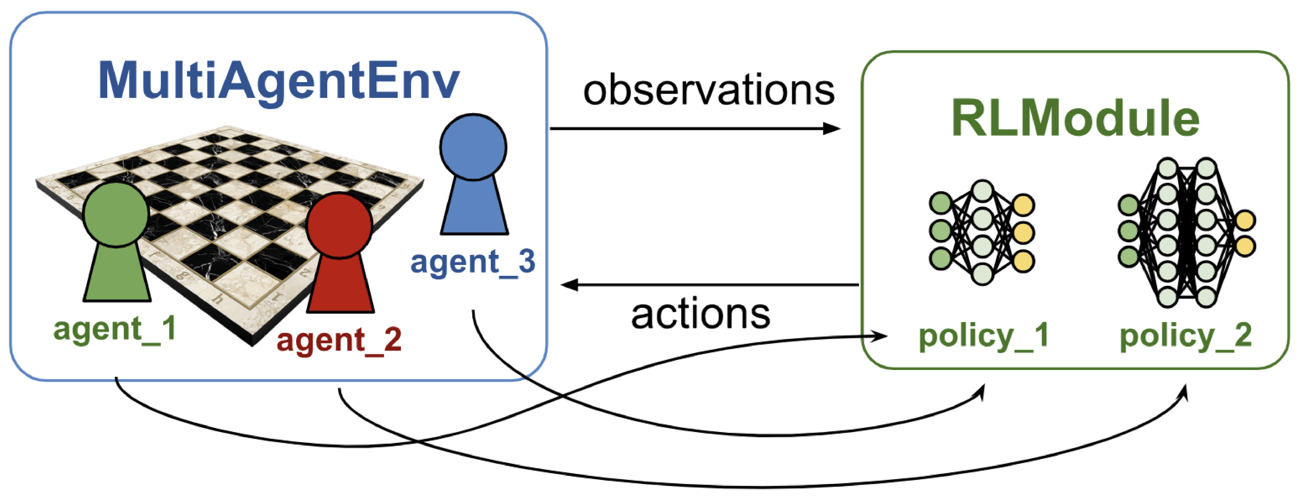
Multi-Agent Reinforcement Learning

Seouh-won Yi[1], **Sungwoo Cho**[2] and **Daesoon Kim**[3]
[1]uniqueseouh@snu.ac.kr, [2]sxxgwoo@snu.ac.kr, [2]kimds929@snu.ac.kr

Multi-Agent Partially Observable MDP



- A Centralized POMDP is defined by: $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}, P, R, \{\mathcal{O}_i\}, O, \gamma \rangle$
  - $\mathcal{I} = \{1, \dots, N\}$: set of agents
  - $P(s' \mid s, \mathbf{a})$, $R(s, \mathbf{a})$, $O(\mathbf{o} \mid s', \mathbf{a})$: joint state transition / reward / observation
- At each time step $t = 1, \dots, T$:
  1. Each agent i receives private observation $o_t^i \sim O_i(\cdot \mid s_t)$
  2. Each agent selects action a_t^i based on its own history
  3. Joint action $\mathbf{a}_t = (a_t^1, \dots, a_t^N)$ is executed
  4. Environment transitions to $s_{t+1} \sim P(\cdot \mid s_t, \mathbf{a}_t)$
  5. Shared reward $r_t = R(s_t, \mathbf{a}_t)$ is received
  6. Agents observe new private observations o_{t+1}^i
- **Goal**: Maximize $\mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} R(s_t, \mathbf{a}_t)\right]$ over possible polices $\{\pi_i\}$

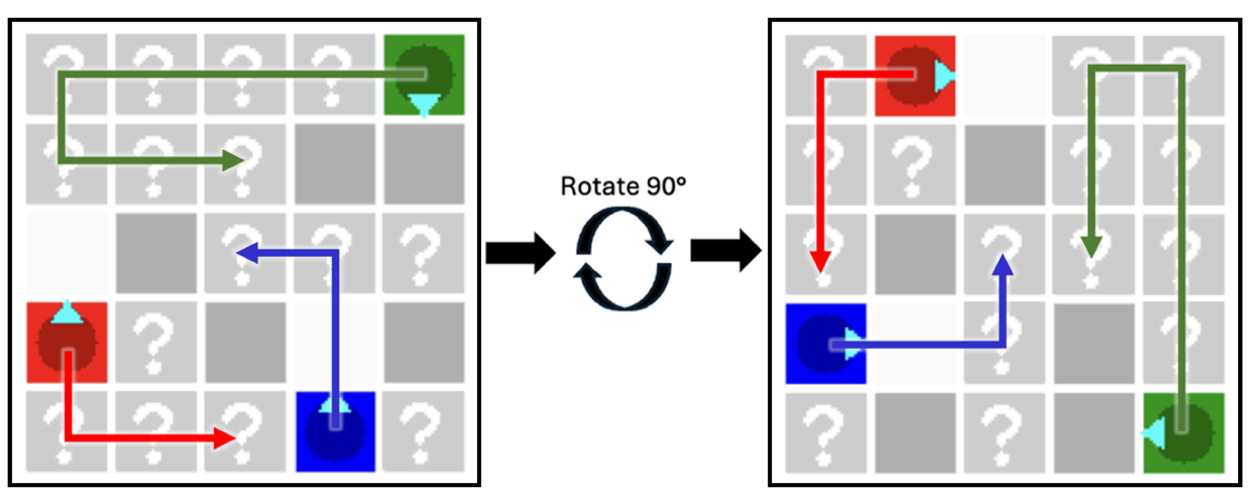
Open Challenges in Multi-Agent RL

- **Decentralized setting:** Each agent selects actions using local observations only
- **Partial observability:** Restrictive observation on state or on other agents
- **Non-stationarity:** Policies of other agents change over time
- **Credit assignment:** Global reward must be attributed to individual actions
- **Observation complexity:** Multi-agent settings often limits the effectiveness of standard algorithms on single-agent environments

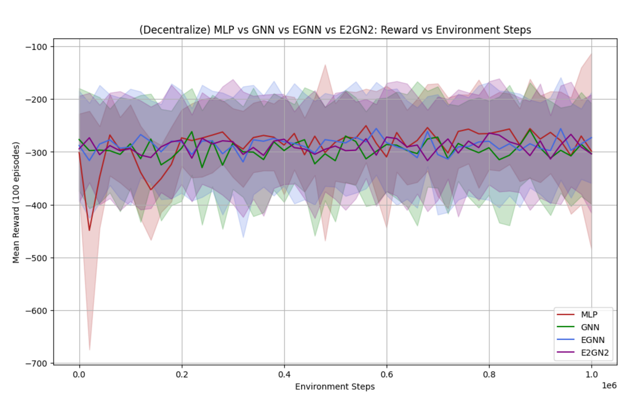
Research Questions:

1. Does communication via graph-based structures improve sample efficiency in Multi-Agent Reinforcement Learning?
   → *Analyze and reproduce recent graph-based models (EGNN[2] and E2GN2[1])*
2. How can we design a partially observable environment for multi-agent tasks?
   → *Design a novel environment and evaluate policy learning in this setting*
3. How does Centralized vs. Decentralized control affect performance, and how crucial is relative information?
   → *Compare them in the MPE (Multi-Agent Particle Environment)*

How to Improve Sample Complexity?



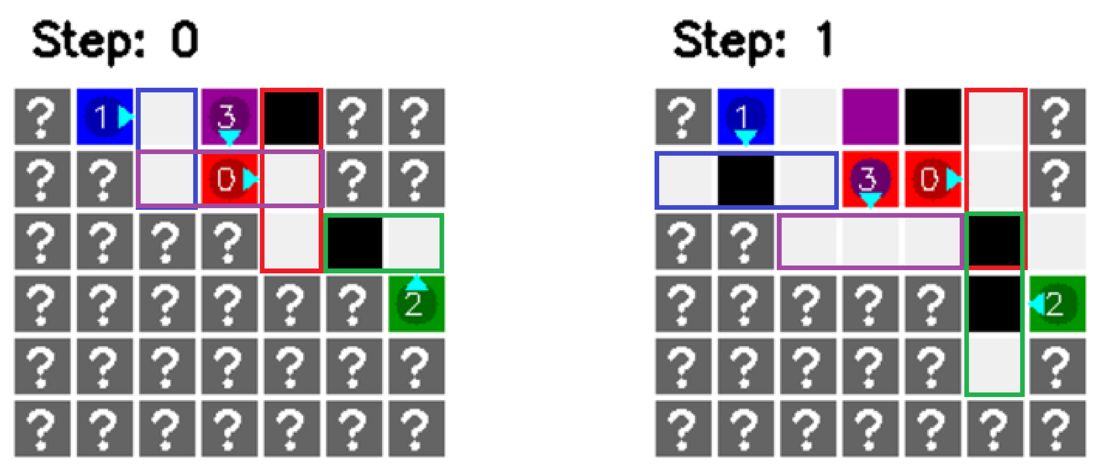
- **State Similarity and Equivalence**: Reinforcement learning algorithms often waste samples exploring equivalent or symmetric states.
  - Graph-based representations can reduce redundancy by encoding structured inductive biases.
  - Edge information can encode relative or absolute features — which is more effective?



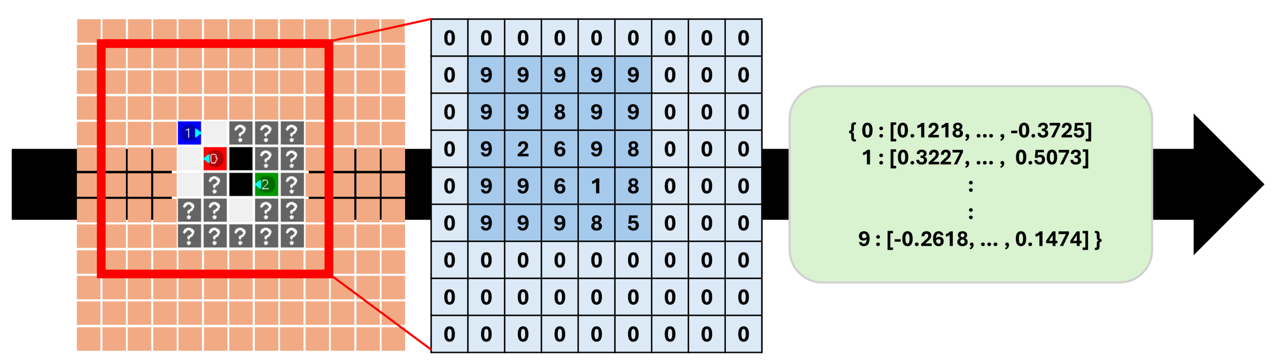
- **Result**
  - Comparison of 4 actor policies
    - PPO(MlpPolicy), GNN, EGNN, E2GN2
  - In the decentralized case, no significant difference in performance was observed.

MA-POMDP Task: Area Coverage

1. Real-world Cleaning robot can **stop, rotate and go forward**.
2. **Reward System**
   - **Global reward:** $+100$ when all tiles are covered / -0.05 at every step
   - **Local reward (per agent):**
     - *Collision penalty:* -1 for wall bump, -2 for agent collision
     - *Coverage:* $+5$ for newly covered tile, -0.1 for revisiting
3. **Partial observability:** each agent sees a local patch (3 blocks in front of the agent) and share observations with the other agents.

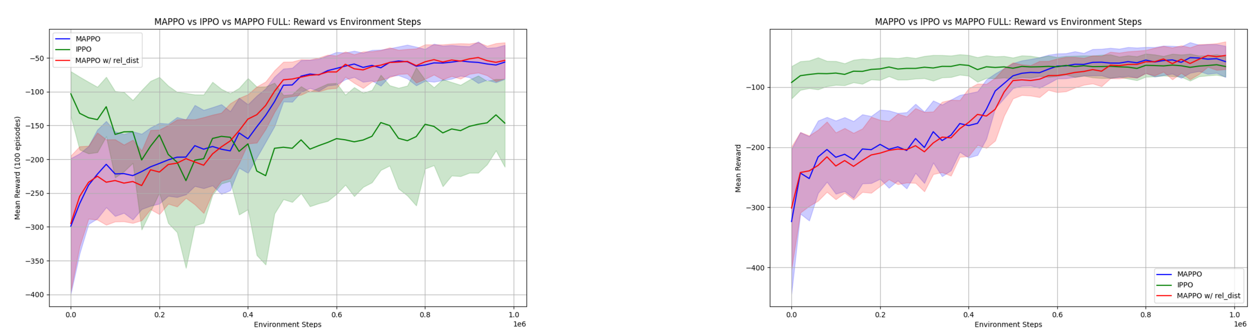


4. **State & Observation Implementation**
   - Each agent receives an observation of shape (grid size$_y$, grid size$_x$, 5)
   - The observation includes:
     – grid-wise state information (1∼9 ; e.g., wall, covered, undiscovered)
     – absolute and relative positions (e.g., x/y differences between each grid cell and the agent, considering agent's orientation)
5. **Objective of the Project**
   - Policy for efficient coverage in a **fixed-grid environment** (PPO + MlpPolicy)
   - Efficient cooperative policy that covers **arbitrary, dynamic environments**
   - **Forwarding idea**:
     - Leverage graph concepts to preprocess and utilize observations
     - Add padding to center the agent (orientation fixed), update cell-wise state information, then map to learnable embeddings



Centralized vs. Decentralized

- **Why Centralization?** Centralized critics or shared policies can stabilize learning in cooperative settings, but how much information is really needed?
- **Result Discussion**:



  - **MPE Simple Spread**: A cooperative environment where agents work together to navigate and reach landmarks.
  - **Decentralized PPO** with absolute coordinates performs worse with higher variance; adding relative coordinates improves stability and performance.
  - **Centralized PPO** performs similarly with or without relative coordinates.
  - Centralized PPO achieves much better performance than Decentralized PPO, highlighting the effectiveness of centralized training in cooperative settings.

Discussion

- Graph models underperformed MLPs, indicating a need for better feature design or tuning to realize their strengths.
- Leveraging environment-specific priors enhanced the model to learn effective policies.
- Centralized policies show better stability and performance than decentralized ones, highlighting the value of shared information.

Link to Result Examples

References

[1] Joshua McClellan, Naveed Haghani, John Winder, Furong Huang, and Pratap Tokekar. Boosting sample efficiency and generalization in multi-agent reinforcement learning via equivariance. *ArXiv*, 2024.
[2] Kieran Nehil-Puleo, Co D. Quach, Nicholas C. Craven, Clare McCabe, and Peter T. Cummings. E(n) equivariant graph neural network for learning interactional properties of molecules. *The Journal of Physical Chemistry B*, 2024.