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

香港中文大學
The Chinese University of Hong Kong

Zhenghao Peng[§], Quanyi Li[‡], Chunxiao Liu[†], Bolei Zhou[§]

[§] CUHK [†] SenseTime Research [‡] Centre for Perceptual and Interactive Intelligence, CUHK

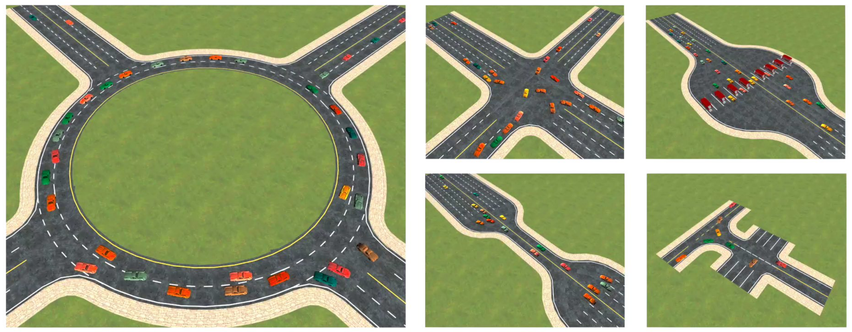
Code and demo video
decisionforce.github.io/CoPO/

Task

Simulating **traffic** via multi-agent reinforcement learning (**MARL**)

- Realistic Crowd Actions
- Safe Driving
- Social Behaviors

Environments powered by:
MetaDrive



Method

Step 1: Local Coordination for each policy

Step 2: Global Coordination to update global LCF

Local Coordination

Individual Reward r_1^I

Mixed by Local Coordination Factor (LCF) Φ

Neighborhood Reward

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

Global Reward

$$r^G = \sum_{i\in\mathcal{E}} r_i$$

Global Coordination

Local Coordination: Update policies to maximize coordinated reward

Coordinated Reward: $r_1^C = \cos(\Phi) r_1^I + \sin(\Phi) r_1^N$, where $\Phi \in [-90°, 90°]$ is LCF

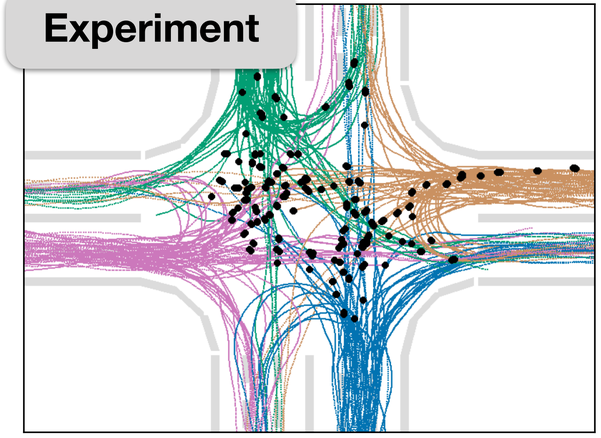
Maximize this reward via PPO loss

Global Coordination: Adjust LCF to maximize global reward

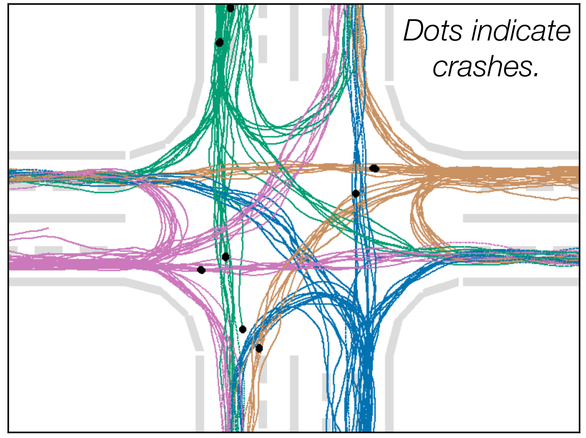
Meta-gradient to update Φ: $\nabla_\Phi J^G(\theta^{new}) = \nabla_{\theta^{new}} J^G(\theta^{new})\, \nabla_\Phi \theta^{new}$

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

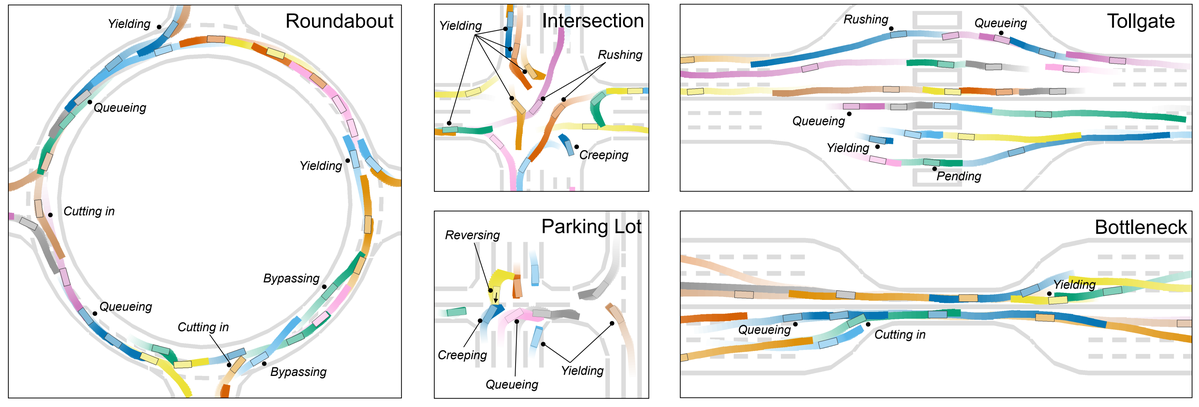
Experiment



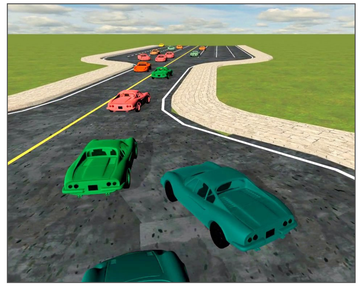
Independent PPO



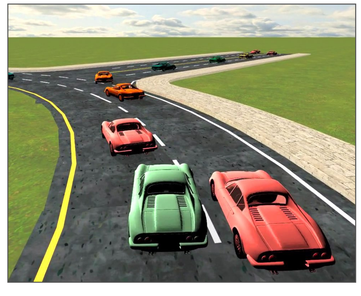
CoPO (Ours)



Detailed Behaviors in CoPO Populations



Cooperation: Yielding



Competition: Cutting in