NEURAL INFORMATION PROCESSING SYSTEMS

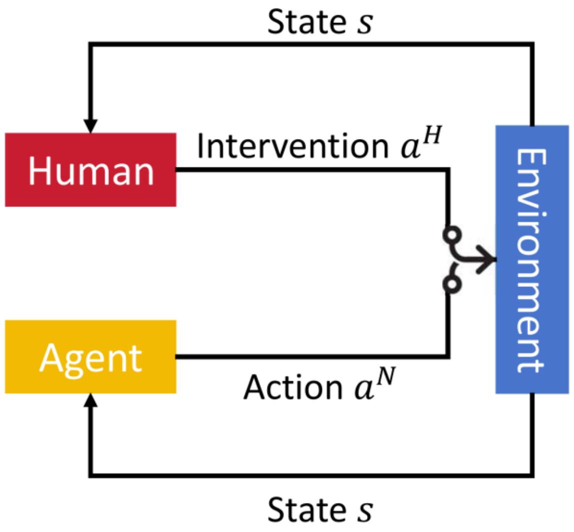
Faithful Dynamic Imitation Learning from Human Intervention with Dynamic Regret Minimization

Bo Ling[1], Zhengyu Gan[1], Wanyuan Wang[1], Guanyu Gao[2], Weiwei Wu[1], Yan Lyu[1]

[1]Southeast University, [2]Nanjing University of Science and Technology

Background

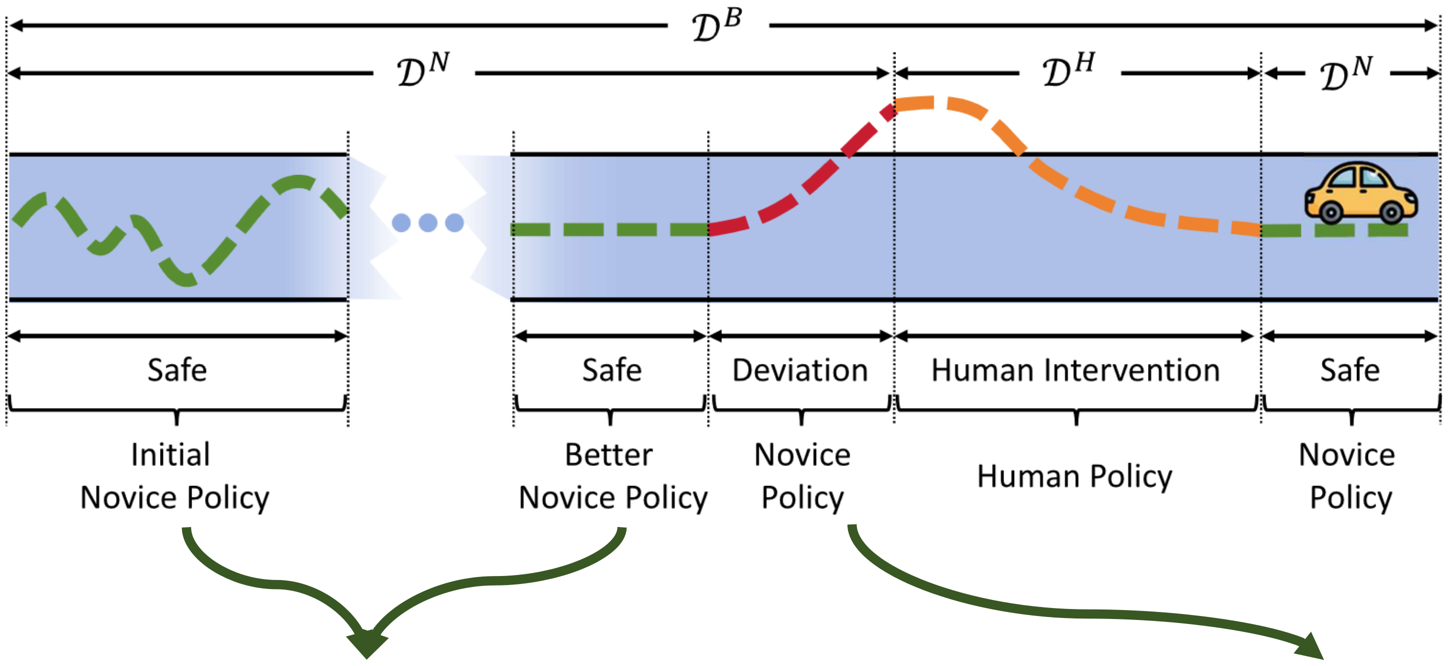
Human-in-the-loop (HIL) imitation learning enables agents to better align with human preferences and directly enhances training-time safety.



(a) Imitation learning from human intervention.

(b) Human supervises the driving agent, and provide real-time corrections when necessary.

Motivation



Challenge 1:
As the policy improves, its trajectory **Distribution Shifts**. Cannot treat them equally.

Challenge 2:
Trajectories can be flawed due to **Human Reaction Delays**.
Cannot be imitation objective.

Key Idea

NEURAL INFORMATION PROCESSING SYSTEMS

- **Solving Distribution shifts:** we claim that learning from human intervention problem is fundamentally an online learning problem.

$$R_D = \sum_{i=1}^{M} \ell\left(\pi_i^N, D_i^B, D_i^H\right) - \sum_{i=1}^{M} \ell\left(\pi_i^*, D_i^B, D_i^H\right)$$

- **Solving human reaction delays:** we focuses on imitating the human expert policy while excluding bias from agent-generated trajectories.

$$D_{\mathrm{KL}}(d^\pi(s,a) \| d^H(s,a)) =$$

$$\mathbb{E}_{(s,a)\sim d^\pi}\left[\log \frac{d^B(s,a)}{d^H(s,a)}\right] + D_{\mathrm{KL}}(d^\pi(s,a) \| d^B(s,a))$$

Method

NEURAL INFORMATION PROCESSING SYSTEMS

Faithful Imitation Objective with Behavior Trajectory

We focus on faithfully imitating only the human expert, while still leveraging novice data for data efficiency.

- Faithful imitation objective:

$$D_{\mathrm{KL}}(d^{\pi}(s,a)\|d^{H}(s,a)) = \mathbb{E}_{(s,a)\sim d^{\pi}}\left[\log\frac{d^{B}(s,a)}{d^{H}(s,a)}\right] + D_{\mathrm{KL}}(d^{\pi}(s,a)\|d^{B}(s,a))$$

- Reformulate the objective into a tractable optimization over a value function :

$$\min_{V}(1-\gamma)\mathbb{E}_{s\sim d_0}V(s) + \mathbb{E}_{(s,a)\sim d^{B}}[f^{*}(\mathcal{T}_{\tilde{r}}V(s,a) - V(s))]$$

Method

Faithful Imitation Objective with Behavior Trajectory

We focus on faithfully imitating only the human expert, while still leveraging novice data for data efficiency.

- Extract the policy using weighted behavior cloning:

$$\omega^{\star}(s,a)=\frac{d^{\star}(s,a)}{d^{B}(s,a)}=\max\left(0,,(f')^{-1}\left(\mathcal{T}_{\tilde{r}}V^{\star}(s,a)-V^{\star}(s)\right)\right)$$

$$\pi^{\star}=\arg\max_{\pi}\mathbb{E}_{(s,a)\sim d^{B}}[\omega^{\star}(s,a)\log\pi(a|s)]$$

Method

Dynamic Imitation Learning with Dynamic Regret Minimization

We employ an ensemble learning framework of FTPL-D+ designed for non-convex online learning to optimize for dynamic regret..

- At each round, policy is updated using FTPL:

$$\pi_{i+1}^{N} = \arg\min_{\pi} \left(\sum_{j=\mu_\tau}^{i} \ell(\pi, \mathcal{D}_j^B, \mathcal{D}_j^H) + \sigma_i^\top \theta_\pi \right)$$

- We use a meta algorithm to adaptively assigns weights to each learner:

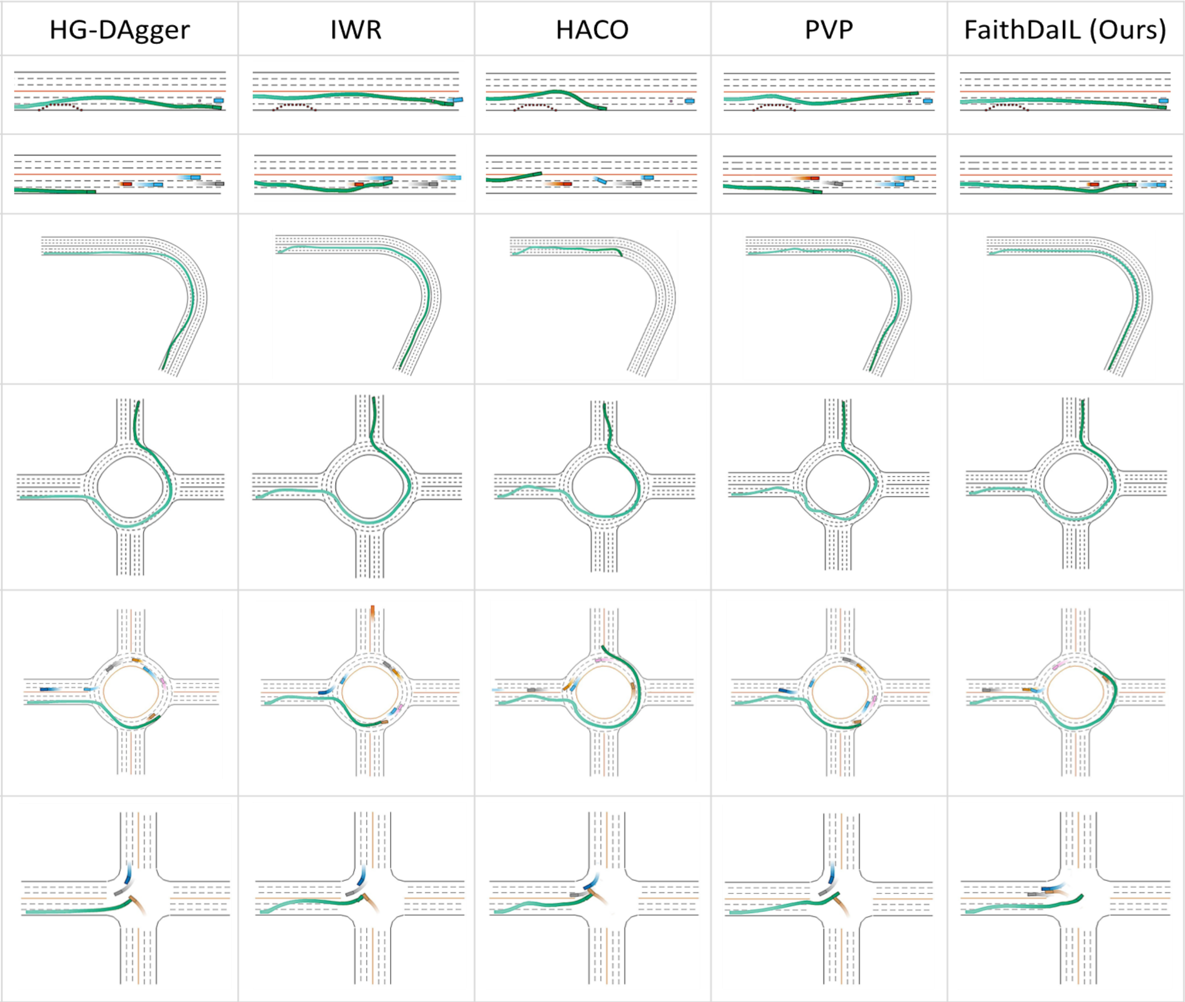
$$\alpha_{i+1,k} = \frac{\alpha_{i,k} e^{-\rho \ell(\pi_{i,k}^N, \mathcal{D}_i^B, \mathcal{D}_i^H)}}{\sum_{k'=1}^{K} \alpha_{i,k'} e^{-\rho \ell(\pi_{i,k'}^N, \mathcal{D}_i^B, \mathcal{D}_i^H)}}$$

Experiments

Method	MetaDrive-Keyboard						CARLA-Wheel			
	Training			Testing			Training		Testing	
	Human Data	Total Data	Total Safety Cost	Episodic Return	Episodic Safety Cost	Success Rate	Human Data	Total Data	Route Comp.	Success Rate
PPO	-	1M	26.4K	327.33	3.31	0.76	-	1M	0.24	0.0
TD3	-	1M	1.90K	317.45	**1.44**	0.58	-	1M	0.11	0.0
Human	-	-	-	374.73	0.39	0.98	-	-	0.99	1.0
BC	30K	-	-	129.60	17.40	0.12	5K	-	0.42	0.20
HG-DAgger	7.5K	30K	143	297.60	7.05	0.59	6.8K	24K	0.64	0.47
IWR	6.1K	30K	112	327.32	9.16	0.75	5.7K	24K	0.69	0.60
HACO	9.9K	30K	76	239.41	4.29	0.26	4.8K	24K	0.52	0.40
PVP	7.0K	30K	54	343.86	2.51	0.85	6.6K	24K	0.92	0.73
FaithDaIL	**4.8K**	30K	55	**354.35**$\pm$3.43	**1.47**$\pm$0.28	**0.91**$\pm$0.04	**4.2K**	24K	**0.95**$\pm$0.02	**0.91**$\pm$0.03

We achieve the best performance in MetaDrive-Keyboard and CARLA-Wheel!

Experiments



Qualitative comparison of agent trajectories.

Method	MetaDrive-Keyboard			CARLA-Wheel	
	Episodic Return	Episodic Safety Cost	Success Rate	Route Comp.	Success Rate
FaithDaIL w/o DRM	346.06 ± 5.42	2.29 ± 0.29	0.87 ± 0.04	0.91 ± 0.03	0.81 ± 0.01
FaithDaIL w/o FOP	350.26 ± 3.57	1.78 ± 0.51	0.89 ± 0.05	0.86 ± 0.04	0.73 ± 0.07
FaithDaIL (Ours)	**354.35** ± 3.43	**1.47** ± 0.28	**0.91** ± 0.04	**0.95** ± 0.02	**0.91** ± 0.03

Ablation Study

- We conducted ablation studies to assess effectiveness of key components.

- Qualitative comparison also shows that our method produces the smoothest and safest trajectories.

Conclusion

- We proposed Faithful Dynamic Imitation Learning framework, **FaithDaIL**, that first formally formulates learning from human intervention as an online non-convex learning problem.
- We propose an unbiased objective for faithful human expert imitation from mixed data, and achieve it by using proxy rewards.
- Extensive experiments to assess the outstanding performance of our method, which closely matching expert performance.

Thank you!

Faithful Dynamic Imitation Learning from Human Intervention with Dynamic Regret Minimization

Bo Ling[1], Zhengyu Gan[1], Wanyuan Wang[1], Guanyu Gao[2], Weiwei Wu[1], Yan Lyu[1]

Code: https://github.com/William-island/FaithDaIL