EFFICIENT LEARNING OF SAFE DRIVING POLICY VIA HUMAN-AI COPILOT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human intervention is an effective way to inject human knowledge into the training loop of reinforcement learning, which can bring fast learning and ensured training safety. Given the very limited budget of human intervention, it remains challenging to design when and how human expert interacts with the learning agent in the training. In this work, we develop a novel human-in-the-loop learning method called Human-AI Copilot Optimization (HACO). To allow the agent's sufficient exploration in the risky environments while ensuring the training safety, the human expert can take over the control and demonstrate how to avoid probably dangerous situations or trivial behaviors. The proposed HACO then effectively utilizes the data both from the trial-and-error exploration and human's partial demonstration to train a high-performing agent. HACO extracts proxy state-action values from partial human demonstration and optimizes the agent to improve the proxy values meanwhile reduce the human interventions. The experiments show that HACO achieves a substantially high sample efficiency in the safe driving benchmark. HACO can train agents to drive in unseen traffic scenarios with a handful of human intervention budget and achieve high safety and generalizability, outperforming both reinforcement learning and imitation learning baselines with a large margin. Code and demo video are included in the supplementary materials.

1 INTRODUCTION

How to effectively inject human knowledge into the learning system is the key challenge to learn reliable autonomous agents for safety-critical applications. In reinforcement learning (RL), researchers can inject their intentions in the carefully designed reward function. The learning agent freely explores the environment to collect the data and learn the desired behaviors induced by the reward function. However, RL methods bear two drawbacks which limit the application of RL in the safety-critical tasks: First, the nature of trial-and-error exploration exposes RL agent to dangerous situations (Saunders et al., 2017). Second, it is difficult to integrate all the intended behaviors to be learned into the reward function. Taking driving vehicle as an example, an ideal policy should obtain a set of skills, such as overtaking, yielding, emergent stopping and negotiation with other vehicles. It is intractable to manually design a reward function that can lead to all those behaviors. To mitigate these two challenges, practitioners enforce the human intentions through imitation learning (IL) where the agent is trained to imitate the expert generated state and action sequences. The demonstration data collected beforehand eliminates the risky interactions between the premature agents and the environment and thus increase the training safety of the agent. High quality expert demonstrations provide direct the optimal solution for the agent to imitate from. However, IL paradigm suffers from the distributional shift problem (Ross & Bagnell, 2010; Ross et al., 2011) while the induced skills are not sufficiently robust with respect to changes in the control task (Camacho & Michie, 1995).

Different from vanilla RL or IL, human-in-the-loop learning is an alternative paradigm to inject human knowledge where a human subject accompanies the agent and oversees its learning process. Depending on the tasks, the human subject can provide guidance such as evaluation Reddy et al. (2018); Warnell et al. (2018), preference (Christiano et al., 2017; Sadigh et al., 2017; Palan et al., 2019), and intervention (Kelly et al., 2019; Zhang & Cho, 2016) to the learning agent. Previous works design the human to either passively advise which action is good (Mandel et al., 2017) or

A. Human-AI Copilot **B**. Human Interface **C**. Copilot Trajectory

Figure 1: **A**. In the proposed human-in-the-loop method, human subject will decide whether to intervene the execution of the learning agent and demonstrate the correct action. Such intervention incurs a cost to the agent such that it learns to minimize the total cost during the training. **B**. The human interface for copilot during the training. The human expert can steer the wheel or press the paddle to start the intervention and then the demonstration data will be recorded. **C.** An illustrative driving trajectory where the green segments are generated by the agent and the blue segments are the demonstrations of the human expert to overcome the dangerous situations.

evaluate the collected trajectories (Christiano et al., 2017; Guan et al., 2020). This kind of passive human involvement still exposes the human-AI system to risks. Some other works require the human to merely intervene the exploration by terminating the episode (Saunders et al., 2018). However, it is not practical to terminate and reset the environment instantly in the real world (Xu et al., 2020). In a human-in-the-loop system, the budget of human intervention is always precious and limited. It is essential to design when and how the human expert involves in the learning process.

In this work, we propose an efficient human-in-the-loop learning method called ***H**uman-AI **C**opilot **O**ptimization (**HACO**)*. The key feature of HACO is that it can learn to minimize the human intervention and adjust the level of automation to the learning agent adaptively during the training. As shown in Figure 1 A, HACO allows the human expert to take over the human-AI system in a proactive manner. If the human decides to intervene the action of the agent, he/she should demonstrate the correct actions to overcome current undesired situations to the learning agent. The human intervention and the partial demonstration are two sources of informative training data. We use offline RL technique to maintain a proxy value function of the human-AI mixed behavior policy even though the agent doesn't have access to the environmental reward during training. To encourage the exploration in the area that does not violating human intention, we also maximize the entropy of action distribution of the agent. Since the human intervention indicates that agent has done something violating human intentions, we set minimizing the expected intervention cost as a learning objective of HACO.

Experiments in the safe driving benchmark show that, with an economic human budget, HACO outperforms RL and IL baselines with a substantial margin in terms of sample efficiency, performance, safety, and generalizability in the unseen testing environment. Thus the human-AI copilot optimization is an efficient learning paradigm to inject human knowledge in an online setting.

2 Related Work

Learning from Demonstration. Though it might be difficult to describe the detailed behaviors of some tasks, it is sometimes more convenient to demonstrate those behaviors. A substantial body of work has explored the problem of learning from demonstration. Passive imitation learning such as behavior cloning (Widrow, 1964; Osa et al., 2018) and recently proposed offline RL methods (Kumar et al., 2020; Fujimoto et al., 2018; Wu et al., 2019) train agent from a out-of-the-shelf data set with safety guaranteed, since no interaction with environment is needed. Apart from the passive methods, Inverse RL methods (Ng et al., 2000; Abbeel & Ng, 2004; Fu et al., 2017; Bloem & Bambos, 2014) learn a reward function first from the human demonstration, and then use it to incentivize the agents to pick up the intended behaviors. More recently, GAIL (Ho & Ermon, 2016) and its variants (Song et al., 2018; Sasaki et al., 2018; Kostrikov et al., 2018) and SQIL (Reddy et al., 2019) are proposed to compare the trajectory similarity between agents and humans. These imitation learning methods

require the agent to interact with the environment. Similar to RL methods, this paradigm exposes the agent to potentially dangerous situations.

Human-in-the-loop Learning ethods. Many works focus on incorporating human in the training loop of conventional RL or IL paradigms. DAgger (Ross et al., 2011) and its extended methods (Kelly et al., 2019; Zhang & Cho, 2016; Hoque et al., 2021) correct the compounding error (Ross & Bagnell, 2010) of behavior cloning by periodically requesting expert to provide more demonstration. Human-Gated DAgger (HG-DAgger) (Kelly et al., 2019) utilizes an expert to intervene exploration and carry the agent to safe states before giving back the control. However, HG-Dagger does not impose constraints to reduce human intervention and does not utilize the data from free exploration. Other forms of human participation includes providing human preferences based evaluative feedback on two behavior sequences generated by the agent (Christiano et al., 2017; Sadigh et al., 2017; Palan et al., 2019; Ibarz et al., 2018; Cui & Niekum, 2018), and more.

Human-AI copilot or shared autonomy is one form of the human-in-the-loop methods. The AI agent and human are working together simultaneously to achieve a common goal. By giving human guidance and feedback at run-time instantly, the explorable state and action spaces can be greatly narrowed down (Saunders et al., 2018). The learning goal can further match the task objective by providing extra human feedback combined with reward function (Reddy et al., 2018; Warnell et al., 2018; Wu et al., 2021; Cederborg et al., 2015; Arumugam et al., 2019). Human-AI copilot is helpful and practical when applying RL to real world tasks where it is critical to learn the policy under safety constraints (Garcıa & Fernández, 2015; Amodei et al., 2016; Bharadhwaj et al., 2020).

A major challenge of human-in-the-loop methods is that frequently querying the human result in tremendous cognitive cost (Zhang et al., 2021), which might lead to incorrect feedback and damage the training. A feasible solution is to learn an intervention function that imitates human intervention signals (Kelly et al., 2019; Zhang & Cho, 2016; Saunders et al., 2017; Abel et al., 2017). The function learns to stop the catastrophic actions of agents and save the attention of the human during training. In this work we consider the human cognitive cost as one part of the learning objective to minimize.

3 HUMAN-AI COPILOT OPTIMIZATION

For human-in-the-loop learning, it is essential to design when and how to engage human subjects. The major issue is the *cognitive costs* of the human subject. Frequent querying might exhaust the human expert and brings tremendous cognitive costs (Zhang et al., 2021), which cause incorrect or delayed feedback and hinder the training. Thus the proposed pipeline aims to minimize the human intervention cost during the training as well as adaptively reduce the reliance on the expert's demonstration and improve the learning agent's autonomy.

3.1 HUMAN-AI COPILOT TRAINING PARADIGM

We aim to learn an autonomous agent with policy $\pi_n(a_n|s)$ that can make informed action a_n in state s. As shown in Fig. 1, we frame the *human-AI copilot paradigm* that extends standard reinforcement learning diagram by incorporating a human expert. At each step, the human expert oversees current state and decides whether to intervene. If necessary, he/she will execute human action a_h to overwrite the agent's action a_n. Denoting the human policy as π_h, the actual trajectories occurred during training are derived from a shared behavior policy π_b:

$$\pi_b(a|s) = \pi_n(a|s)I(s,a) + \pi_h(a|s)(1 - I(s,a)). \tag{1}$$

The human intervention is represented by a Boolean indicator $I(s, a_n)$. We call the transition sequences during the takeover $\{(s_t, a_{n,t}, a_{h,t}, \hat{a}_t, s_{t+1}), ...\}$ as the *partial demonstration*, wherein the safe action $\hat{a}_t = I(s, a_{n,t})a_{h,t} + (1 - I(s, a_{n,t}))a_{n,t}$ is passed to the environment. The human intervention and partial demonstration will be recorded in the replay buffer $\mathcal{B}$ and fed to the training algorithm.

In the human-AI copilot training, the human is obligated to guide the agent learning and safeguard the learning process by proactively taking over the control if necessary. This paradigm rules out the dispensable states and mitigates the safety concern in free exploration of RL and active imitation learning methods (Ross et al., 2011). Different from previous works with fixed dataset (Bojarski

Algorithm 1: The workflow of HACO during training

Initialize an empty replay buffer $\mathcal{B}$
while *Training is not finished* **do**
  while *Episode is not terminated* **do**
    $a_{n,t} \sim \pi_n(\cdot|s_t)$ Retrieve agent's action
    $I(s_t, a_{n,t}) \leftarrow$ Human expert decides whether to intervene by observing current state s_t
    if $I(s_t, a_{n,t})$ *is True* **then**
      $a_{h,t} \leftarrow \pi_h(\cdot|s_t)$ Retrieve human's action
      Apply $a_{h,t}$ to the environment
    else
      Apply $a_{n,t}$ to the environment
    if $I(s_t, a_{n,t})$ *is True and* $I(s_{t-1}, a_{n,t-1})$ *is False* **then**
      $C_I(s_t, a_{n,t}) \leftarrow$ Compute intervention cost following Eq. 6
    else
      $C_I(s_t, a_{n,t}) \leftarrow 0$ Set intervention cost to zero
    Record $a_{n,t}$, $I(s_t, a_{n,t})$ and $a_{h,t}$ (if $I(s_t, a_{n,t})$) to the buffer $\mathcal{B}$
  Update proxy value Q, intervention value Q_I and policy π according to Eq. 5, Eq. 7, Eq. 8 respectively

et al., 2016; Ho & Ermon, 2016; Reddy et al., 2019; Kumar et al., 2020; Fujimoto et al., 2018; Wu et al., 2019) where no closed loop feedback is accessible, the human-AI copilot training produces *partial demonstrations* that bring the necessary human knowledge of overcoming dangerous situations into the learning. The copilot nature alleviates the distributional shift problem, since the human intervenes after the agent performing suspicious behaviors, so that there is a continuity of the state visitation between the agent and the expert.

In next section, we will introduce how we instantiate the human-AI copilot paradigm with a human-efficient algorithm that can effectively optimize the agent toward safe and high-performing policy.

3.2 Learning Objectives

In this section, we introduce the human-efficient training algorithm, referred as ***H**uman-AI **C**opilot **O**ptimization (**HACO**)*, that can train agents from the human intervention and partial demonstrations.

We form three objectives that fully utilize the human data. (1) Agent should learn a *proxy value function* $Q(s, a)$ which reflects human intentions on how to finish the task. (2) Agent should explore thoroughly to visit the state-action subspace permitted by the human to find a optimal policy that maximizes expected $Q(s, a)$. Concretely, we maximize the action distribution entropy $\mathcal{H}(\pi(\cdot|s))$. (3) Agent should maximize the level of automation and reduce the human intervention. Human intervention is measured by an intervention value function $Q_I(s, a)$. Thus the overall learning objective of HACO becomes:

$$\underset{\pi}{\text{maximize}}\, \mathbb{E}[Q(s, a) + \mathcal{H}(\pi) - Q_I(s, a)]. \tag{2}$$

3.3 Implementation Details

In this section, we discuss the practical implementation of aforementioned objectives.

Proxy value function. HACO follows reward-free setting so we can't estimate the expected state-action value based on a ground-truth reward function from the environment. We instead turn to estimate a proxy value function $Q_\phi(s, a)$ (ϕ is model parameters) that captures the ordinal preference of human expert, which implicitly reflects human intentions. Concretely, we utilize the conservative Q-learning (Kumar et al., 2020) and form the optimization problem of the proxy value function as:

$$\underset{\phi}{\text{minimize}} \underset{(s,\hat{a})\sim\mathcal{B}}{\mathbb{E}} \{I(s, \hat{a})[Q(s, \hat{a}) - Q(s, \hat{a})]\}. \tag{3}$$

The above optimization objective can be interpreted as rewarding the state-action pairs deduced by human and penalizing those by agent. Weighting the importance of visited states and implemented

actions, the proxy value function learns to represent the high-value state-action subspace preferred by the human expert.

Entropy regularization. The proxy value is hard to be back-propagated to preceding states that lead to the high-level state, if the learning agent can not visit those human-preferable subspace frequently during free explorable sampling. To encourage exploration, we adopt the entropy regularization technique in (Haarnoja et al., 2018) and forms auxiliary signal to update the proxy value function apart from Eq. 3:

$$\begin{aligned} y &= \gamma \mathop{\mathbb{E}}_{a' \sim \pi_n(\cdot|s_{t+1})} [Q_{\phi'}(s_{t+1}, a') - \alpha \log \pi_n(a'|s_{t+1})] \\ &\underset{\phi}{\text{minimize}} \mathop{\mathbb{E}}_{(s_t, \hat{a}_t, s_{t+1}) \sim \mathcal{B}} |y - Q_\phi(s_t, \hat{a}_t)|^2, \end{aligned} \tag{4}$$

wherein at a_t is the executed action at state s_t, $Q_{\phi'}$ denotes a delay updated target network of Q_ϕ, γ is the discount factor. Since the environment reward is not accessible to HACO, we remove the reward term in the update target y.

Combining Eq. 3 and Eq. 4, the formal optimization objective of the proxy value function becomes:

$$\underset{\phi}{\text{minimize}} \mathop{\mathbb{E}}_{(s_t, a_n, a_h, \hat{a}, s')} \{|y - Q_\phi(s, \hat{a})|^2 + I(s, a_n)[Q_\phi(s, a_n) - Q_\phi(s, a_h)]\} \tag{5}$$

Reducing human interventions. Directly optimizing the agent policy according to the proxy value function will lead to failure when evaluating the agent without human participation. This is because $Q(s, a)$ represents the proxy value of the mixed behavior policy π_b instead of the learning agent's π_n due to the existence of human intervention. It is possible that the agent learns to deliberately abuse human intervention by always taking actions that violating human intentions, such as driving off the road when near the boundary, which force human to take over and provide demonstration. In this case the level of automation for the agent is low and the human subjects become exhausted to provide demonstrations. Ablation study result in Table 2(c) illustrates this phenomenon.

Thus we need to make more economic use of the human budget and reduce the human interventions over time. We measure the cost of human intervention by a function $C_I(s, a_n)$ and set minimizing it as one of the learning objectives. If the human subject is agile enough to capture the faulty actions of agent instantly, a straightforward form of C_I is a constant $+1$ when human expert issues intervention. However, we find that there usually exists a latency ϵ between human intervention and faulty actions so that the intervention cost is given to the agent at a delayed time step $t + \epsilon$. It is possible that the agent's action $a_{n,t+\epsilon}$ is a correct action that saves the agent itself from dangers but mistakenly marked as faulty action that triggers human intervention. To overcome this issue, we punish the agent action that triggers human intervention in a mild manner by using the cosine similarity between agent's action and human's action as the intervention cost function in the form below:

$$C_I(s, a_n) = 1 - \frac{{a_n}^\mathsf{T} a_h}{||a_n||||a_h||}, \;\; a_h \sim \pi_h(\cdot|s). \tag{6}$$

When agent triggers human intention, it will receive large penalty only when its action is significantly different to the expert action in term of cosine similarity. In the ablation study, we find that using the constant cost raises inferior performance compared to the cosine similarity.

As shown in the Line 11-14 of Algorithm 1, we only yield non-zero intervention cost at the first step of human intervention. This is because the human intervention triggered by the exact action $a_{n,t}$ indicates this action violates the underlying intention of human at this moment. Minimizing the chance of those action will increase the level of automation.

To optimize the policy according to C_I improve the level of automation, we form an additional intervention value function $Q_I(s, a_n)$ to approximate the expected accumulative intervention cost:

$$Q_I(s, a_n) \leftarrow C_I(s, a_n) + \gamma \, \mathbb{E}[Q_I(s', a'_n)]. \tag{7}$$

Learning policy. Using the entropy-regularized proxy value function $Q(s, a)$ as well as the intervention value function $Q_I(s, a)$ in above sections, we form the the policy improvement objective as:

$$\underset{\theta}{\text{maximize}} \mathop{\mathbb{E}}_{s_t \sim \mathcal{B}} [Q(s_t, a_n) - \alpha \log \pi_\theta(a_n|s_t) - Q_I(s_t, a_n)], \;\; a_n \sim \pi_n(\cdot|s_t) \tag{8}$$

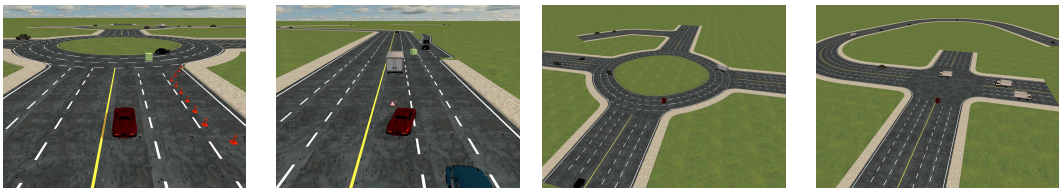

Figure 2: The demonstrations of generated safety environments.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Task. We focus on the driving task in this work. This is because driving is an important decision making problem with a huge social impact, where the safety and training efficiency are critical. Since many researches on autonomous driving employing human in a real vehicle (Bojarski et al., 2016; Kelly et al., 2019), the human safety and human cognitive cost becomes practical challenges that limits the application of learning-based methods in industries. Therefore, the driving task is an ideal benchmark for the human-AI copilot paradigm.

Simulator. Considering the potential risks of employing human subjects in physical experiment, we benchmark the performance of different approaches in the driving simulator. We employ a lightweight driving simulator that preserves the capacity to evaluate the safety and generalizability in unseen environments so it is easy to conduct human-AI copilot experiments. The simulator is implemented based on Panda3D Goslin & Mine (2004) and Bullet Engine that has high efficiency as well as accurate physics-based 3D kinetics. Different to the fixed maps in CARLA (Dosovitskiy et al., 2017), our simulator uses procedural generation to synthesize an unlimited number of driving maps for the split of training and test sets, which is useful to benchmark the generalization capability of different approaches in the context of safe driving. Some generated driving scenes are presented in Fig. 2. The simulator is also extremely efficient and flexible so that we can run the human-AI copilot experiment in real-time. We will release the simulator.

Training Environment. In the simulator, the task for the agent is to steer the target vehicle with low-level control signal, namely acceleration, brake and steering, to reach the predefined destination and receive a *success* flag. The ratio of episodes where agent successfully reaches the destination is called the *success rate*. To increase the difficulty of the task, we scatter obstacles randomly in each driving scene such as movable traffic vehicles, fixed traffic cones, and warning triangles.

The observation contains (1) the current states such as the steering, heading, velocity and relative distance to boundaries *etc.*, (2) the navigation information that guides the vehicle toward the destination, and (3) the surrounding information encoded by a vector of 240 Lidar-like distance measures of the nearby vehicles. Vehicles are controlled by continuous steering and acceleration signals.

Though HACO does not receive environmental reward during training, we provide reward function to train baseline methods and evaluate HACO in test time. The reward function only contains a dense driving reward, speed reward and a sparse terminal reward. The dense reward is the longitudinal movement toward destination. The speed reward is vehicle's current speed at step t, and the sparse reward $+20$ is given when the agent arrives the destination.

Each collision to the traffic vehicles or obstacles yields $+1$ environmental cost. Note that HACO can not access this cost during the training. This cost is used to train safe RL baselines as well as for testing. The episodic cost in test time is the measurement on the safety of a policy, which is independent of the participation of the human subject during training.

We invite the human expert to supervise the real-time exploration of the learning agent with hands on the steering wheel, as shown in the Fig. 1B. When dangerous situation is going to happen, the human takes over the vehicle by pressing the paddle besides the wheel and start controlling the vehicle by steering the wheel and stepping the pedals.

Split of training and test sets. Different from the conventional RL setting where the agent is trained and tested in the same fixed environment, we focus on evaluating the generalization performance

Table 1: The performance of different approaches in safe driving benchmark.

Category	Method	Total Training Cost	Training Data Usage	Test Return	Test Cost	Test Success Rate
Expert	*Human*	-	-	*358.19* $\pm$*86.00*	*0.16*$\pm$*0.61*	*0.98*
RL	SAC-RS	2.76K $\pm$ 0.95K	1M	**386.77** $\pm$35.1	0.73 $\pm$1.18	0.82 $\pm$0.18
	PPO-RS	24.34K $\pm$3.56K	1M	335.39 $\pm$12.41	3.41 $\pm$1.11	0.69$\pm$0.08
Safe RL	SAC-Lag	1.84K $\pm$ 0.49K	1M	351.96 $\pm$101.88	**0.72** $\pm$0.49	0.73 $\pm$0.29
	PPO-Lag	11.64K $\pm$ 4.16K	1M	299.99 $\pm$49.46	1.18 $\pm$0.83	0.51 $\pm$0.17
	CPO	4.36K $\pm$2.22K	1M	194.06 $\pm$108.86	1.71 $\pm$1.02	0.21 $\pm$0.29
Offline RL	CQL	-	36K	156.4 $\pm$31.94	6.82 $\pm$5.1	0.11 $\pm$0.07
IL	BC	-	36K	101.63 $\pm$16.06	1.00 $\pm$0.45	0.01 $\pm$0.03
	GAIL	3.70K $\pm$2.43K	36K	136.08 $\pm$18.73	4.42 $\pm$1.89	0.13 $\pm$0.03
Ours	**HACO**	**30.14** $\pm$ 11.36	**30K***	349.25 $\pm$ 11.45	0.79 $\pm$ 0.31	**0.83** $\pm$ 0.04

* In 8,316 $\pm$ 497.90 steps out of the total 30K steps the human expert intervenes and overwrite the agent's actions. The whole training takes about 50 minutes.

through testing the trained agents in separated test environment. We split the driving scenes into the training set and test set with 50 different scenes in each set. At the beginning of each episode, a scene in the training or test set is randomly selected. After each training iteration, we roll out the learning agent *without guardian* in the test environments and record the percentage of successful episodes over multiple evaluation episodes, called *success rate*. Besides, we also record the episodic cost given by the environment and present it in following tables and figures.

Implementation details. We conduct experiments on the driving simulator and implement algorithms using RLLib (Liang et al., 2018), an efficient distributed learning system. When training the baselines, we host 8 concurrent trials in an Nvidia GeForce RTX 2080 Ti GPU. Each trial consumes 2 CPUs with 8 parallel rollout workers. All baseline experiments are repeated 5 times with different random seeds. The HACO experiments are conducted by three human expert. The main experiments of HACO repeat 3 times. Information about other hyper-parameters is given in Appendix.

4.2 Baseline Comparison

We compare our method to RL methods and Safe RL methods which inject the human intention and constraint through the pre-defined reward function and cost function. We test native RL methods, PPO (Schulman et al., 2017) and SAC (Haarnoja et al., 2018), with cost added to the reward as auxiliary negative reward, called reward shaping (RS). Three common safe RL baselines, namely Constraint Policy Optimization (CPO) (Achiam et al., 2017), PPO-Lagrangian (Stooke et al., 2020), SAC-Lagrangian (Ha et al., 2020), are evaluated.

Apart from the RL methods, we also generate a human demonstration dataset containing one-hour expert's demonstrations where there are about 36K transitions, in the training environments. For the high-quality demonstration data in the dataset, the success rate of the episodes reaches 98% and the cost is down to 0.16. Using this dataset, we evaluate passive IL method Behavior Cloning, active IL method GAIL (Ho & Ermon, 2016) and offline RL method CQL (Kumar et al., 2020).

Training-time Safety. The training-time safety is measured by the total training cost, the total number of critical failures occurring in the training. Note that the environmental cost here is different from the human intervention cost in HACO. As illustrated in Table 1 and Fig. 3, HACO achieves high success in training time safety. Under the protection from the human expert, HACO yields only 30.14 total training cost in the whole training process, an order of magnitude better than the other baselines, even though HACO does not access the environmental cost. The most competitive RL baseline SAC-RS, which achieves similar test success rate, causes averagely 2767.77 training safety violations which are much higher than HACO. The active IL method GAIL also has a significantly higher training cost than HACO and the performance is unsatisfactory.

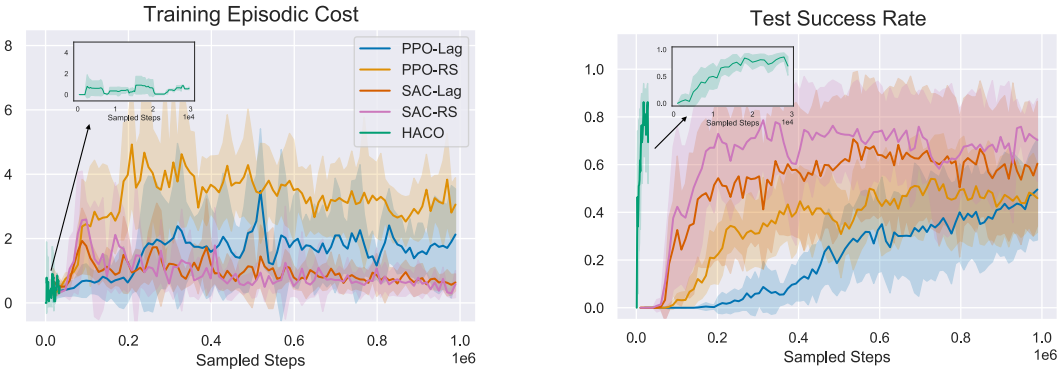

Figure 3: The performance of RL methods and HACO. HACO achieves superior training time safety and high success rate in test time.

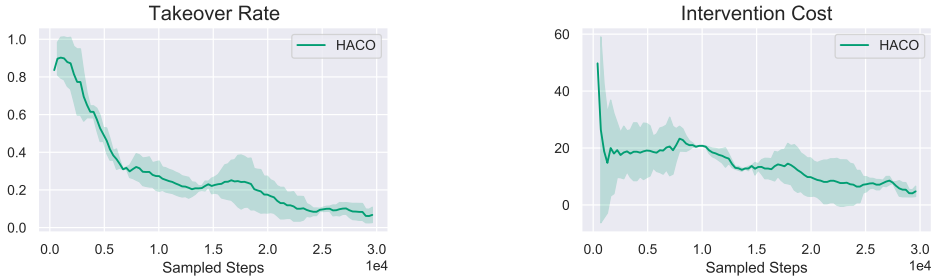

Figure 4: The takeover rate is the proportion of the human takeover steps in an episode. In the course of training, the takeover rate and the episodic intervention cost gradually reduces where the agent obtains higher level of autonomy from the human expert.

We find that the reward shaping technique is inferior compared to the Lagrangian method, both for SAC and PPO variants. PPO costs more safety variation than SAC, probably due to the relative slower convergence speed.

Sample Efficiency and Human Cognitive Cost. Though the human-AI system is protected so well by the human, HACO still achieves superior sample efficiency with limited data usage. HACO achieves 0.83 test success rate by merely interacting with the environment in the 30K steps, wherein only averagely 8,316 steps the human provides safe actions as demonstration. On the contrary, all the learning from demonstration methods fails with the small dataset containing 36K transitions, while RL methods requires much more environment interactions. In Fig. 3, we find that HACO is an order of magnitude more efficient than RL baselines, even though HACO contains shared ideas with RL such as the value-based RL and entropy regularization.

Learning Dynamics. Compared to the method like DAgger where human needs to be focused during data sampling, the intervention minimization mechanism in HACO reduces human cognitive cost. As shown in Fig. 4, the takeover rate gradually decreases in the course of learning. The curve of episodic intervention cost suggests that the human intervention frequency becomes lower and the similarity between agent's action and human's action increases. Only approximately 30 % partial human demonstration, namely 8K steps, exist among the total 30K environmental interactions (See Table1). The precious human budget is economically used and the level of automation is increased.

4.3 Ablation Study

Takeover Policy Analysis. We request the human subject to try two intervention strategies. The first is to take over in a low frequency and produce a long demonstration at each intervention. In this way the intervention cost becomes sparse. The other strategy is to intervene more frequently and provide fragmented demonstration. In Table 2(a), the experiment shows that the proposed HACO

Table 2: The test performance when ablating components in HACO. All experiments here are conducted by human subject. Some experiments only repeat once due to limited human budget.

Experiment	Test Return	Test Cost	Test Success Rate
(a) Human intervenes less frequently	220.03	0.867	0.397
(b) W/o cosine similarity intervention cost	23.23	0.42	0.00
(c) W/o intervention minimization	84.77	1.07	0.00
HACO	349.25 $\pm$ 11.45	0.79 $\pm$ 0.31	0.83 $\pm$ 0.04

works better with dense human intervention signals. Agent trained with long trajectories achieves inferior success rate and mean episodic reward than agents trained with more intervention signals.

Cosine Similarity Cost Function As shown in Table 2(b), we replace the intervention cost function in Eq. 6 to a constant value $+1$ if human intervention happens. We find the agent learns to stay in the spawn points and do not move at all in test time. As discussed in Sec. 3.3, it is possible that the human intervenes in incorrect timing. This makes agent fail to know how to drive correctly. Using the negative cosine similarity to measure the divergence between agent and human's action alleviates this phenomenon since if the agent provides action that adheres human intention, then the human intervention penalty is reduced.

Intervention Minimization. As shown in Table 2(c), when removing the intervention minimization mechanism, the agent learns to drive directly toward the boundary of the road. This causes consistently out-of-the-road failures, resulting in the zero success rate and near 1 episodic cost. This result shows the importance of the intervention minimization.

5 Conclusion

We develop an efficient human-in-the-loop learning method, ***H**uman-AI **C**opilot **O**ptimization (**HACO**)*, which can train agents from the human intervention and partial demonstrations. The method incorporates the human expert in the interaction between agent and environment to ensure safe and efficient exploration. The experiments on safe driving show that the proposed method achieves superior training-time safety, outperforming RL and IL baselines. Besides, it shows a high sample efficiency for rapid learning. The constraint optimization technique is used to prevent the agent from excessively exploiting the human expert, which also decreases the takeover frequency and valuable human budget.

One limitation of this work is that the trained agents behave conservatively when comparing to RL baselines. Aiming to ensure the training time safety of the copilot system, human expert typically drives slowly to rescue the vehicle. This makes the agent tend to drive slowly and exhibit behaviors such as frequent yielding in the intersection. In future work we will explore the possibility of learning more sophisticated skills.

6 Ethics Statement

The proposed Human-AI Copilot Optimization algorithm aims at developing a new human-friendly human-in-the-loop training framework. We successfully increase the level of automation after very human-efficient training. We believe this work has great positive social impact which advances the development of more intelligent AI systems that costs less human burdens.

We employs human subjects to participate in the experiments. Human subjects can stop the experiment if any discomforts happens. Since we test different approaches in the driving simulator, no injury would happens. The human subjects earn hourly salary more than average in our community. Each experiment lasts near one hour. Human participants will rest at least three hours after one experiment. During training and data processing, no personal information is revealed in the collected dataset or the trained agents.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079*, 2017.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 22–31. JMLR. org, 2017.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. *arXiv preprint arXiv:2010.14497*, 2020.

Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE conference on decision and control*, pp. 4911–4916. IEEE, 2014.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Rui Camacho and Donald Michie. Behavioral cloning a correction. *AI Magazine*, 16(2):92–92, 1995.

Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Paul Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

Yuchen Cui and Scott Niekum. Active reward learning from critiques. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6907–6914. IEEE, 2018.

Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: an open urban driving simulator. *CoRR*, abs/1711.03938, 2017. URL `http://arxiv.org/abs/1711.03938`.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. *arXiv preprint arXiv:1812.02900*, 2018.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Mike Goslin and Mark R Mine. The panda3d graphics engine. *Computer*, 37(10):112–114, 2004.

Lin Guan, Mudit Verma, Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati. Explanation augmented feedback in human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2006.14804*, 2020.

Sehoon Ha, Peng Xu, Zhenyu Tan, Sergey Levine, and Jie Tan. Learning to walk in the real world with minimal human effort, 2020.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *arXiv preprint arXiv:1606.03476*, 2016.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning, 2021.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8022–8034, 2018.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8077–8083. IEEE, 2019.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Rllib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*, pp. 3053–3062, 2018.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. Where to add actions in human-in-the-loop reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.

Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*, 2019.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Shared autonomy via deep reinforcement learning. *Robotics: Science and Systems*, 2018.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*, 2019.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. 2017.

Fumihiro Sasaki, Tetsuya Yohira, and Atsuo Kawaguchi. Sample efficient imitation learning for continuous control. In *International conference on learning representations*, 2018.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2067–2069. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. *arXiv preprint arXiv:1807.09936*, 2018.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pp. 9133–9143. PMLR, 2020.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Bernard Widrow. Pattern recognition and adaptive control. *IEEE Transactions on Applications and Industry*, 83(74):269–277, 1964.

Jingda Wu, Zhiyu Huang, Chao Huang, Zhongxu Hu, Peng Hang, Yang Xing, and Chen Lv. Human-in-the-loop deep reinforcement learning with application to autonomous driving. *arXiv preprint arXiv:2104.07246*, 2021.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Kelvin Xu, Siddharth Verma, Chelsea Finn, and Sergey Levine. Continual learning of control primitives: Skill discovery via reset-games. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450*, 2016.

Ruohan Zhang, Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems*, 35(2):1–39, 2021.

APPENDIX

We provide a **demonstrative video** showing the complete training process of HACO and the behavior of trained agents. Please visit this link: https://drive.google.com/file/d/1g6Fr90vc-uJhxTlT0Y8eIFFKaz3EAcX-/view?usp=sharing.

We also provide a theorem describes the **training safety guarantee** from the introduction of human expert. Concretely, we prove an upper bound of the expected cumulative probability of failure of the behavior policy. Please refer to the theorem and proof at: https://drive.google.com/file/d/1qalYrwPSo0uf9YYrCtwJdYfgnNYP8x3b/view?usp=sharing.

The **code** of HACO and baselines is provided at: https://drive.google.com/file/d/1QgqHUEB93lS6bn7hY5WSAv_72_wnpPqC/view?usp=sharing.

A HYPER-PARAMETERS

Table 3: HACO

Hyper-parameter	Value
Discounted Factor γ	0.99
τ for Target Network Update	0.005
Learning Rate	0.0001
Environmental Horizon T	1000
Steps before Learning Start	100
Steps per Iteration	100
Train Batch Size	1024
CQL Loss Temperature	10.0
Target Entropy	2.0

Table 4: PPO/PPO-Lag

Hyper-parameter	Value
KL Coefficient	0.2
λ for GAE Schulman et al. (2018)	0.95
Discounted Factor γ	0.99
Number of SGD epochs	20
Train Batch Size	4000
SGD mini batch size	100
Learning Rate	0.00005
Clip Parameter ϵ	0.2
Cost Limit for PPO-Lag	1

Table 5: SAC/SAC-Lag/CQL

Hyper-parameter	Value
Discounted Factor γ	0.99
τ for target network update	0.005
Learning Rate	0.0001
Environmental horizon T	1500
Steps before Learning start	10000
Cost Limit for SAC-Lag	1
BC iterations for CQL	200000
CQL Loss Temperature β	5
Min Q Weight Multiplier	0.2

Table 6: BC

Hyper-parameter	Value
Dataset Size	36,000
SGD Batch Size	32
SGD Epoch	200000
Learning Rate	0.0001

Table 7: CPO

Hyper-parameter	Value
KL Coefficient	0.2
λ for GAE Schulman et al. (2018)	0.95
Discounted Factor γ	0.99
Number of SGD epochs	20
Train Batch Size	8000
SGD mini batch size	100
Learning Rate	0.00005
Clip Parameter ϵ	0.2
Cost Limit	1

Table 8: GAIL

Hyper-parameter	Value
Dataset Size	36,000
SGD Batch Size	64
Sample Batch Size	12800
Generator Learning Rate	0.0001
Discriminator Learning Rate	0.005
Generator Optimization Epoch	5
Discriminator Optimization Epoch	2000
Clip Parameter ϵ	0.2