Safe Driving via Expert Guided Policy Optimization
The Appendix

Zhenghao Peng[†*], **Quanyi Li**[§*], **Chunxiao Liu**[‡], **Bolei Zhou**[†]
[†] The Chinese University of Hong Kong, [‡] SenseTime Research,
[§] Centre for Perceptual and Interactive Intelligence

A Rationale on the Evaluation

Evaluation on driving simulator. The major focus of this work is the safety. However, in the domain of autonomous driving, evaluating systems' safety in real robot is costly and even unavailable. Thus we benchmark the safety performance of baseline methods and the proposed EGPO method in driving simulator. Using driving simulator to prototype allows us to focus on the algorithmic part of the problem. The exact reproducible environments and vehicles allow safe and effective evaluation of different safe training algorithms. In this work, we conduct experiments on the driving simulator MetaDrive [1] instead of CARLA because we want to evaluate the generalization of the different safe exploration methods. Different to the fixed maps in CARLA, MetaDrive 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 reinforcement learning in the context of safe driving. MetaDrive also supports scattering diverse obstacles in the driving scenes such as fixed or movable traffic vehicles, traffic cones and warning triangles. The simulator is also extremely efficient and flexible. The above unique features of MetaDrive driving simulator enables us to develop new algorithms and benchmark different approaches. We intend to validate and extend the proposed method with real data in the following two ways.

Extension to the human-in-the-loop framework. We are extending the proposed method to replace the pre-trained policy in the guardian with real human. A preliminary experiment is provided in Appendix B. We invite human expert to supervise the real-time exploration of the learning agent with hands on the steering wheel. When dangerous situation is going to happen, the human guardian takes over the vehicle by pressing the paddle and steering the wheel. Such trajectories will be explicitly marked as "intervention occurred". EGPO can incorporate the data generated by either a virtual policy or human being. Therefore, EGPO can be applied to such human-in-the-loop framework directly. We are working on further improvement of the sample efficiency of the proposed method to accommodate the limited budget of human intervention.

Extension to the mobile robot platform. We design the workflow to immigrate EGPO to real robot in future work. Our system includes several components: (1) a computer controlling the vehicle remotely and training the agent with EGPO; (2) a human expert steering vehicle and watching the images from camera on the robot; and (3) an UGV robot simulating a full-scale vehicle (as shown in Fig. 1). During exploration, the on-board processor receives the low-level actions from human and queries the policy network for agent's action. Then the on-board processor executes the action on the robot and receives new sensory data. The data is recorded and used to train the agent. EGPO algorithm can train such real-world robot based on the above workflow.

To summarize, the essential ideas proposed in the work, such as *expert as guardian*, *intervention minimization*, *learning from partial demonstration*, are sufficiently evaluated through the safe driving experiments in the driving simulator. With on-going efforts, we are validating our method with real data from human-in-the-loop framework and extending our method for the real-world mobile robot experiments.

5th Conference on Robot Learning (CoRL 2021), London, UK.

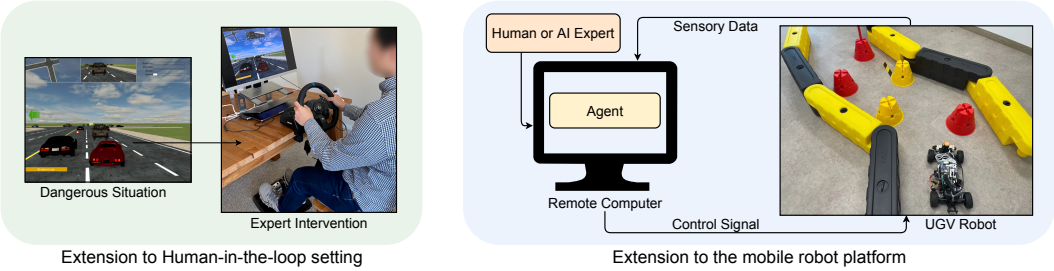

Figure 1: We extend the proposed EGPO to Human-in-the-loop setting and real mobile robot platform.

B Preliminary Human-in-the-loop Experiment

To further demonstrate the capacity of the proposed framework, in this experiment, a human staff supervises the learning progress of the agent in a single training map. The expert takes over once he/she feels necessary by pressing the paddle in the wheel. At this time, an intervention cost is yielded and the action sequences of the expert are recorded and fed into the replay buffer.

Table 1 captures the result of this experiment. We find that EGPO with a human expert can achieve a high success rate in merely 15,000 environmental steps, while SAC-Lagrangian (with PID update) takes 185,000 steps to achieve similar results. We also ask the expert to generate 15,000 steps demonstrations (note that in EGPO experiment, only a small part of the 15,000 steps is given by the expert) and train a BC agent based on those demonstrations. However, BC fails to learn a satisfactory policy. This experiment shows the applicability of the proposed framework even with human experts.

Table 1: Human-in-the-loop experiment results

Experiment	Total Training Cost	Test Reward	Test Cost	Test Success Rate
Human expert **(20 episodes)**	-	219.50 $\pm$39.53	0.30 $\pm$0.550	0.95
Behavior Cloning	-	33.21 $\pm$5.46	0.990 $\pm$0.030	0.000 $\pm$0.000
PPO-Lagrangian **(200K steps)**	285.1	197.76 $\pm$7.90	0.427 $\pm$0.043	0.598 $\pm$0.029
SAC-Lagrangian **(185K steps)**	452.5	221.381 $\pm$7.90	0.060 $\pm$0.049	0.940 $\pm$0.049
EGPO (with human expert) **(15K steps)**	6.14	221.058 $\pm$32.562	0.120 $\pm$0.325	0.900 $\pm$0.300

C Proof of Main Theorem

In this section, we derive the upper bound of the discounted probability of failure of EGPO, showing that we can bound the training safety with the guardian.

Notations. Before starting, we firstly recap and describe the notations. The switch function used in this work is:

$$\mathcal{T}(s,a,\mathcal{E}) = (\hat{a},\hat{c}) = \begin{cases} (a,0), & \text{if } a \in \mathcal{A}_\eta(s) \\ (a^E \sim \mathcal{E}(\cdot|s),1), & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, at a given state, we can split the action space into two parts: where intervention will happen or will not happen if we sample action in it. We denote the *confident action space* as $\mathcal{A}_\eta(s) = \{a : \mathcal{E}(a|s) \geq \eta\}$, which is related to the expert as well as η. We also define the ground-truth indicator I denoting whether the action will lead to unsafe state. This unsafe state is determined by the environment and is not revealed to learning algorithm:

$$I(s,a) = \begin{cases} 1, & \text{if } s' = \mathcal{P}(s'|s,a) \text{ is an unsafe state,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, at a given state s the step-wise probability of failure for arbitrary policy π is $\mathbb{E}_{a\sim\pi(\cdot|s)} I(s,a) \in [0,1]$.

Now we denote the *cumulative discounted probability of failure* as $V^{\pi}(s_t) = \mathbb{E}_{\pi} \sum_{t'=t} \gamma^{t'-t} I(s_{t'}, a_{t'})$, counting for the chance of entering dangerous states in current time step as well as in future trajectories deduced by the policy π. We use $V^{\mathcal{E}} = \mathbb{E}_{\tau \sim \mathcal{E}} V^{\mathcal{E}}(s_0)$ to denote the expected cumulative discounted probability of failure of the expert $\mathcal{E}$.

For simplicity, we can consider the actions post-processed by the guardian mechanism during training are sampled from a mixed policy $\hat{\pi}$, whose action probability can be written as:

$$\begin{aligned} \hat{\pi_\theta}(a|s) &= \pi_\theta(a|s)\mathbf{1}_{a \in \mathcal{A}_\eta(s)} + \mathcal{E}(a|s) \int\limits_{a' \notin \mathcal{A}_\eta(s)} \pi_\theta(a'|s) da' \\ &= \pi_\theta(a|s)\mathbf{1}_{a \in \mathcal{A}_\eta(s)} + \mathcal{E}(a|s) F(s). \end{aligned} \tag{3}$$

Here the second term captures the situation that the learning agent takes arbitrary action a' that triggers the expert to take over and chooses the action a. For simplicity, we use a shorthand $F(s) = \int_{a' \notin \mathcal{A}_\eta(s)} \pi_\theta(a'|s) da'$.

Following the same definition as $V^{\mathcal{E}}$, we can also write the expected cumulative discounted probability of failure of the behavior policy as: $\hat{V} = \mathbb{E}_{\tau \sim \hat{\pi}} \hat{V}(s_0) = \mathbb{E}_{\hat{\pi}} \sum_{t=0} \gamma^t I(s_t, a_t)$.

Assumption. Now we introduce one important assumption on the expert.

Assumption 1. *For all states, the step-wise probability of expert producing unsafe action is bounded by a small value $\epsilon < 1$:*

$$\mathop{\mathbb{E}}_{a \sim \mathcal{E}(\cdot|s)} I(s, a) \leq \epsilon. \tag{4}$$

This assumption does not impose any constrain on the structure of the expert policy.

Lemmas. We propose several useful lemmas and the correspondent proofs, which are used in the main theorem.

Lemma 1 (The performance difference lemma)**.**

$$\hat{V} = V^{\mathcal{E}} + \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim P_{\hat{\pi}}} \mathop{\mathbb{E}}_{a \sim \hat{\pi}} [A^{\mathcal{E}}(s, a)]. \tag{5}$$

Here the $P_{\hat{\pi}}$ means the states are subject to the marginal state distribution deduced by the behavior policy $\hat{\pi}$. $A^{\mathcal{E}}(s, a)$ is the advantage of the expert in current state action pair: $A^{\mathcal{E}}(s, a) = I(s, a) + \gamma V^{\mathcal{E}}(s') - V^{\mathcal{E}}(s)$ and $s' = \mathcal{P}(s, a)$ is the next state. This lemma is proposed and proved by Kakade and Langford [2] and is useful to show the behavior policy's safety. In the original proposition, the V and A represents the expected discounted return and advantage w.r.t. the reward, respectively. However, we replace the reward with the indicator $\mathcal{I}$ so that the value function $\hat{V}$ and $V^{\mathcal{E}}$ presenting the expected cumulative failure probability.

Lemma 2. *Only a small subspace of the confident action space of expert covers the ground-truth unsafe actions:*

$$\int\limits_{a \in \mathcal{A}_\eta(s)} I(s, a) da \leq \frac{\epsilon}{\eta}.$$

Proof. According to the Assumption, we have:

$$\epsilon \geq \int\limits_{a \in \mathcal{A}} \mathcal{E}(a|s) I(s, a) da = \int\limits_{a \in \mathcal{A}_\eta(s)} \mathcal{E}(a|s) I(s, a) da + \int\limits_{a \notin \mathcal{A}_\eta(s)} \mathcal{E}(a|s) I(s, a) da. \tag{6}$$

Following the definition of $\mathcal{A}_\eta(s)$, we get $\mathcal{E}(a|s) \geq \eta, \forall a \in \mathcal{A}_\eta(s)$. Therefore:

$$\epsilon \geq \int\limits_{a \in \mathcal{A}_\eta(s)} \eta I(s, a) da + 0 = \eta \int\limits_{a \in \mathcal{A}_\eta(s)} I(s, a) da. \tag{7}$$

Therefore $\int_{a \in \mathcal{A}_\eta(s)} I(s, a) da \leq \frac{\epsilon}{\eta}$ is hold. □

Lemma 3. *The cumulative probability of failure of the expert $V^{\mathcal{E}}(s)$ is bounded for all state:*

$$V^{\mathcal{E}}(s) \leq \frac{\epsilon}{1-\gamma}$$

Proof.

$$V^{\mathcal{E}}(s_t) = \mathop{\mathbb{E}}_{\mathcal{E}}[\sum_{t'=t}^{\infty} \gamma^{t'-t} I(s_{t'}, a_{t'})] = \sum_{t'=t}^{\infty} \gamma^{t'-t} \mathop{\mathbb{E}}_{\mathcal{E}}[I(s_{t'}, a_{t'})] \leq \sum_{t'=t}^{\infty} \gamma^{t'-t} \epsilon = \frac{\epsilon}{1-\gamma} \tag{8}$$

□

Theorem. We introduce the main theorem of this work, which shows that the training safety is related to the safety of the expert ϵ and the confidence level η.

Theorem 4 (Upper bound of the training risk)**.** *The expected cumulative probability of failure $\hat{V}$ of the behavior policy $\hat{\pi}$ in EGPO is bounded by the step-wise failure probability of the expert ϵ as well as the confidence level η:*

$$\hat{V} \leq \frac{\epsilon}{1-\gamma}(1 + \frac{1}{\eta} + \frac{\gamma}{1-\gamma} K'_{\eta}),$$

*wherein $K'_{\eta} = \max_s \int_{a \in \mathcal{A}_{\eta}(s)} da$ is **negatively correlated** to η.*

Proof. We use the performance difference lemma to show the upper bound. At starting, we first decompose the advantage by splitting the behavior policy:

$$\mathop{\mathbb{E}}_{a \sim \hat{\pi}(\cdot|s)} A^{\mathcal{E}}(s,a) = \int_{a \in \mathcal{A}} \pi(a|s) \mathbf{1}_{a \in \mathcal{A}_{\eta}(s)} A^{\mathcal{E}}(s,a) da + \int_{a \in \mathcal{A}} \mathcal{E}(a|s) F(s) A^{\mathcal{E}}(s,a) da \tag{9}$$

The second term is equivalent to $F(s)\, \mathbb{E}_{a \sim \mathcal{E}}[A^{\mathcal{E}}(s,a)]$, which is equal to zero, according to the definition of advantage. So we only need to compute the first term. Firstly we split the integral over whole action space into the confident action space and non-confident action space (which removed by the $\mathbf{1}$ operation), then we expand the advantage into detailed form, we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{a \sim \hat{\pi}(\cdot|s)} A^{\mathcal{E}}(s,a) = \int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) A^{\mathcal{E}}(s,a) da \\
&= \int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s)[I(s,a) + \gamma V^{\mathcal{E}}(s') - V^{\mathcal{E}}(s)] da \\
&= \underbrace{\int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) I(s,a) da}_{\text{(a)}} + \underbrace{\int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) \gamma V^{\mathcal{E}}(s') da}_{\text{(b)}} - \underbrace{\int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) V^{\mathcal{E}}(s) da}_{\text{(c)}}
\end{aligned} \tag{10}$$

Following the Lemma 2, the term (a) can be bounded as:

$$\int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) I(s,a) da \leq \int\limits_{a \in \mathcal{A}_{\eta}(s)} I(s,a) da \leq \frac{\epsilon}{\eta} \tag{11}$$

Following the Lemma 3, the term (b) can be written as:

$$\int\limits_{a \in \mathcal{A}_{\eta}(s)} \pi(a|s) \gamma V^{\mathcal{E}}(s') da \leq \gamma \int\limits_{a \in \mathcal{A}_{\eta}(s)} V^{\mathcal{E}}(s') da \leq \frac{\gamma \epsilon}{1-\gamma} \int\limits_{a \in \mathcal{A}_{\eta}(s)} da = \frac{\gamma \epsilon}{1-\gamma} K_{\eta}, \tag{12}$$

wherein $K_{\eta} = \int_{a \in \mathcal{A}_{\eta}(s)} da$ denoting the area of feasible region in the action space. It is a function related to the expert and η. If we tighten the guardian by increasing η, the confident action space

determined by the expert $\mathcal{A}_\eta(s)$ will shrink and the K_η will decrease. **Therefore K_η is negatively correlated to η.** The term (c) is always non-negative, so after applying the minus to term (c) will make it always ≤ 0.

Aggregating the upper bounds of three terms, we have the bound on the advantage:

$$\mathop{\mathbb{E}}_{a\sim\hat{\pi}} A^{\mathcal{E}}(s,a) \leq \frac{\epsilon}{\eta} + \frac{\gamma\epsilon}{1-\gamma}K_\eta \tag{13}$$

Now we put Eq. 13 as well as Lemma 3 into the performance difference lemma (Lemma 1), we have:

$$\begin{aligned}\hat{V} &= V^{\mathcal{E}} + \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim P_{\hat{\pi}}}\mathop{\mathbb{E}}_{a\sim\hat{\pi}}[A^{\mathcal{E}}(s,a)] \\ &\leq \frac{\epsilon}{1-\gamma} + \frac{1}{1-\gamma}[\frac{\epsilon}{\eta} + \frac{\gamma\epsilon}{1-\gamma}K'_\eta] \\ &= \frac{\epsilon}{1-\gamma}[1 + \frac{1}{\eta} + \frac{\gamma}{1-\gamma}K'_\eta].\end{aligned} \tag{14}$$

Here we have $K'_\eta = \max_s \int_{a\in\mathcal{A}_\eta(s)} da$. Now we have proved the upper bound of the cumulative probability of failure for the behavior policy in EGPO.

□

D Detail on Simulator and the Safe Driving Environments

The MetaDrive simulator is implemented based on Panda3D [3] and Bullet Engine that has high efficiency as well as accurate physics-based 3D kinetics. Some traffic cones and broken vehicles (with warning triangles) are scattered in the road network, as shown in Fig. 2. Collision to any object raises an environmental cost $+1$. The cost signal can be used to train agents or to evaluate the safety capacity of the trained agents.

In all environments, the observation of vehicle contains (1) 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 length of 240 Lidar-like cloud points with $50m$ maximum detecting distance measures of the nearby vehicles.

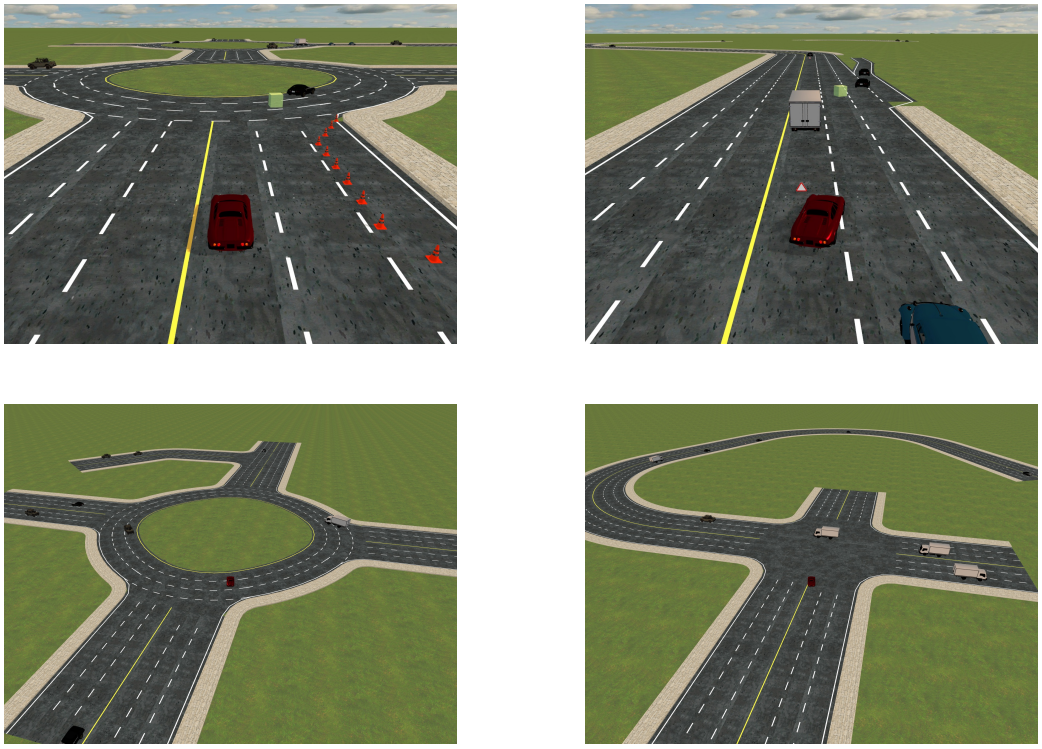

Figure 2: The demonstrations of generated safety environments.

E Learning Curves

Fig. 3 and Fig. 4 present the detailed learning curves of different approaches. Note that in CQL, the first 200,000 steps is for warming up and it uses the behavior cloning to train. In each DAgger iteration, a mixed policy will explore the environment and collect new data aggregated into the dataset. The mixed policy chooses action following $a_{\text{mixed}} = \beta a_{\text{expert}} + (1-\beta) a_{\text{agent}}$, where the parameter β anneals from 1 to 0 during training. Therefore DAgger agent achieves high training success rate at the beginning. In DAgger experiment, we only plot the result after each DAgger iteration.

We find that EGPO achieves expert-level training success rate at the very beginning of the training, due to the takeover mechanism. Besides, the test success rate improves drastically and achieves similar results as the expert. On the contrary, other baselines show inferior training efficiency.

In term of safety, due to the guardian mechanism, EGPO can constrain the training cost to a minimal value. Interestingly, during test time, EGPO agent shows even better safety compared to the expert. However, according to the main table in paper and the curves in Fig. 4, BC agent can achieve lower cost than EGPO agent. We find that the reason is because BC agent drives the vehicle conservatively in low velocity, while EGPO agent drives more naturally with similar velocity as the expert.

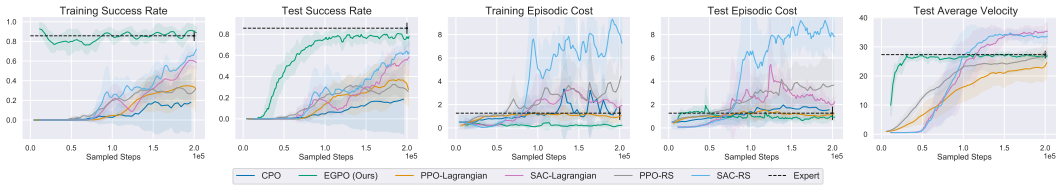

Figure 3: Detailed learning curves of EGPO and Safe RL baselines.

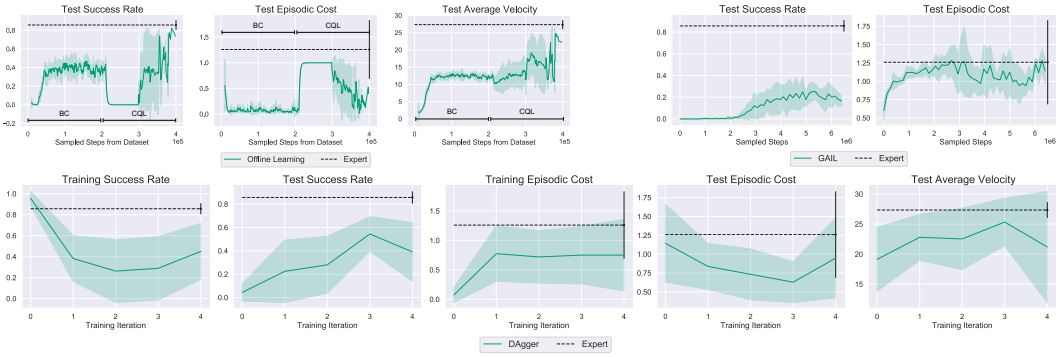

Figure 4: Detailed learning curves of BC, CQL, GAIL and DAgger.

F Hyper-parameters

Table 2: EGPO

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
Intervention Occurrence Limit C	20
Number of Online Evaluation Episode	5
K_p	5
K_i	0.01
K_d	0.1
CQL Loss Temperature β	3.0

Table 3: PPO/PPO-Lag

Hyper-parameter	Value
KL Coefficient	0.2
λ for GAE [4]	0.95
Discounted Factor γ	0.99
Number of SGD epochs	20
Train Batch Size	2000
SGD mini batch size	100
Learning Rate	0.00005
Clip Parameter ϵ	0.2
Cost Limit for PPO-Lag	1

Table 4: 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 5: BC

Hyper-parameter	Value
Dataset Size	250000
SGD Batch Size	32
SGD Epoch	200000
Learning Rate	0.0001

Table 6: DAgger

Hyper-parameter	Value
SGD Batch Size	64
SGD Epoch	2000
Learning Rate	0.0005
Number of DAgger Iteration	5
Initial β	0.3
Batch Size to Aggregate	5000

Table 7: GAIL

Hyper-parameter	Value
Dataset Size	250000
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

References

[1] Q. Li, Z. Peng, Z. Xue, Q. Zhang, and B. Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *arXiv preprint arXiv:2109.12674*, 2021.

[2] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[3] M. Goslin and M. R. Mine. The panda3d graphics engine. *Computer*, 37(10):112–114, 2004.

[4] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.