



**Carnegie
Mellon
University**

SACHA: Soft Actor-Critic with Heuristic-Based Attention for Partially Observable Multi-Agent Path Finding

Paper presentation by Kailash Jagadeesh (kailashj)

16-832: Integrated Planning & Learning (Spring 2026)

Date : 2nd Feb 2026



Paper Logistics

Authors: Qiushi Lin and Hang Ma

Year: 2023

ArXiv: <https://arxiv.org/abs/2307.02691>

Codebase: <https://github.com/Qiushi-Lin/SACHA>

Short Summary:

The authors proposed a multi-agent reinforcement learning method that can solve MAPF problems in a decentralized approach with partial observability of other agents.



What?



What is being proposed?

- A decentralized learning-based MAPF solver
- Works under partial observability
- Learns cooperative behavior
- Scales to many agents
- Uses:
 - Soft Actor-Critic
 - Heuristic-based attention
 - Agent-centered critic



Why?

Why MAPF is difficult?

Challenges:

- MAPF is **NP-hard**
- Centralized planners:
 - Need full observability
 - Scale poorly with agent count
- Real robots:
 - Limited sensing (FOV)
 - No global planner at runtime



Why Naive RL Fails

Problems with naive decentralized RL:

- Sparse rewards
- Long horizons
- Non-stationarity
- Selfish behavior
- Poor credit assignment
- Deadlocks & livelocks

Why this method is better?

Core insight of SACHA:

Cooperation requires **intent awareness**, not just collision avoidance.

How?

- Shortest-path heuristics encode **intent**
- Attention selects **which agents matter**
- Agent-centered critic assigns **local credit**



Related Works

Related Work Overview p1

Learning-Based MAPF

- Goal: learn **decentralized policies** for MAPF
- Trained centrally, executed locally (CTDE)
- Avoid replanning from scratch
- Key challenge: **cooperation under partial observability**

Methods Discussed:

- **PRIMAL** : imitation + actor-critic
- **DHC** : IQL + heuristic + broadcast communication
- **DCC** : selective communication on top of DHC

Related Work Overview p2

Cooperative Multi-Agent RL

Challenges:

- Non-stationarity
- Credit assignment
- Scalability

Common solutions:

- Centralized critics (MADDPG, COMA)
- Attention-based critics (MAAC)

Limitations:

- Input grows with agent count
- Poor generalization to new team sizes
- Often require full observability

Why SACHA is Different from these?

Method	Learning Framework	Communication	Single-Agent Guidance	Cooperative Guidance
PRIMAL [8]	A3C (RL) + Behavior Cloning (IL)	Inapplicable	Goal Direction	Goal Directions of Neighbouring Agents
DHC [9]	IQL	Required	Shortest Path Distances	—
DCC [10]	IQL	Required	Shortest Path Distances	—
SACHA (Ours)	Multi-Agent Soft Actor-Critic	Optional	Shortest Path Distances	Shortest Path Distances of Neighbouring Agents



Assumptions?



Main Assumptions + Model Info

Assumptions Made:

1. Each agent has access to static global map of the environment
2. Each agent can compute the shortest path distance by using methods like backward uniform cost search (dijkstra's , BFS)
3. Agents can see other agents in field of view and communicate with them.
4. Deterministic transitions, deterministic observations with a grid world setup

Model Used:

- Decentralised POMDP
- Shared reward objective
- Partial observability with fixed FOV



How does it work?

Methodology pt1: Multi-Agent Heuristic Maps

What each agent sees:

- Its own shortest-path distance map
- Neighbors' shortest-path distance maps (within FOV)

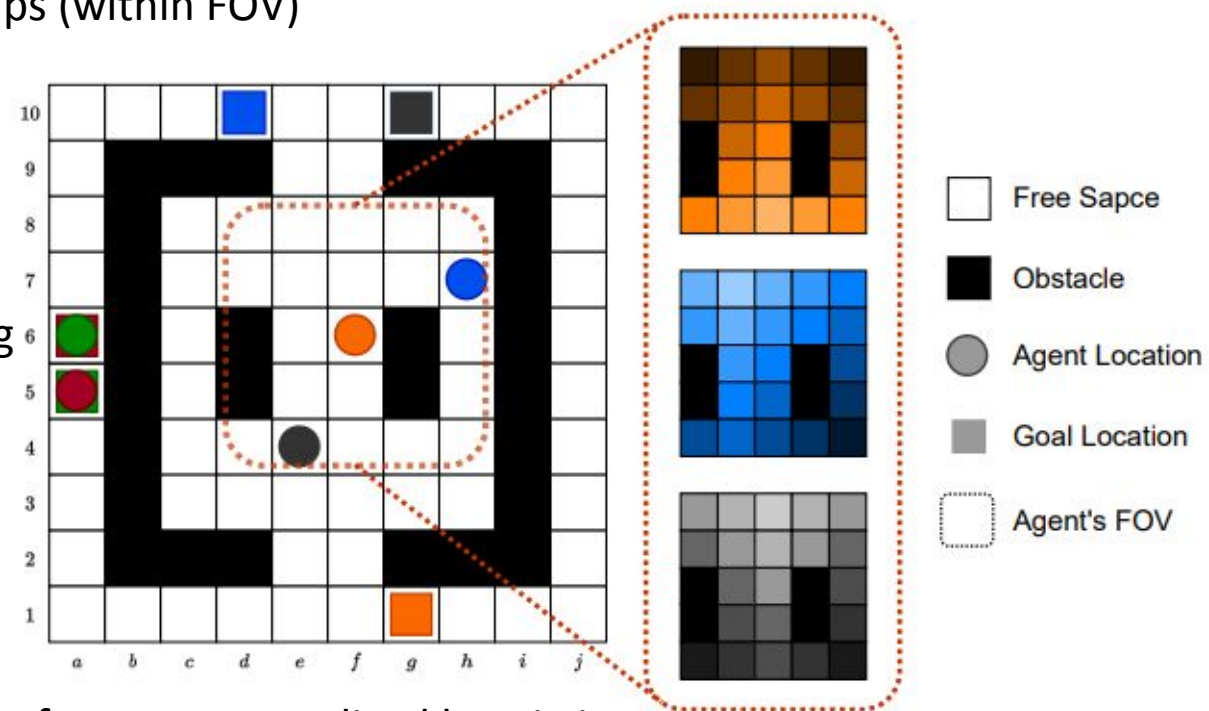
Why:

- Encodes **goal intent**
- Enables cooperation without planning

For each agent, a feature map is created,

$$F_i^j \in \mathbb{R}^{L \times L \times 3} \quad \text{with } K-1 \text{ feature maps} \\ \text{for } 1 \text{ agent.}$$

where each channel consists of a binary matrix of free space, matrix with positions of agents, normalised heuristic distances.



Methodology pt2: Reward Design

Baseline reward:

- Time penalty
- Collision penalty
- Goal reward

Heuristic shaping:

- Add **negated normalized distance to goal**
- Dense, smooth progress signal
- Does **not** change optimal policy

Action	Reward
Move (up / down / left / right)	-0.075
Wait (on goal, away goal)	0, -0.075
Collision (obstacles or agents)	-0.5
Reaching Goal	3

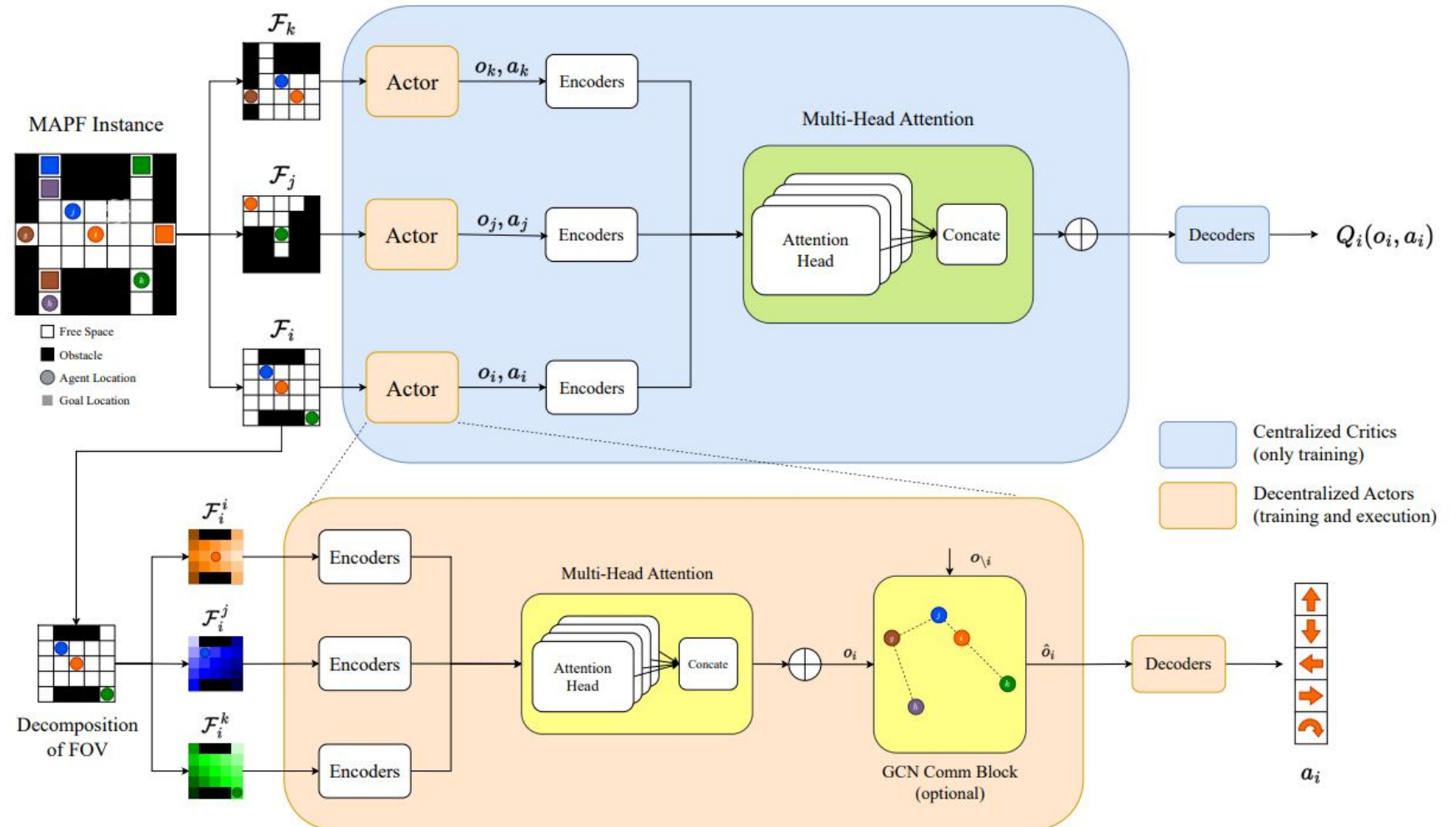
Individual Rewards
Based on DHC

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

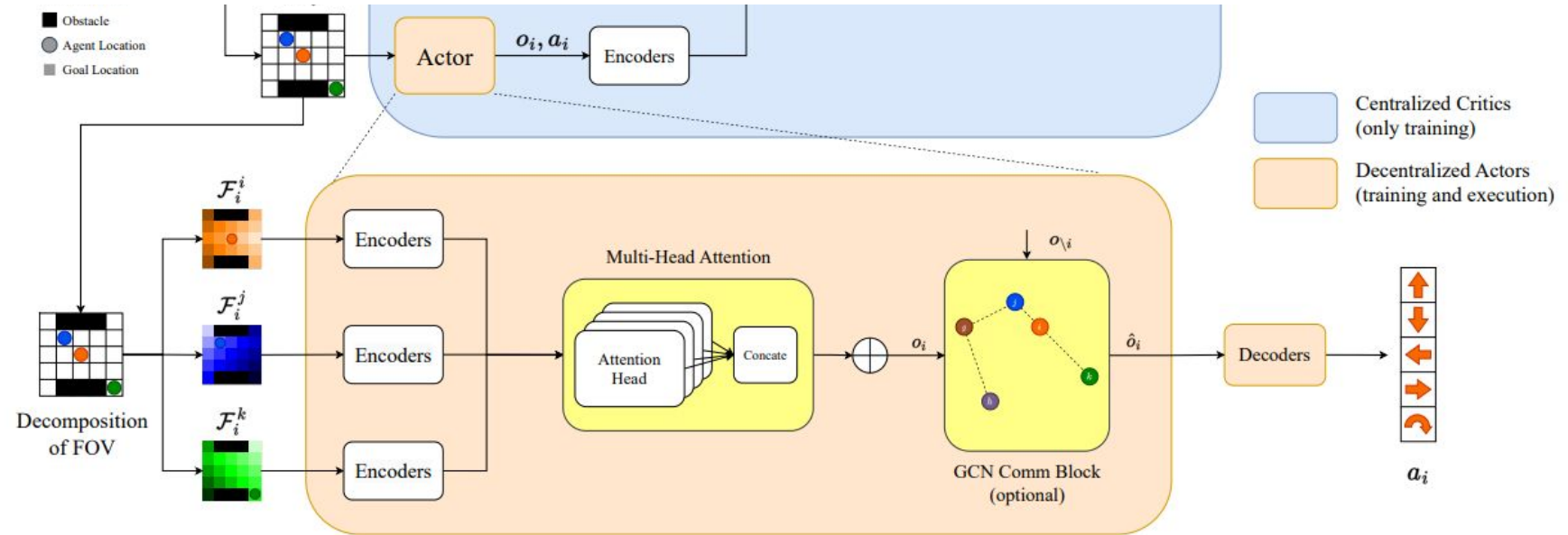
Modifying reward function using heuristic guidance based on HuRL

Methodology pt3: Model Design

Starts with an observation graph of agents



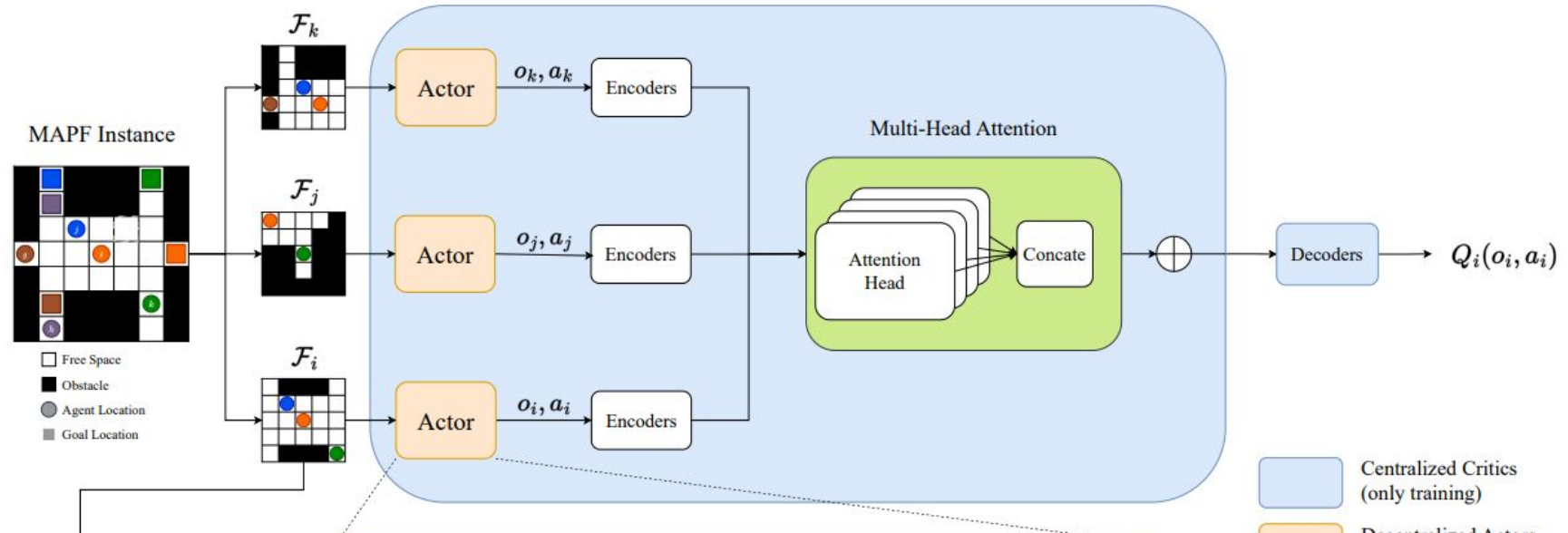
Methodology pt3: Model Design



Actor Architecture:

1. CNN + GRU encoders (per-agent maps)
2. Multi-head attention over agents
3. Aggregation
4. Action decoder

Methodology pt3: Model Design



Critic Architecture:

- Per-agent encoders (shared with actor)
- Action-conditioned attention over neighbors
- Local aggregation
- Agent-centered Q-value head

Methodology pt4: Learning Rule

Learning Rule used: Soft Actor Critic (SAC) with Multi-Agent Advantage Function

Why Soft Actor-Critic:

- Stable
- Encourages exploration
- Handles non-stationarity

Counterfactual baseline:

- Measures agent's marginal contribution
- Reduces variance
- Enables fair credit assignment

Action value temporal difference

$$\delta_i = Q_i^\psi(o_i, a_i) - \tilde{r}_i - \gamma \mathbb{E}_{a'_i \sim \pi_{\bar{\theta}}(o'_i)} [Q_i^\psi(o'_i, a'_i) - \alpha \log(\pi_{\bar{\theta}}(a'_i|o'_i))]$$

MSE loss optimisation by Critic Network

$$L_Q(\psi) = \mathbb{E}_{e \sim \mathcal{D}} \left[\frac{1}{M(e)} \sum_{i=1}^{M(e)} \delta_i^2 \right]$$

Policy gradient update for Actor Network

$$\nabla_{\theta_i} J(\theta) = \mathbb{E}_{o_i \sim D, a_i \sim \pi_{\theta_i}(o_i)} [\nabla_{\theta_i} \log(\pi_{\theta_i}(a_i|o_i)) \underbrace{(Q_i^\psi(o_i, a_i) - b(o_i, a_{\setminus i}) - \alpha \log(\pi_{\theta_i}(a_i|o_i)))}_{\text{advantage}}]$$



Methodology pt5: Communication Module

SACHA(C):

- Shares learned embeddings
- Uses GCN for message passing
- Extends information beyond FOV



Guarantees?



Theoretical guarantees applicable?

What SACHA does NOT guarantee

- No optimality guarantee
- No convergence guarantee to global optimum
- No bounded suboptimality (unlike CBS / wCBS)

What SACHA does guarantee

- Unbiased policy gradient
- Equivalence to Independent Q Learning (IQL) objective
- Correct credit assignment via counterfactual baseline



Experimental Analysis

Analysis

Training:

- Curriculum learning
- Random maps
- Up to 72 agents

Testing:

- Standard MAPF benchmarks
- 4–64 agents
- Strict time limits

Key findings:

- Centralized planners(CBS,ODrM*, PBS) time out
- SACHA > PRIMAL, DHC, DCC
- SACHA(C) best at large scale

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



Limitations?

Limitations?

Limitations:

- Not optimal
- Requires known map & goals
- Heuristic computation required
- No formal convergence guarantees

Takeaways:

- Heuristic-based attention enables cooperation
- Agent-centered critic improves generalization
- Works at scales where centralized planning fails



Thank you!