Predictive Preference Learning from Human Interventions

Haoyuan Cai, Zhenghao Peng, Bolei Zhou

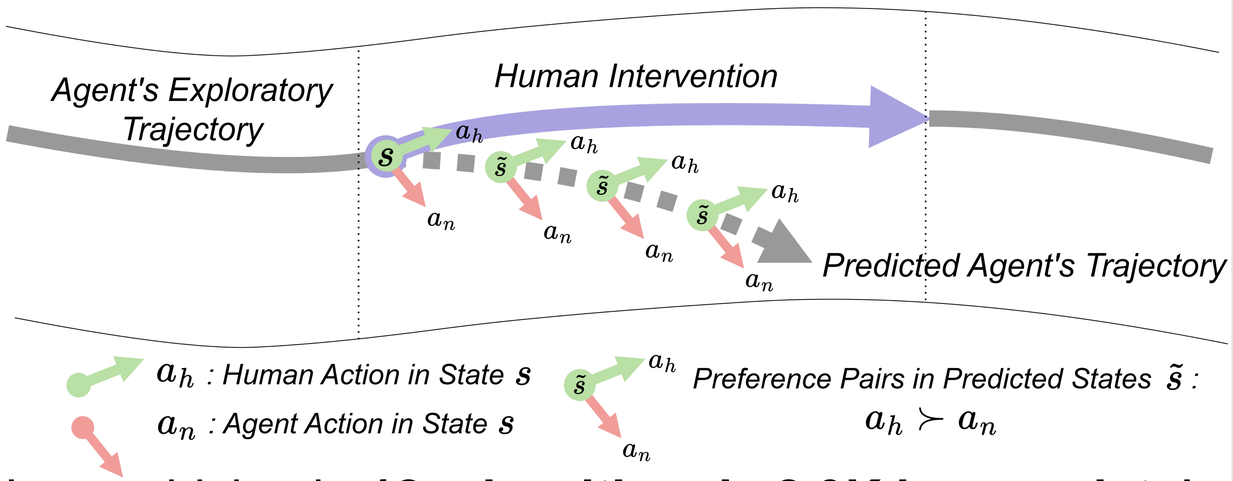
Website Paper

Predictive Preference Learning: Motivation

- Traditional IL: **1 Intervention = 1 BC Data**

Human Foresees Failure | Human Demonstrates | Behavior Cloning Learner

- Our PPL: **1 Intervention = Multiple Preference Pairs**

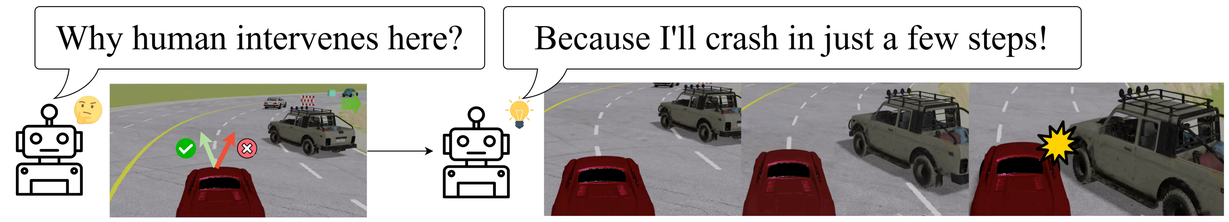


- Learn driving in **12 min with only 2.9K human data!**

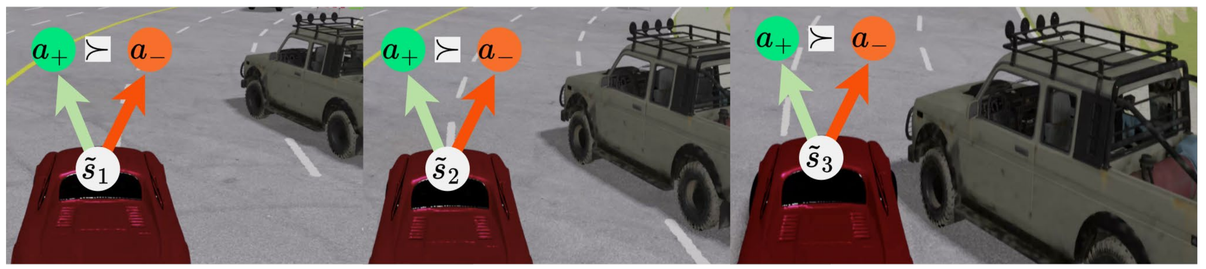
Key Idea: Online RLHF in Predicted Future States

- Incorporate human expertise in predicting future failures
- Construct multiple preference pairs for each intervention
- Online preference learning in future rollouts with DPO loss

1. Human Interventions Reveal Imminent Failures



2. Interventions Imply Preferences in Future Rollouts



3. DPO-Like Preference Loss Boosts Training Efficiency

Traditional IL: BC Loss

$$\mathcal{L}_{\mathrm{BC}}(\pi) = -\mathop{\mathbb{E}}_{(s,a_+)}\left[\log \pi(a_+ \mid s)\right]$$

Ours: Contrastive Preference Loss

$$\mathcal{L}_{\mathrm{pref}}(\pi) = \mathop{\mathbb{E}}_{(\tilde{s},a^+,a^-)}\left[\log \sigma\left(\beta \log \frac{\pi(a^+ \mid \tilde{s})}{\pi(a^- \mid \tilde{s})}\right)\right]$$

One intervention teaches multiple future states!

Real-Human Experiments: PPL Achieves **2X Sample Efficiency Improvement** Across Multiple Tasks

Driving | Test Success Rate | Wiping | Test Success Rate | Peg Insertion | Test Success Rate

MetaDrive Statistics (10K Total Data)

Methods	Training	Testing		
	Human Data Usage	Success Rate	Episodic Return	Route Completion
Human Expert	-	**0.95**	**349.2**	**0.98**
PVP	4.9K	0.46	267.3	0.71
Thrifty-DAgger	3.2K	0.45	221.5	0.62
PPL (Ours)	**2.9K**	**0.76**	**324.8**	**0.90**