









Code is available at <u>metadriverse.github.io/pvp</u>

# Learning from Active Human Involvement through Proxy Value Propagation

Zhenghao (Mark) Peng<sup>1</sup>, Wenjie Mo<sup>1</sup>, Chenda Duan<sup>1</sup>, Quanyi Li<sup>2</sup>, Bolei Zhou<sup>1</sup> <sup>1</sup>UCLA, <sup>2</sup>University of Edinburgh



PVP (Ours)



PVP can train a high-level controller for a goalconditioned low-level policy.

## Human Friendly

User Study	HGDAgger	IWR	HACO	P١
Compliance	3.0	4.0	3.0	4
Performance	2.2	3.7	3.3	4
Stress	3.2	4.5	2.3	4

PVP agents make human feels better (*compliance*), stronger (*performance*) and less stressful (*stress*) in shared control. It also makes human takes over less (right table).





MiniGrid-Keyboard





## Safety, Efficiency, Performance

MetaDrive-	Training			Testing	
Keyboard Env	Human Data	Total Data	Total SafetyCost	Episodic SafetyCost	Success Rate
SAC	—	1M	2.76K	0.73	0.82
PPO	—	1M	24.34K	3.41	0.69
TD3	—	1M	1.74K	0.47	0.70
SAC-Lag	—	1M	1.84K	0.72	0.73
PPO-Lag	—	1M	11.64K	1.18	0.51
CPO	—	1M	4.36K	1.71	0.21
HumanDemo	30K	-	39	0.39	0.97
BC	30K	-	—	2.17	0.07
GAIL	30K	2M	25.90K	1.30	0.0
HGDagger	39.0K	51K	56	1.97	0.04
IWR	35.8K	45K	52	1.45	0.46
HACO	19.2K	40K	130	1.64	0.13
PVP (Ours)	14.6K	40K	76.8	0.89	0.85

- Human-in-the-loop makes training much safer, even than Safe RL baselines.
- PVP learns most performant agent.

This work was supported by the National Science Foundation under Grant No. 2235012. The human experiment in this study is approved through the IRB#23-000116 at UCLA.