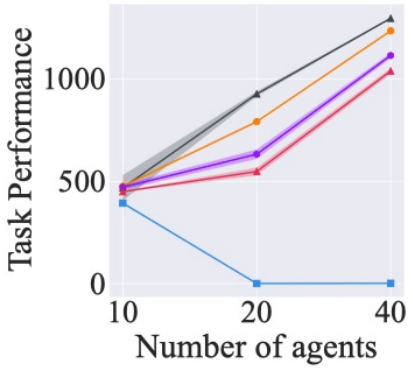
(b) Performance-one



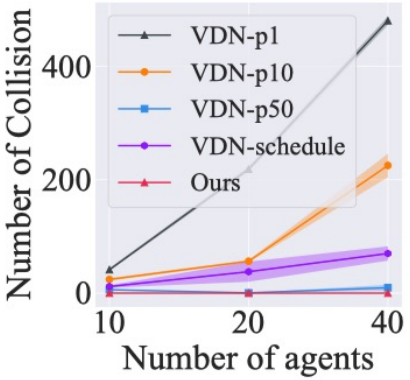
(c) Collision-one



(d) map_two



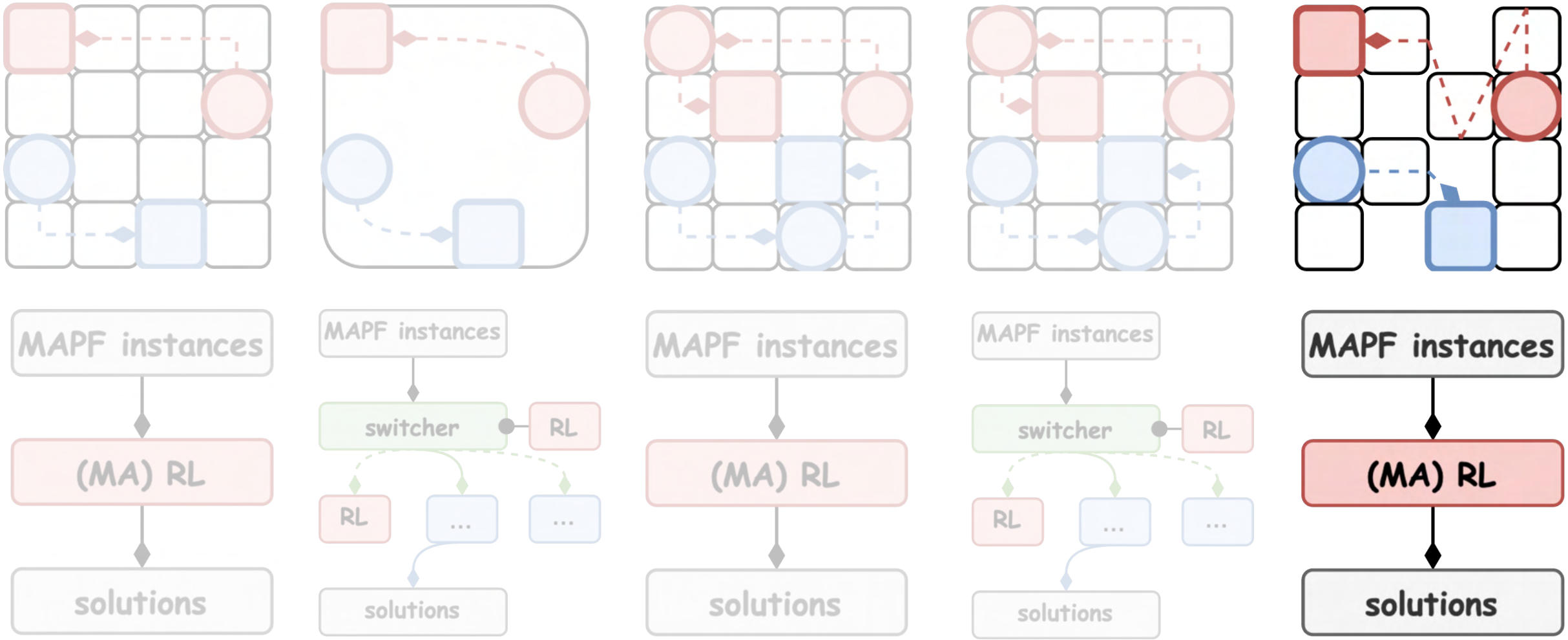
(e) Performance-two



(f) Collision-two

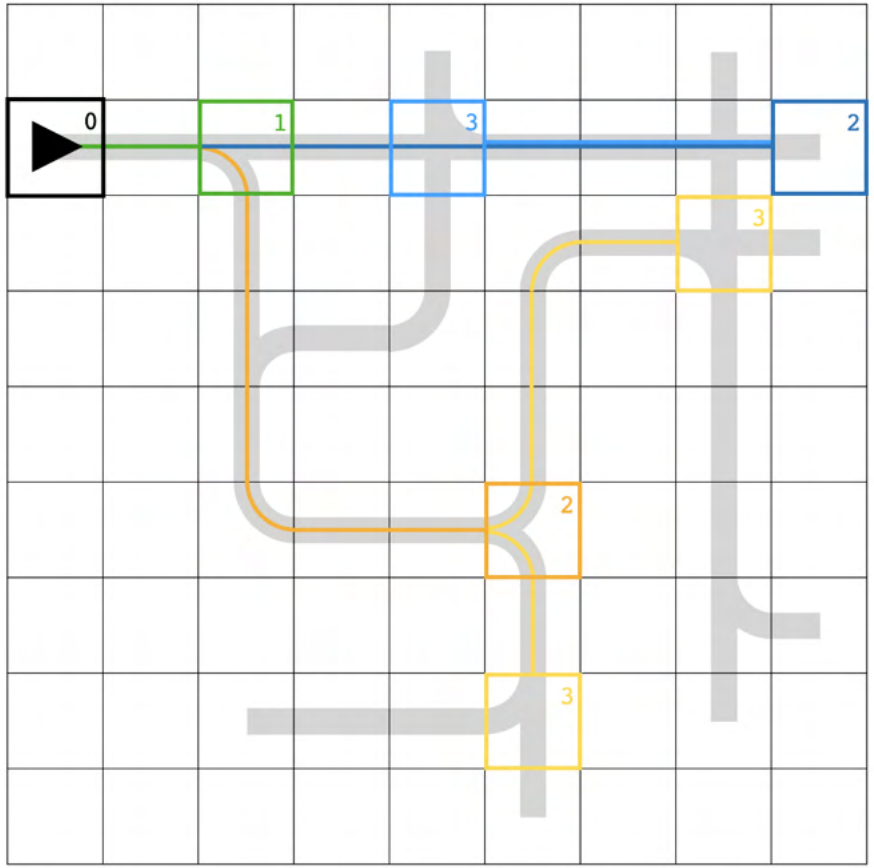
Chen et al. 2023c

Reinforcement Learning as the Graphic MAPF Solver



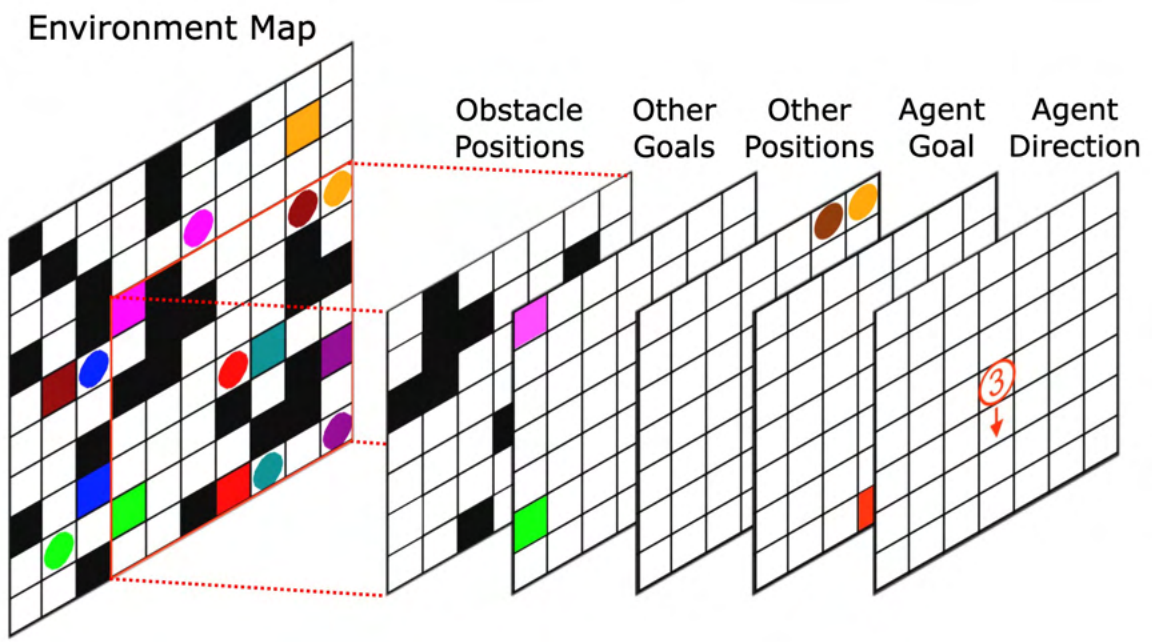
Reinforcement Learning as the Graphic MAPF Solver

Static, local information



Flatland-RL (Mohanty et al. 2020)

Global, local information

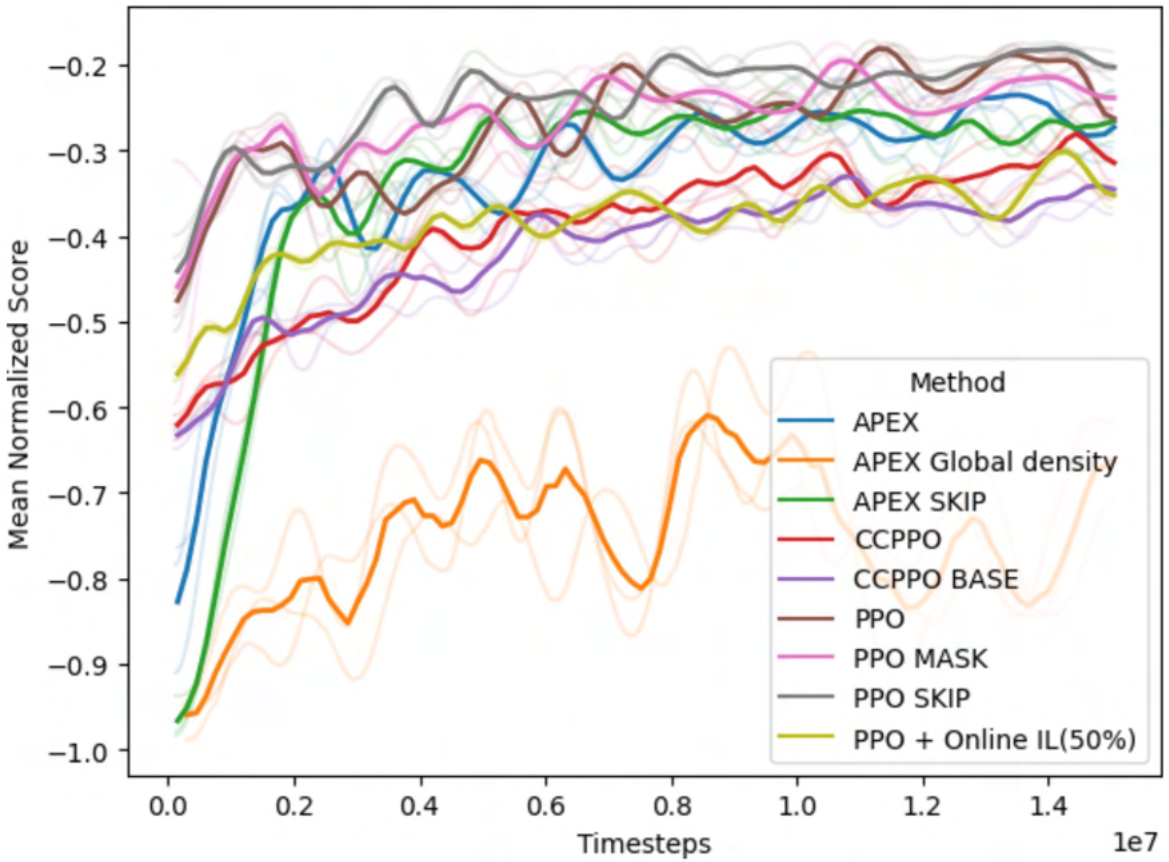


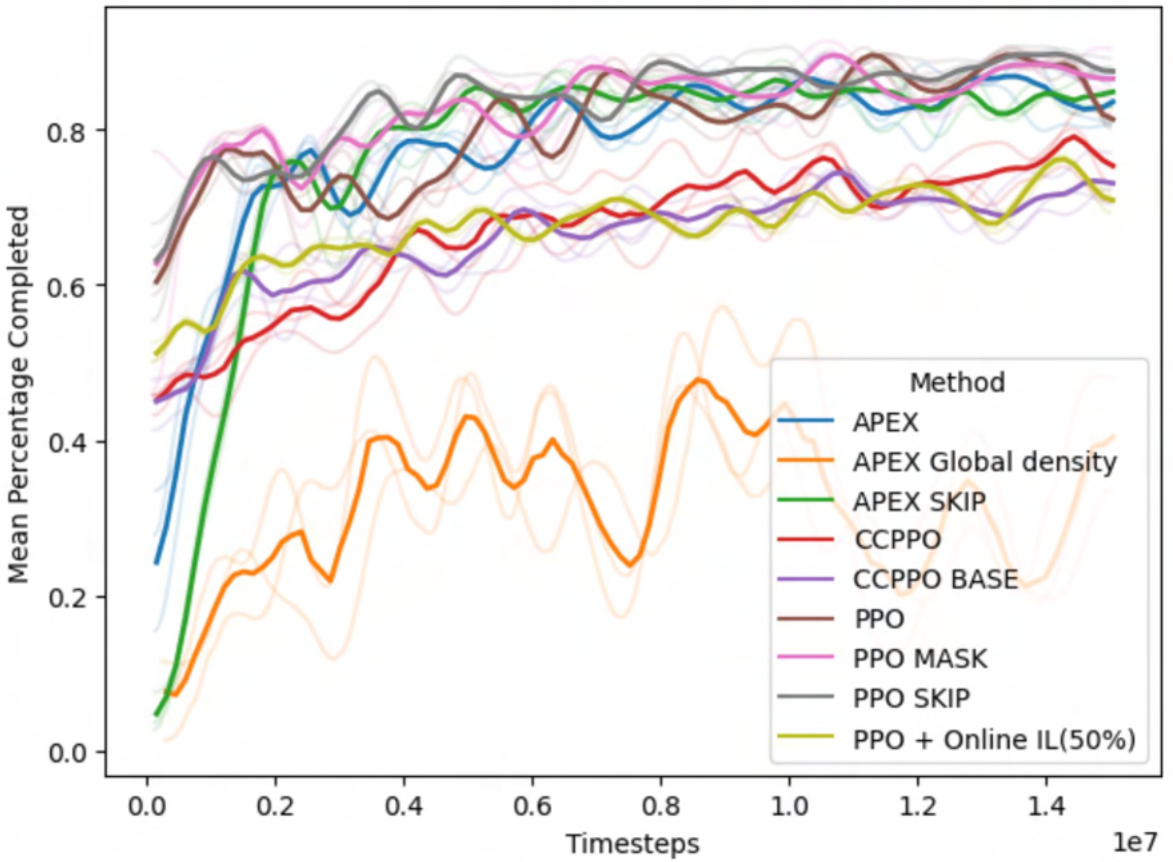
Search-based local information construction

Reinforcement Learning as the Graphic MAPF Solver

Independent learning | cooperation | imitation learning

5 × 25 cells and 5 agents



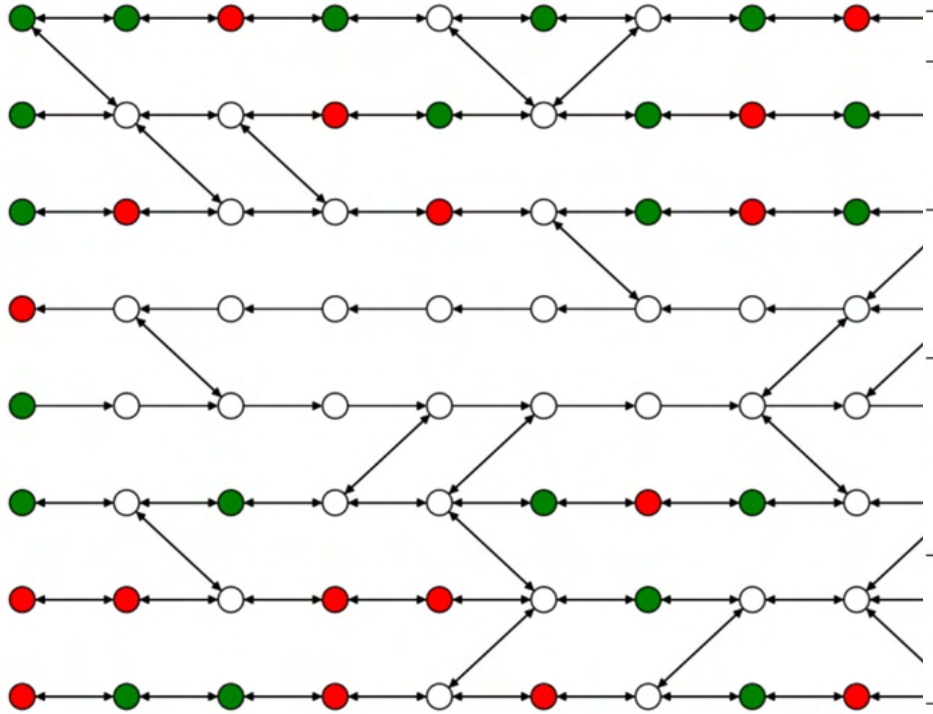


Flatland-RL (Mohanty et al. 2020)

Reinforcement Learning as the Graphic MAPF Solver

Independent learning

DQN



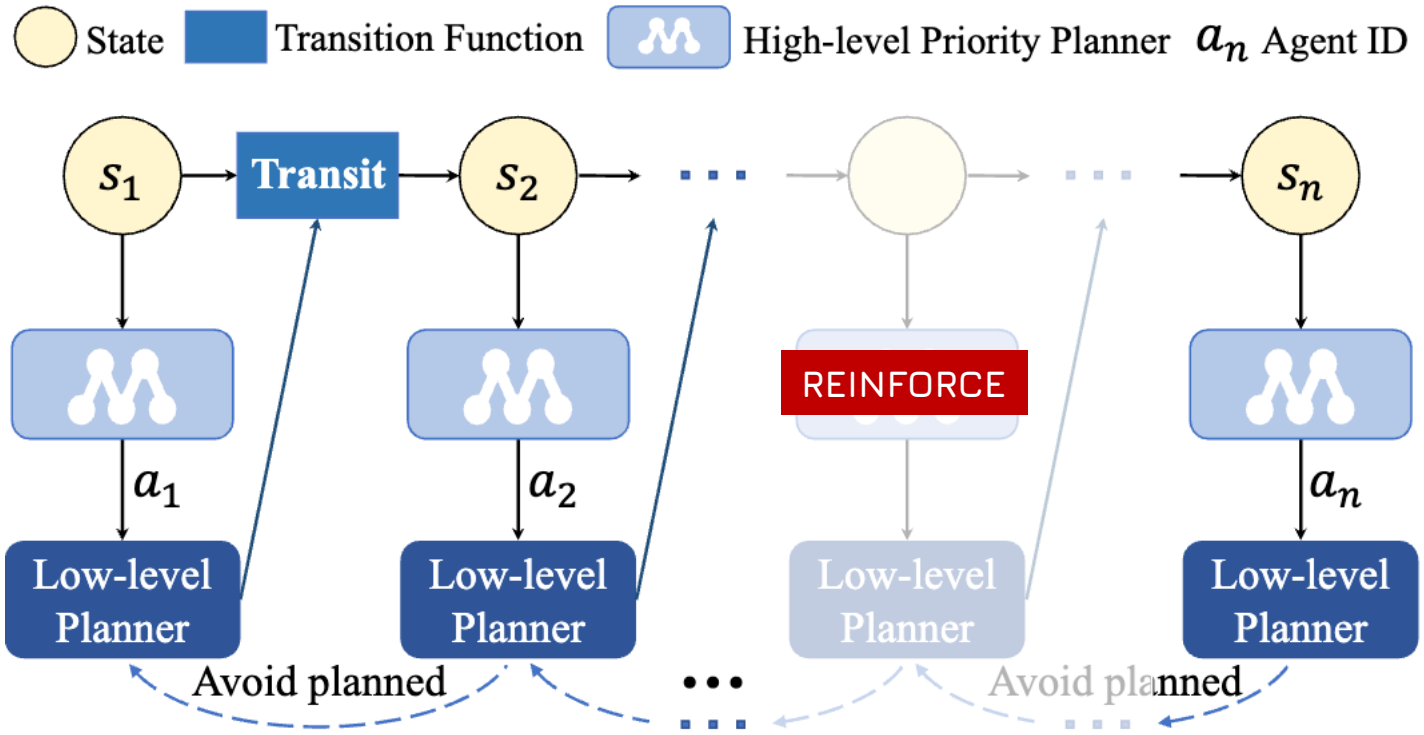
GNN-based feature extraction

Map	Size	A*	ICBS	ICTS	PRIMAL	RL
Berlin_1_256	256x256	577	892	903	873	888
Boston_0_256	256x256	484	718	720	659	701
Paris_01_256	256x256	534	805	799	773	780
brc202d	481x530	198	252	265	242	240
den312d	81x65	467	577	568	560	558
lak303d	194x194	233	377	382	425	403
random-32-32-10	32x32	732	1027	1035	1107	1076
random-32-32-20	32x32	589	862	863	887	865
random-64-64-10	64x64	225	450	492	477	461
random-64-64-20	64x64	618	1078	1112	1045	1127
room-32-32-4	32x32	385	469	480	592	472
room-64-64-16	64x64	542	629	624	686	643
room-64-64-8	64x64	295	360	345	433	381

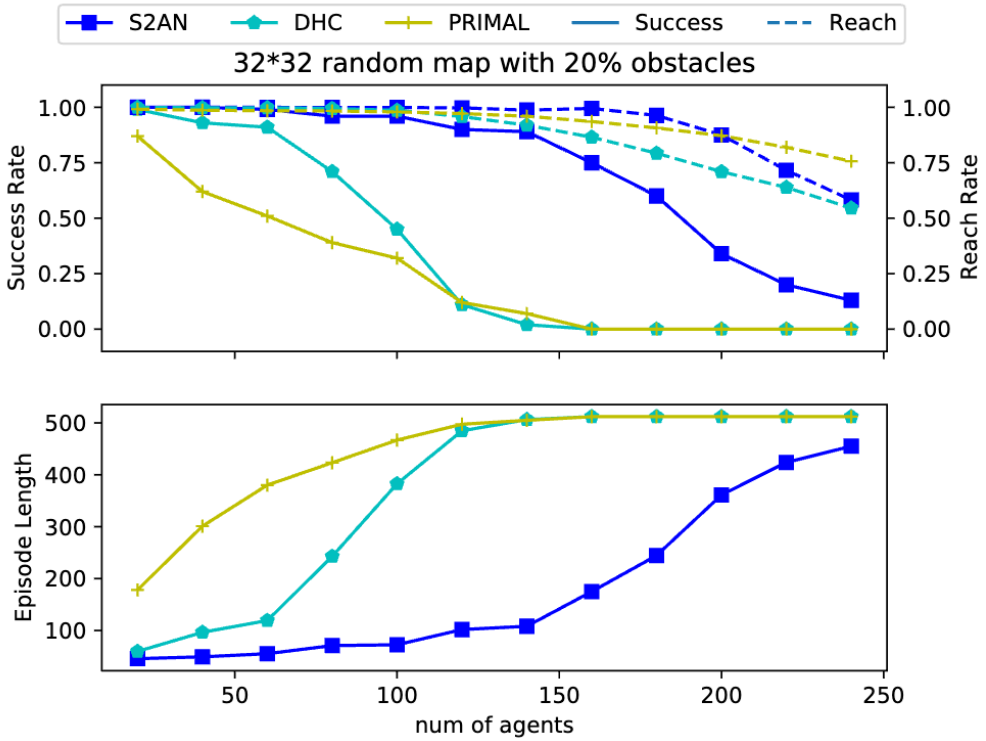
Knippenberg et al. 2021

Reinforcement Learning as the Assistant

Formulating prioritized planning as a Markov Decision Process and introduce a reinforcement learning based prioritized planning paradigm



S2AN (Yang et al. 2024)

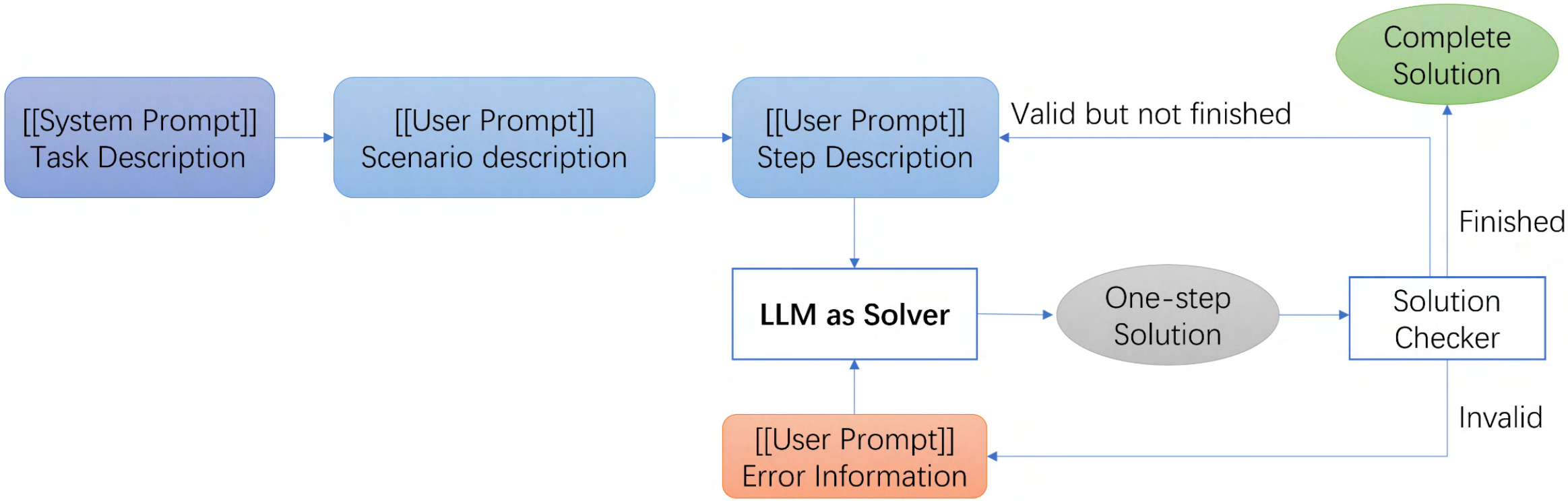


Takeaways

- RL methods are diverse but lack unified benchmarks and evaluation metrics. There is a lack of in-depth comparison between different methods;
- RL methods mainly focus on discrete, one-shot MAPF problems under ideal conditions and do not consider hardware failures, delays, etc., that exist in real-world tasks;
- Transfer, constrained, offline, or model-based RL is very suitable for solving MAPF problems, but there is currently little exploration;
- Regarding the problem scale, the most advanced conventional solvers can reach 1000-3000 agents, but currently, RL methods have conducted fewer large-scale simulation experiments.

One More Thing...

Large Language Models for MAPF



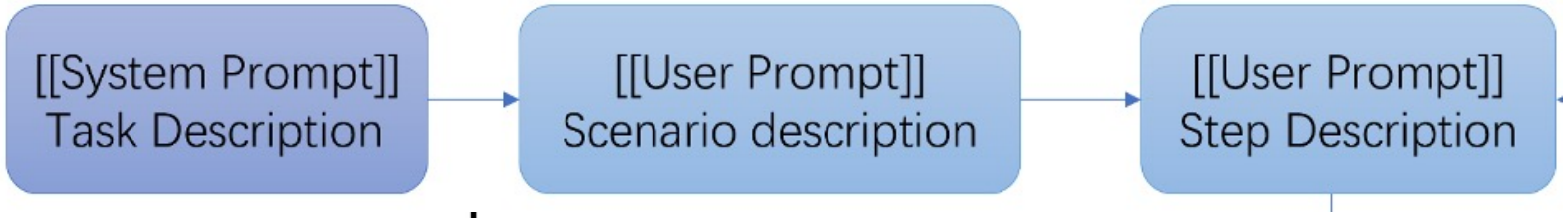
RL paradigm with LLM-based policy-pseudo-gradient

1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

Agent 1 is currently in (0,2), and wants to go to (3,1).
Agent 2 is currently in (1,3), and wants to go to (2,0).
The map is as follows, where '@' denotes a cell with an obstacle that an agent cannot pass, and '.' denotes an empty cell that an agent can pass.
The bottom-left cell is (0,0) and the bottom-right cell is (31,0):
....
...@
....
.@..
In the next step:
Agent 1 can move ['stay at (0, 2)', 'right to (1, 2)', 'up to (0, 3)', 'down to (0, 1)'].
Agent 2 can move ['stay at (1, 3)', 'left to (0, 3)', 'right to (2, 3)', 'down to (1, 2)'].



1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

Agent 1 is currently in (0,2), and wants to go to (3,1).
Agent 2 is currently in (1,3), and wants to go to (2,0).
The map is as follows, where '@' denotes a cell with an obstacle that an agent cannot pass, and '.' denotes an empty cell that an agent can pass.
The bottom-left cell is (0,0) and the botton
(31,0):
....
...@
....
.@..
In the next step:
Agent 1 can move ['stay at (0, 2)', 'right to (1,
3)', 'down to (0, 1)'].
Agent 2 can move ['stay at (1, 3)', 'left to (0
(2, 3)', 'down to (1, 2)'].

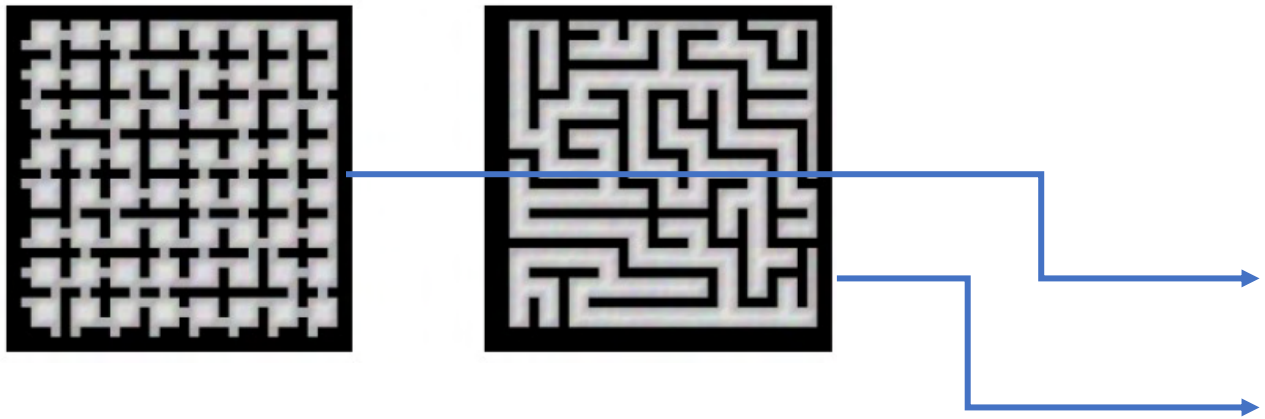
[[User Prompt]]
Error Information

[[Success]]
Good job. Keep moving. In the next step:
Agent 1 can move ['stay at (0, 2)', 'right to (1, 2)', 'up to (0, 3)', 'down to (0, 1)'].
Agent 2 can move ['stay at (1, 3)', 'left to (0, 3)', 'right to (2, 3)', 'down to (1, 2)'].
[[Failure]]
You are wrong. Agent (1,2), (4,5) are colliding with each other. Please correct the current step.
You are wrong. Agent 2,4 is colliding with obstacles. Please correct the current step.

1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF



Map	n	Success Rate (%)	
		OS	SBS
Empty	2	10	100
	4	0	100
	8	0	100
	16	0	60
Room	2	10	100
	4	0	80
	8	0	20
Maze	2	0	0

77% of the failures occurred because the LLM agents began to oscillate in a specific area of the map, while the remaining failures were due to excessively long detours.

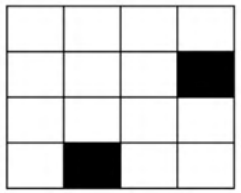
1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

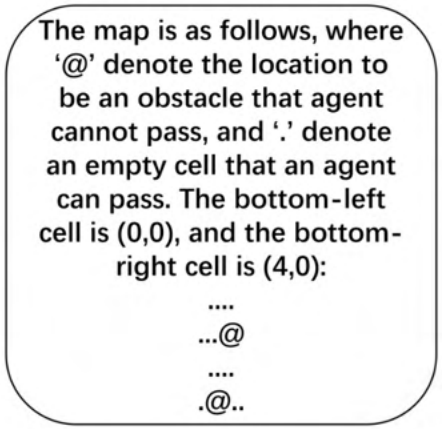
77% of the failures occurred because the LLM agents began to oscillate in a specific area of the map, while the remaining failures were due to excessively long detours.

- Reasoning and Understanding



MM

TOO



TOM

n\Success Rate	MM	TOO	TOM
2	100	100	100
4	20	60	80
8	0	0	20

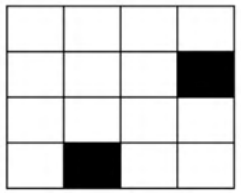
1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

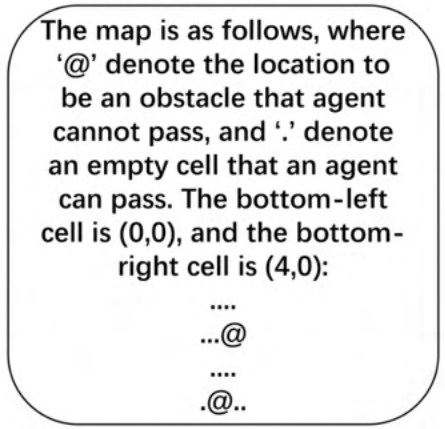
77% of the failures occurred because the LLM agents began to oscillate in a specific area of the map, while the remaining failures were due to excessively long detours.

- Reasoning and Understanding



MM

TOO



TOM

Model	n	Success Rate		Avg. Iterations	
		GO	GO+SSO	GO	GO+SSO
GPT-4-8K	2	80	100	2.7	1.6
	4	20	60	3.0	2.3
	8	0	0	N/A	N/A
GPT-4-128K	2	100	100	2.1	1.2
	4	60	80	2.7	1.4
	8	0	20	N/A	2.4

1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

77% of the failures occurred because the LLM agents began to oscillate in a specific area of the map, while the remaining failures were due to excessively long detours.

- Reasoning and Understanding
- Context Length Limit

1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

One More Thing...

Large Language Models for MAPF

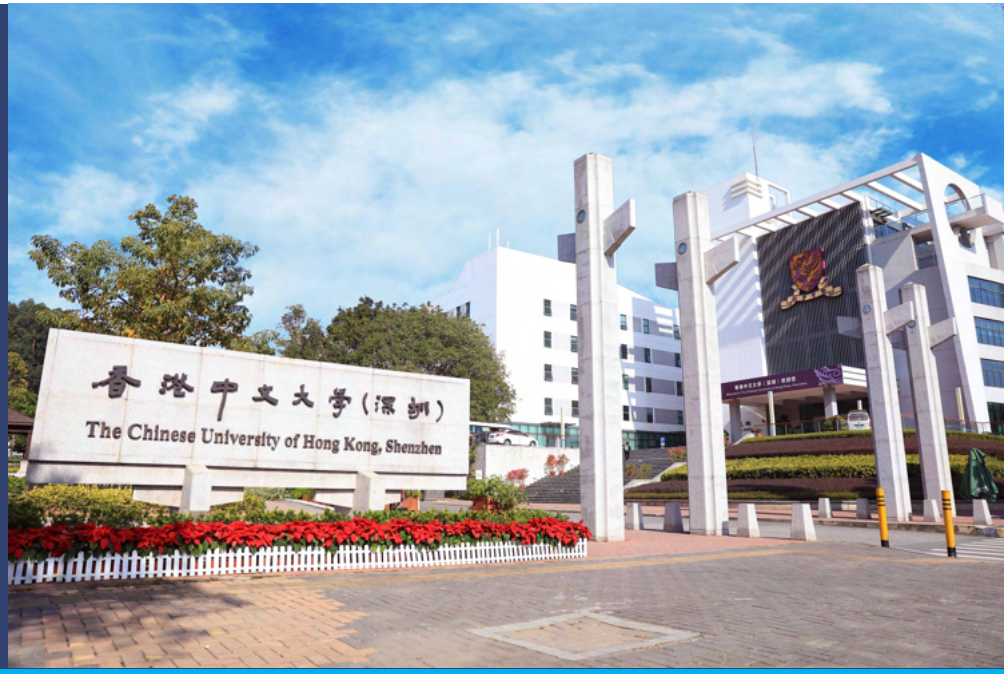
77% of the failures occurred because the LLM agents began to oscillate in a specific area of the map, while the remaining failures were due to excessively long detours.

- Reasoning and Understanding
- Context Length Limit
- Latency

1. Chen, Weizhe, Sven Koenig, and Bistra Dilkina. "Why solving multi-agent path finding with large language model has not succeeded yet." arXiv preprint arXiv:2401.03630 (2024).

RL for MAPF

Thanks for your attention!
Q&A



Xiangfeng Wang, Junjie Sheng (East China Normal University)
Wenhao Li (The Chinese University of Hong Kong, Shenzhen)
Contact email: liwenhao@cuhk.edu.cn