Efficient Learning of Safe Driving Policy via Human-AI Copilot Optimization

Quanyi Li[1]*, Zhenghao Peng[2]*, Bolei Zhou[3]

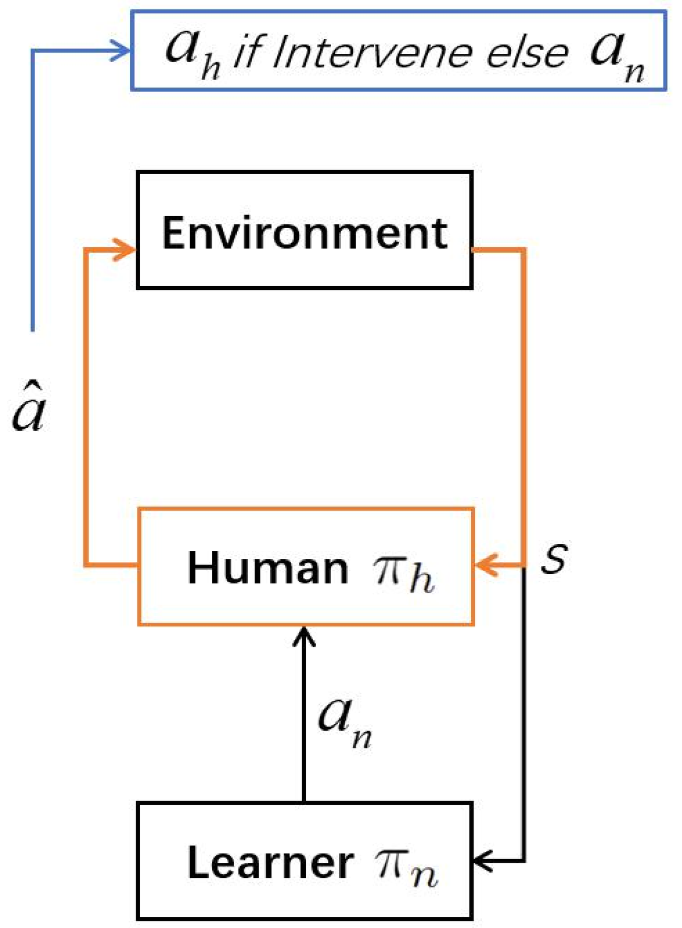
[1]Centre for Perceptual and Interactive Intelligence, [2]The Chinese University of Hong Kong, [3]University of California, Los Angeles

* indicates joint first author

香港中文大學 The Chinese University of Hong Kong

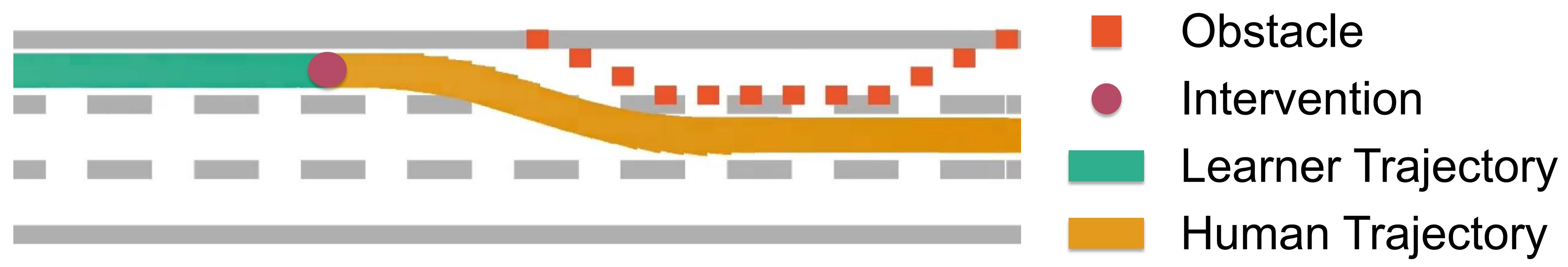
UCLA

Human-in-the-loop RL



$a_h \sim \pi_h$ *Human policy* $a_n \sim \pi_n$ *Learner policy* $\hat{a}$ *Behavior action*

We authorize human expert to take over or intervene, when Reinforcement Learning (RL) agents are in training. This paradigm is referred as **Human-in-the loop RL.**



Copilot trajectory

When training with the learner policy, human's duties are:

1. **Protecting** the agent as a guardian
2. **Teaching** the agent by providing demonstration

Human-AI-Copilot Optimization (HACO)

1. Learning from Demonstration

HACO learns from human-provided demonstrations by applying CQL loss to train proxy value function:

$$\min_{\phi} \mathop{\mathbb{E}}_{\mathcal{B}}[I(s,a_n)(Q(s,a_n;\phi) - Q(s,a_h;\phi))]$$

which is updated through the TD-target.

2. Intervention Minimization

To minimize intervention, HACO additionally learns a intervention cost value function to estimate expected accumulative intervention cost:

$$Q^I(s,a_n) = C(s,a_n) + \gamma \mathop{\mathbb{E}}_{a' \sim \pi_n(\cdot|s')}[Q^I(s',a')]$$

The intervention cost is calculated by cosine similarity:

$$C(s,a_n) = 1 - \frac{{a_n}^{\mathsf{T}} a_h}{||a_n|| \, ||a_h||}, \quad a_h \sim \pi_h(\cdot|s)$$

3. Policy Optimization

The policy optimization goal is to maximize proxy value and minimize the intervention cost:

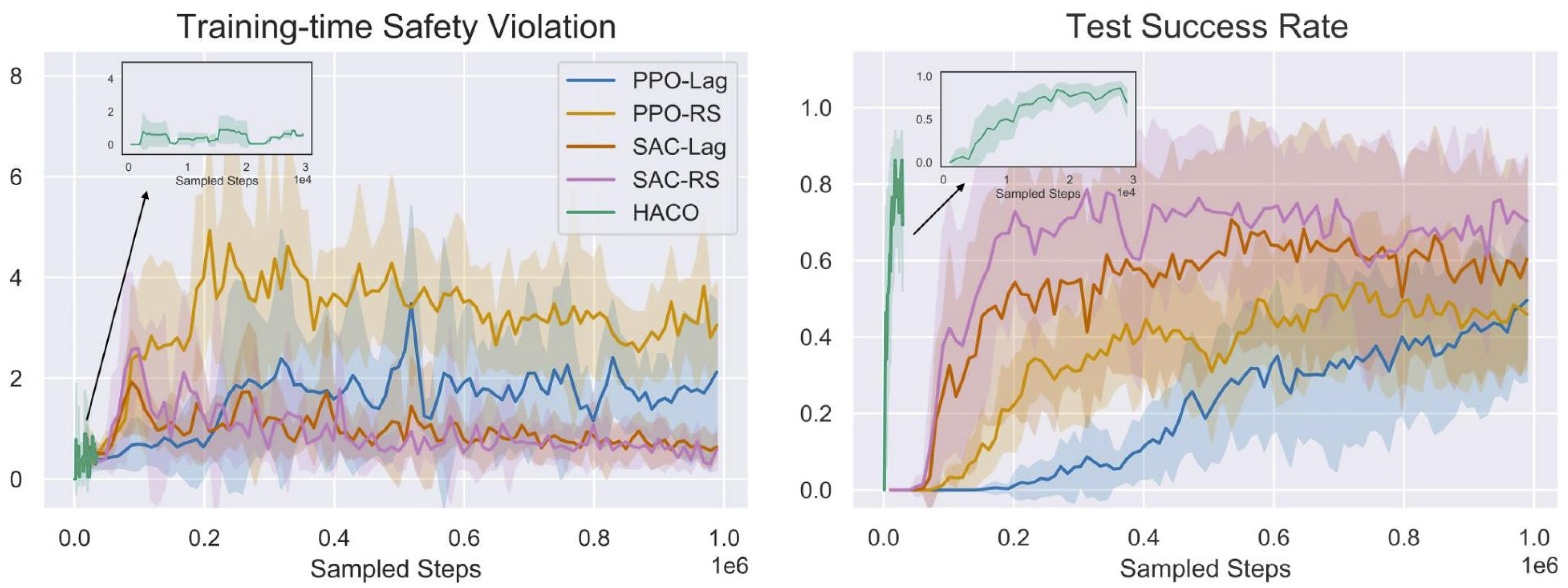
$$\max_{\theta} \mathop{\mathbb{E}}_{\mathcal{B}}[Q(s,a_n) - Q^I(s,a_n)], \quad a_n \sim \pi_n(\cdot|s;\theta)$$

Experiment Results

MetaDrive Safe RL Environment

We evaluate HACO on safe RL suite of MetaDrive:

Comparison with RL baselines



Category	Method	Total Training Safety Violation	Training Data Usage	Test Return	Test Safety Violation	Test Success Rate
RL	SAC-RS	2.76K ± 0.95K	1M	**386.77** ±35.1	0.73 ±1.18	0.82 ±0.18
	PPO-RS	24.34K ±3.56K	1M	335.39 ±12.41	3.41 ±1.11	0.69±0.08
Safe RL	SAC-Lag	1.84K ± 0.49K	1M	351.96 ±101.88	**0.72** ±0.49	0.73 ±0.29
	PPO-Lag	11.64K ± 4.16K	1M	299.99 ±49.46	1.18 ±0.83	0.51 ±0.17
	CPO	4.36K ±2.22K	1M	194.06 ±108.86	1.71 ±1.02	0.21 ±0.29
Ours	**HACO**	**30.14** ± 11.36	**30K***	349.25 ± 11.45	0.79 ± 0.31	**0.83** ± 0.04

Comparison with IL/Offline RL, CARLA Experiments and videos is availabe at: https://decisionforce.github.io/HACO/