Learning to Simulate Self-Driven Particles System with Coordinated Policy Optimization

Zhenghao Peng[⋆], Quanyi Li[§], Ka Ming Hui[⋆], Chunxiao Liu[†], Bolei Zhou[⋆]
[⋆]The Chinese University of Hong Kong, [†]SenseTime Research
[§]Centre for Perceptual and Interactive Intelligence

NeurIPS 2021

Preliminaries

- What are Self-Driven Particles (SDP) Systems?

Birds Flock

Fish School

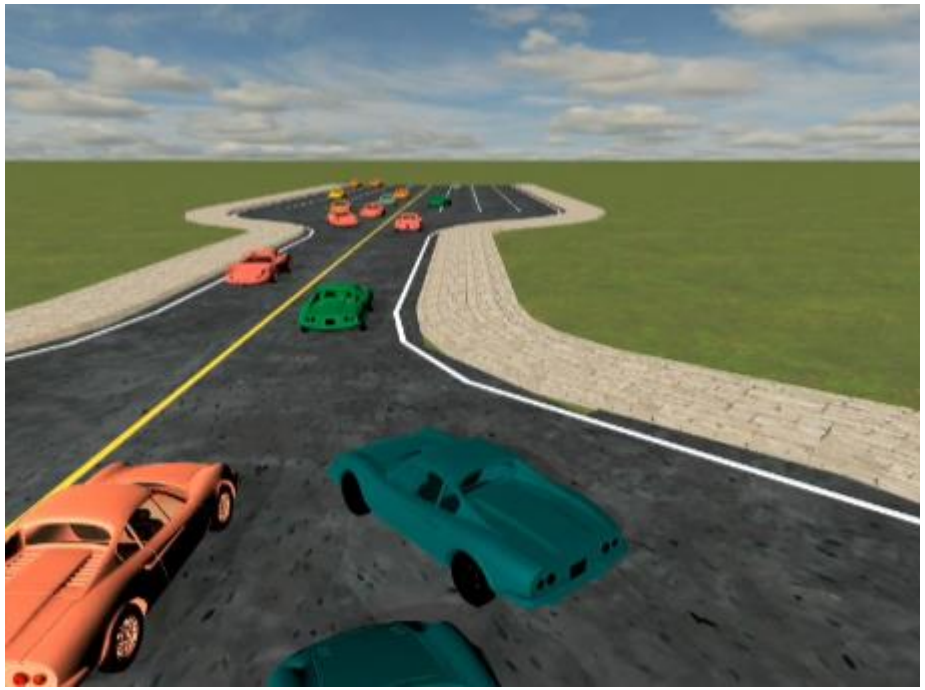
Human Crowd

Traffic System

Motivation

Features of SDP systems

1. Each individual agent is self-interested.
2. The relationship between agents are time-varying.



Cooperation: Yield to other

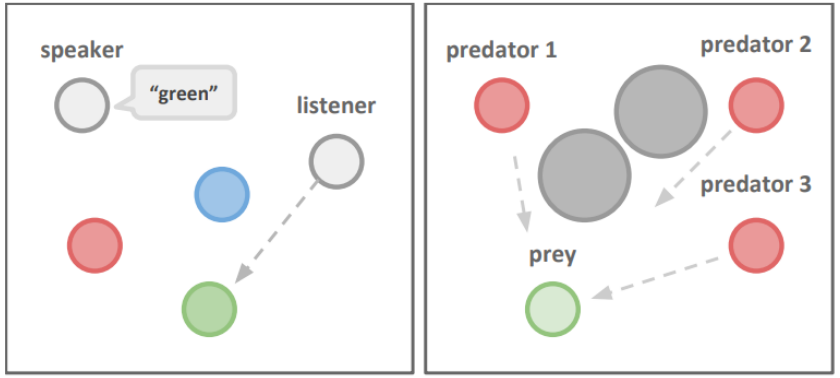
Competition: Cut in

Motivation

Multi-Agent Reinforcement Learning

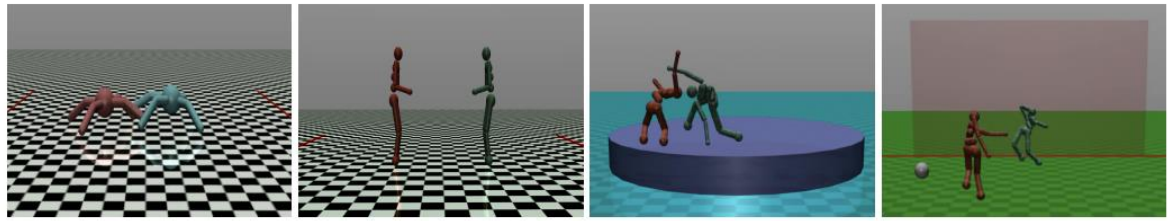
Cooperative(within team)

StarCraft Multi-Agent Challenge(SMAC)



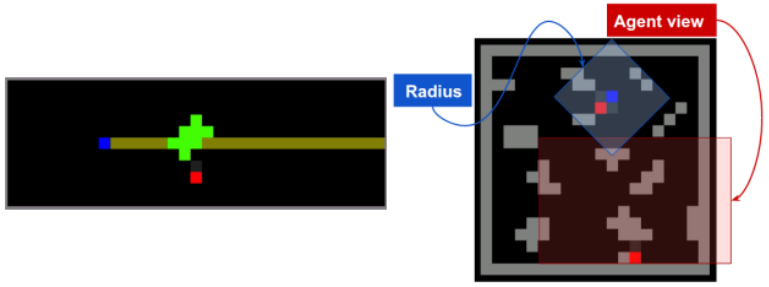
Multi-Agent Particle Env.

Competitive



Competitive Multi-Agent Environments

mixed-motive RL



Sequential Social Dilemma Games

Coordinated Policy Optimization (CoPO)

Framework

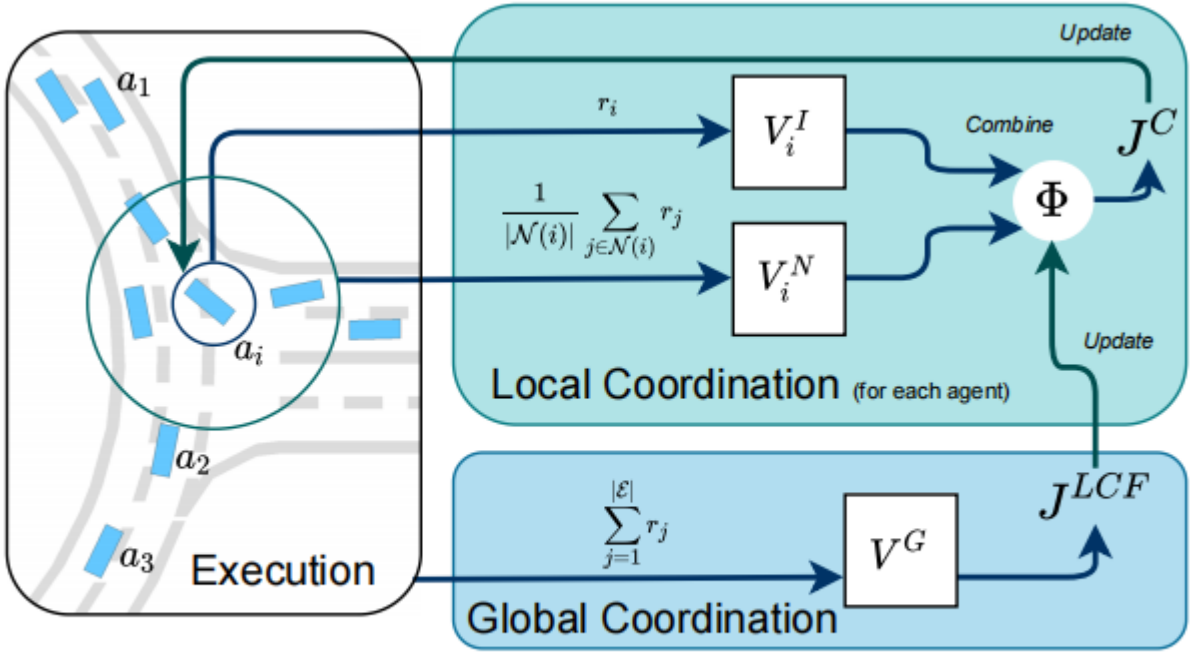


Figure 2: The framework of the CoPO method.

Coordinated Policy Optimization (CoPO)

Local Coordination

Local Coordination

Individual Reward r_1^I

Mixed by Local Coordination Factor

Neighborhood Reward $r_1^N = \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} r_i$

Local Coordination Factor (LCF)

$\Phi \in [-90°, 90°]$

Coordinated Reward

$$r_1^C = \cos(\Phi) r_1^I + \sin(\Phi) r_1^N$$

$$J_i^C(\theta_i, \Phi) = \mathop{\mathbb{E}}_{(s,\mathbf{a})}[\min(\rho A_{\Phi,i}^C, \text{clip}(\rho, 1-\epsilon, 1+\epsilon) A_{\Phi,i}^C)]$$

PPO

Update policy!

How to determine Φ ?

1. $\Phi = 0°$: $r_1^C = r_1^I$ ⟶ egoistic
2. $\Phi = 90°$: $r_1^C = r_1^N$ ⟶ altruistic
3. $\Phi = -90°$: $r_1^C = -r_1^N$ ⟶ sadistic

Coordinated Policy Optimization (CoPO)

Global Coordination

Global Objective: $J^G = \sum_{\forall i} \sum_{\forall t} r^I_{i,t}$ sum of reward of all agents at all steps

Meta-gradient to update ϕ :

$$\nabla_\Phi J_i^G(\theta_i^{\text{new}}) = \underbrace{\nabla_{\theta_i^{\text{new}}} J_i^G(\theta_i^{\text{new}})}_{\text{1st Term}} \underbrace{\nabla_\Phi \theta_i^{\text{new}}}_{\text{2nd Term}}$$

policy gradient

$$\nabla_\Phi(\theta_i^{\text{old}} + \alpha \nabla_{\theta_i^{\text{old}}} J_i^C(\theta_i^{\text{old}}, \Phi))$$

where

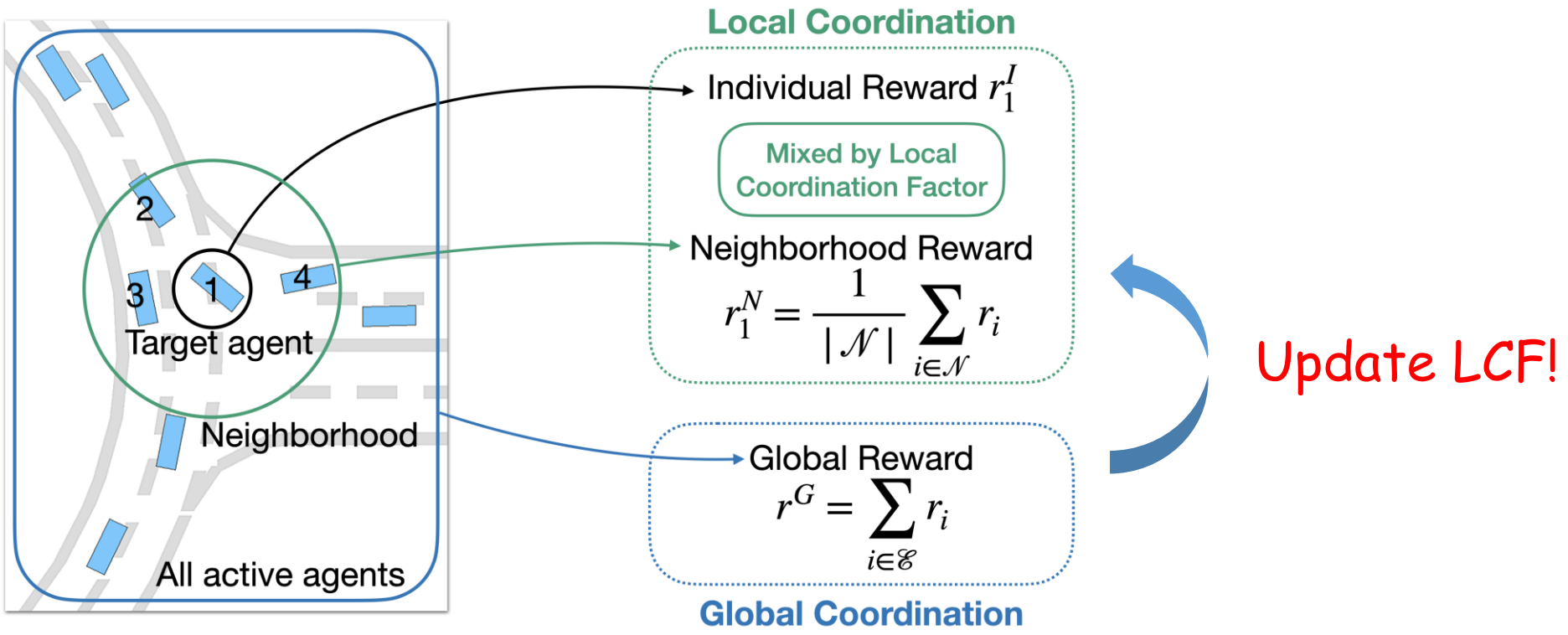
$$J_i^C \sim A_i^C = \cos(\Phi) A_i^I + sin(\Phi) A_i^N$$

$$\nabla_\Phi J^G(\Phi) = \nabla_\Phi \mathbb{E}[A^G \nabla_{\theta_i^{new}} \log \pi_{\theta_i^{new}}(a|s)][(\cos(\Phi) A_i^I + sin(\Phi) A_i^N) \nabla_{\theta_i^{old}} \log \pi_{\theta_i^{old}}(a|s)]$$

Denote the parameters of policies before and after optimizing Local Coordination as θ_i^{old} and θ_i^{new}

Coordinated Policy Optimization (CoPO)

Summary



Experiments-environments

Experiments-Main Results

Table 1: Success rate of different approaches.

	Roundabout	Intersection	Tollgate	Bottleneck	Parking Lot
IPO	70.81 ±1.95	60.47 ±5.79	82.90 ±2.81	72.43 ±3.79	61.05 ±2.81
MFPO	64.27 ±3.68	67.74 ±4.19	81.05 ±3.07	67.40 ±4.77	53.96 ±4.65
CL	65.48 ±3.96	62.03 ±4.41	73.72 ±3.46	68.81 ±4.39	60.62 ±2.25
CoPO (Ours)	**73.67** ±3.71	**78.97** ±4.23	**86.13** ±1.76	**79.68** ±2.91	**65.04** ±1.59

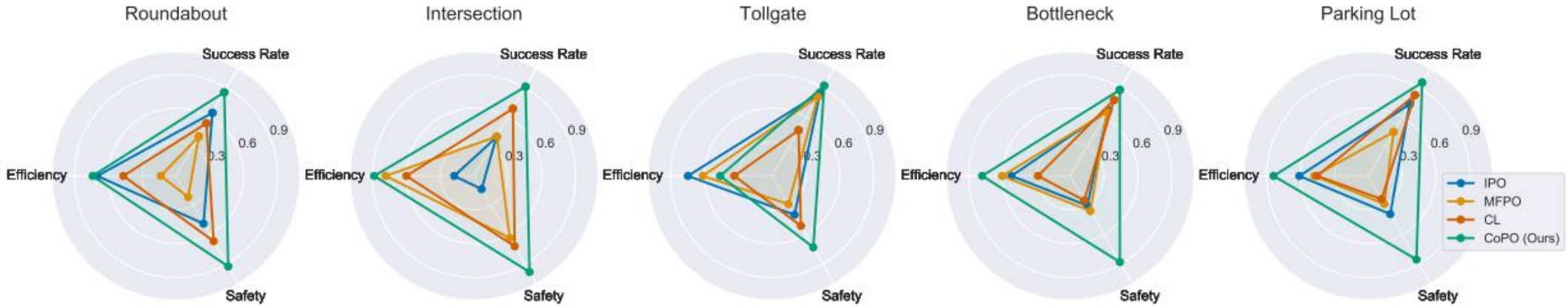
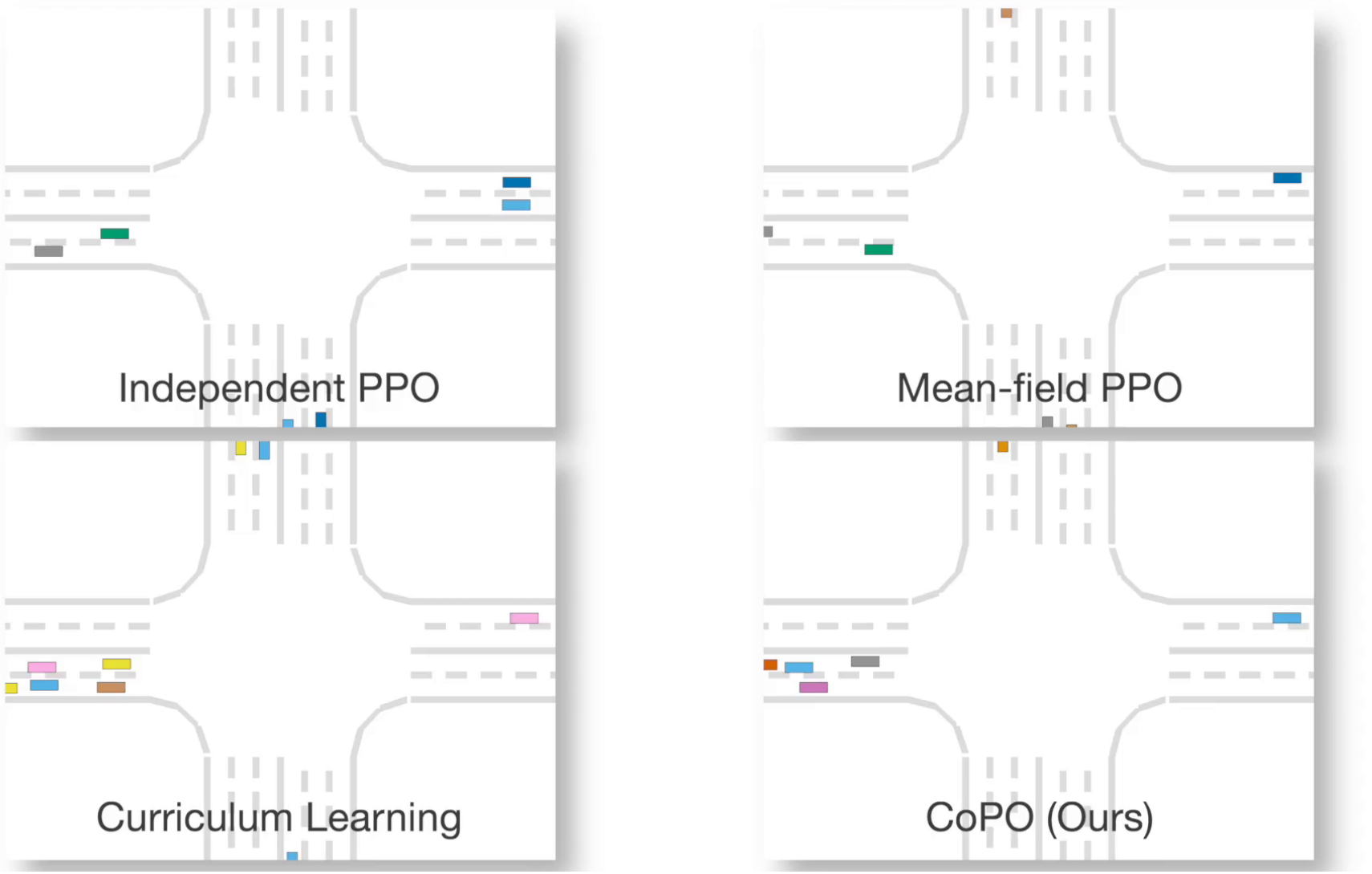
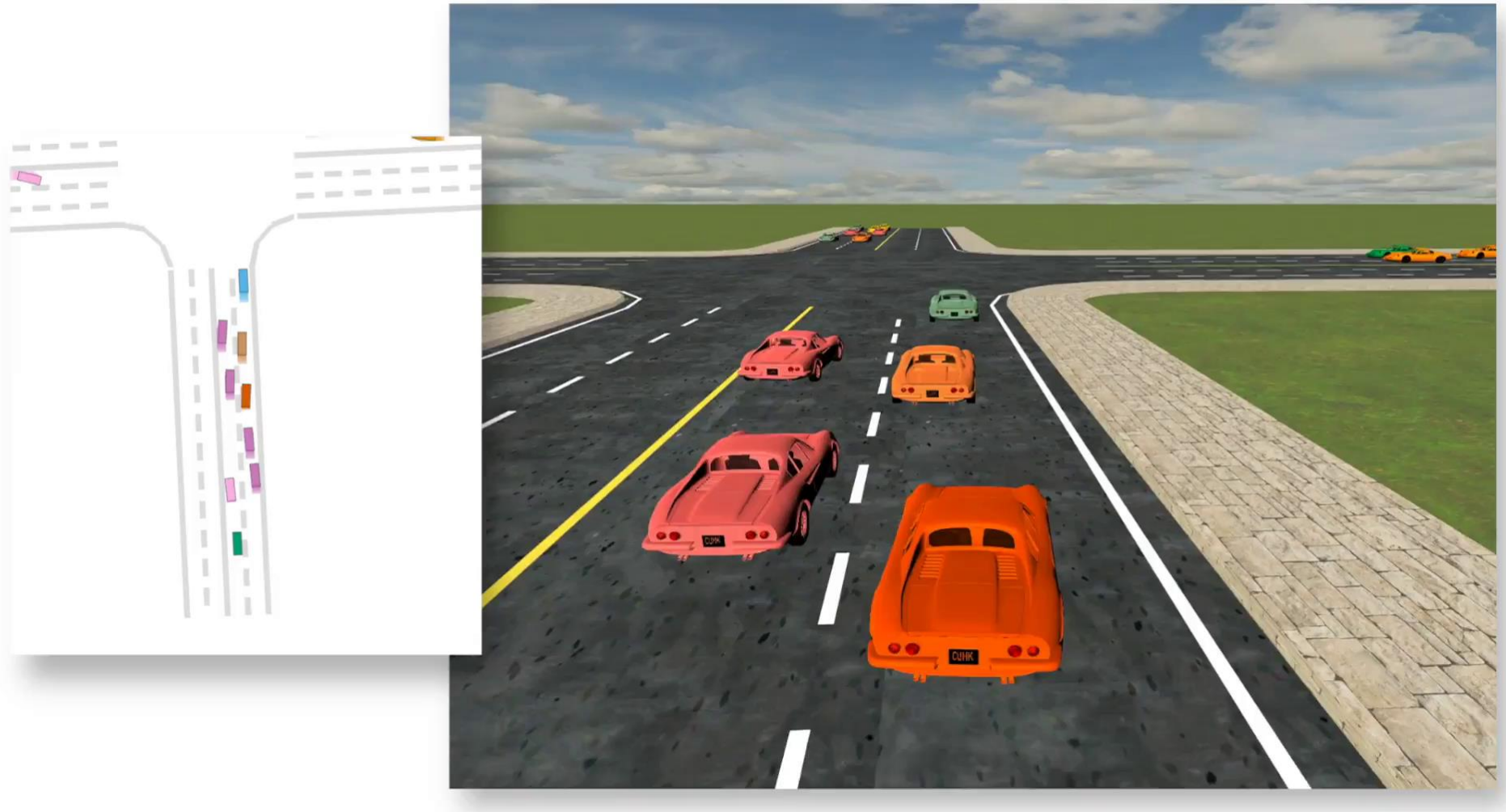


Figure 4: Performance of the trained populations from different MARL methods.

Experiments-Main Results



Experiments-Behavioral Analysis



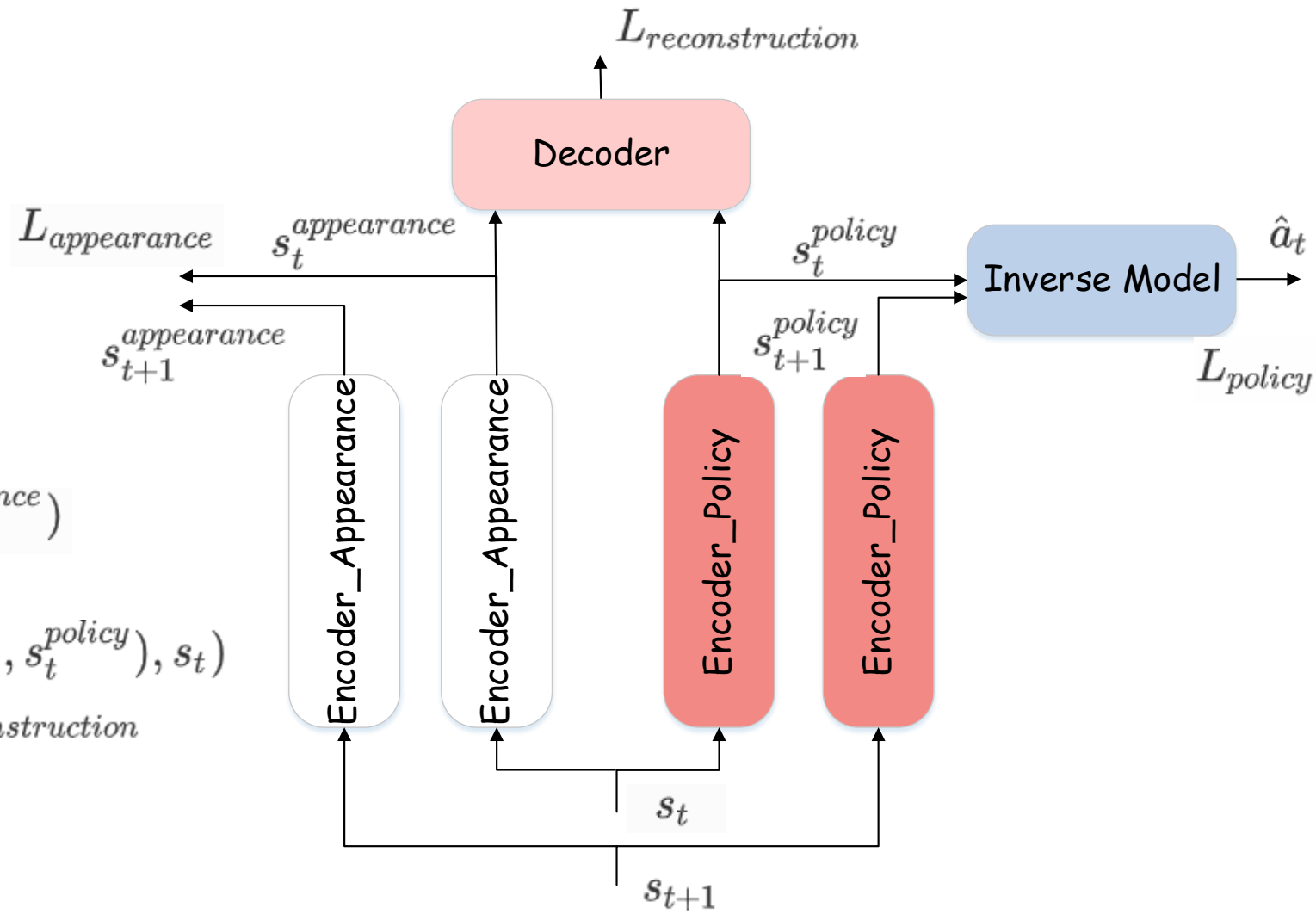
Experiments-Ablation Studies

Table 2: Ablation study on the effectiveness of global coordination. Average success rate is provided.

Experiment	Roundabout	Intersection	Tollgate	Bottleneck	Parking Lot
(a) sample ϕ from $\mathcal{N}(0, 0.1^2)$	62.38 $\pm$7.26	70.23 $\pm$2.72	60.47 $\pm$5.80	71.29 $\pm$3.27	59.14 $\pm$1.29
(b) maximize global reward	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
(c) maximize neighborhood reward	65.70 $\pm$ 2.21	66.83 $\pm$ 2.16	57.30 $\pm$ 3.46	71.62 $\pm$ 1.22	6.34 $\pm$ 1.44
CoPO: sample ϕ from $\mathcal{N}(\phi_\mu, \phi_\sigma^2)$	73.67 $\pm$3.71	78.97 $\pm$4.23	86.13 $\pm$1.76	79.68 $\pm$2.91	65.04 $\pm$1.59

Discussion

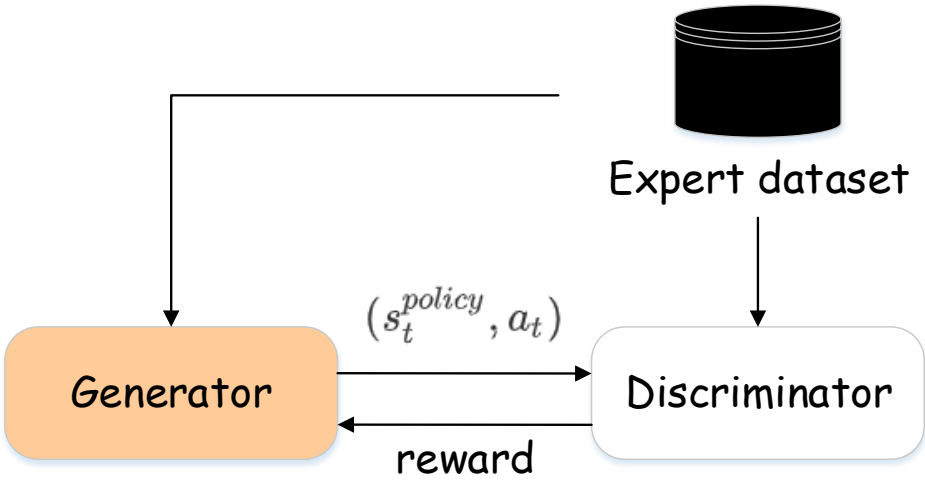
- EncDecInv

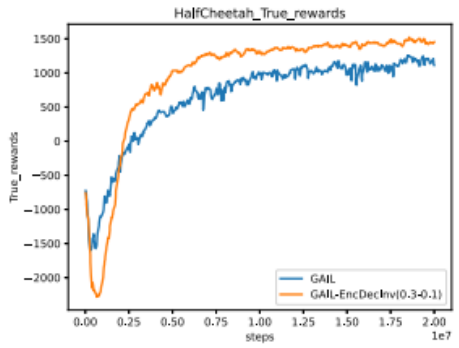
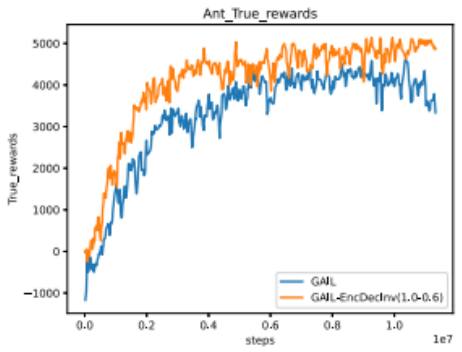


$$L_{appearance} = cosin(s_t^{appearance}, s_{t+1}^{appearance})$$
$$L_{policy} = mse(\hat{a}_t, a_t)$$
$$L_{reconstruction} = L_1(Decoder(s_t^{appearance}, s_t^{policy}), s_t)$$
$$L_{total} = L_{policy} + \lambda L_{appearance} + \beta L_{reconstruction}$$

Discussion

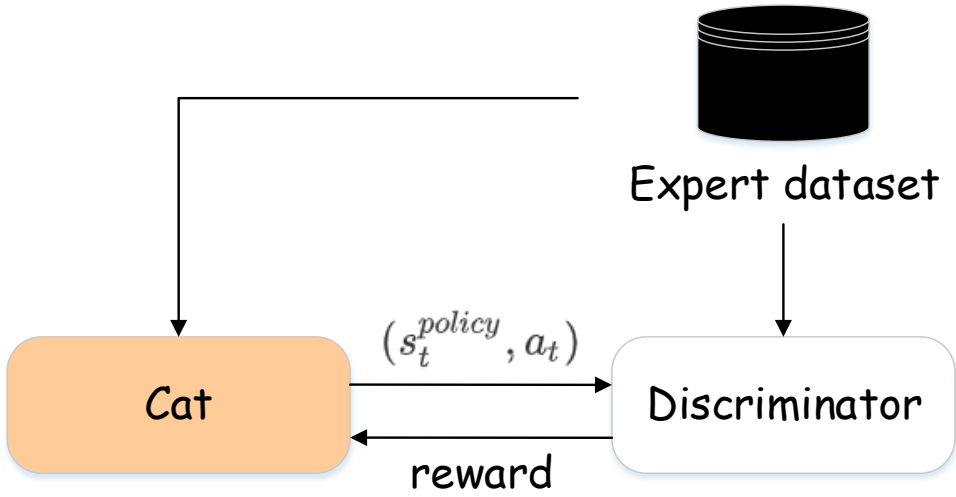
- EncDecInv-GAIL

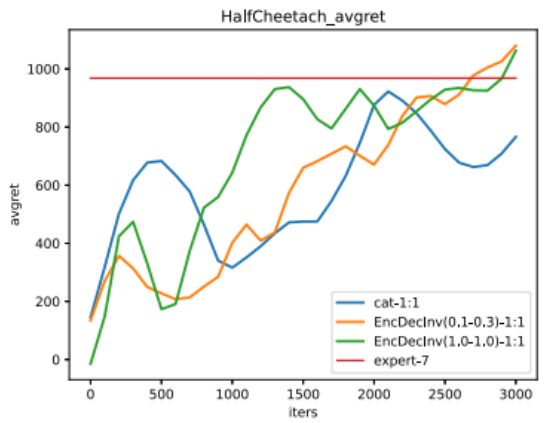
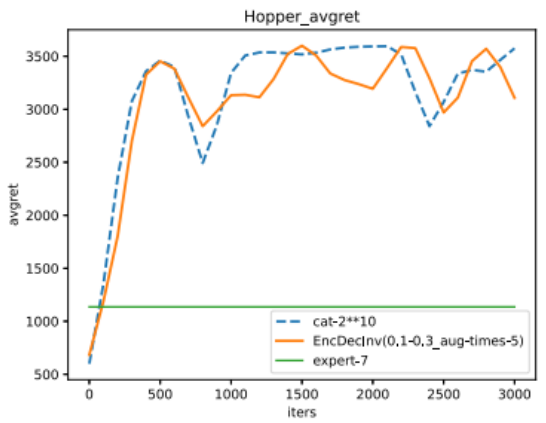
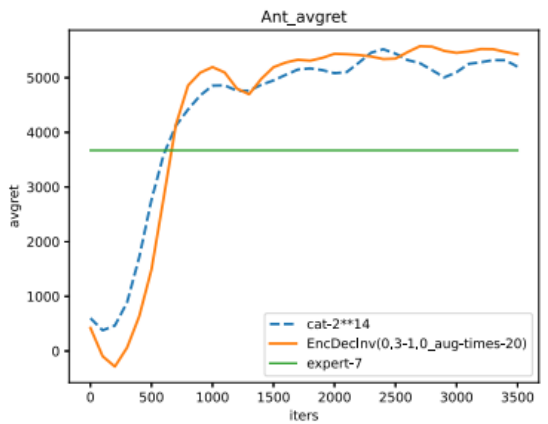


HalfCheetah	Hopper	Ant	door
			

Discussion

- EncDecInv-Cat



HalfCheetah	Hopper	Ant
		

Thanks