Safe CoR: A Dual-Expert Approach to Integrating Imitation Learning and Safe Reinforcement Learning Using Constraint Rewards

Hyeokjin Kwon[1], Gunmin Lee[2], Junseo Lee[1], Songhwai Oh[1,2]

Abstract— **In the realm of autonomous agents, ensuring safety and reliability in complex and dynamic environments remains a paramount challenge. Safe reinforcement learning addresses these concerns by introducing safety constraints, but still faces challenges in navigating intricate environments such as complex driving situations. To overcome these challenges, we present the safe constraint reward (Safe CoR) framework, a novel method that utilizes two types of expert demonstrations—reward expert demonstrations focusing on performance optimization and safe expert demonstrations prioritizing safety. By exploiting a constraint reward (CoR), our framework guides the agent to balance performance goals of reward sum with safety constraints. We test the proposed framework in diverse environments, including the safety gym, metadrive, and the real-world Jackal platform. Our proposed framework improves algorithm performance by 39% and reduces constraint violations by 88% on the real-world Jackal platform, highlighting its effectiveness. Through this innovative approach, we expect significant advancements in real-world performance, leading to transformative effects in the realm of safe and reliable autonomous agents.**

I. Introduction

Autonomous driving technology aims to revolutionize transportation by providing safer, more efficient, and accessible options. Researchers have proposed rule-based controllers [1], [2] and imitation learning methods [3], [4] to ensure safety and reliability in diverse environments. However, these methods struggle with scenarios beyond predefined rules or training data [5], limiting comprehensive coverage.

Reinforcement learning (RL) [6], [7] offers an alternative by allowing agents to learn optimal behaviors through trial and error, enhancing adaptability in complex situations. However, the exploratory nature of RL, which often requires agents to make mistakes to learn, poses a significant risk in real-world driving contexts where safety is crucial. This fundamental concern highlights the need for innovative approaches to balance exploration with safety.

Safe reinforcement learning (safe RL) [8], [9] addresses these concerns by integrating safety constraints into the optimization process. It enhances the agent's ability to adhere to safety constraints, thereby improving safety during both the training phase and the final deployment. Despite these advancements, challenges persist in the application of safe RL algorithms for training agents to navigate complex driving environments safely.

To overcome these challenges, we propose a method called safe CoR, which combines two types of expert demonstrations. The first, reward expert demonstrations, maximizes rewards without considering safety constraints. The second, safe expert demonstrations, prioritizes safety over rewards. By calculating a constraint reward (CoR) that measures how closely the agent aligns with these demonstrations, the agent emulates the reward expert for performance while using the safe expert to ensure safety. This dual-expert framework enhances the agent's ability to navigate complex driving scenarios, balancing performance with safety standards.

Experimental results show that safe CoR significantly improves performance and reduces constraint violations in platforms like the metadrive simulator [10] and safety gym environments [11]. The framework also outperformed the baseline methods on the sim-to-real Jackal platform [9], demonstrating its potential to advance safe RL.

The contributions of this paper are summarized as follows:

- We propose a framework called safe CoR, which uniquely integrates reward-centric and safety-conscious expert data to refine and enhance the performance of existing safe RL algorithms in the autonomous driving domain.
- We show empirical evidence demonstrating that agents, under the guidance of the safe CoR framework, outperform traditional safe RL algorithms by achieving superior performance metrics, especially in the real-world platform, with reduced rates of constraint violations in the training phase.
- We validate the superiority of the proposed algorithm in real-world scenarios utilizing the Jackal robot platform, thereby affirming the framework's applicability and robustness across diverse operational environments.

II. Related Work

A. Imitation learning

Imitation learning is a key approach for developing autonomous driving agents, guiding them to mimic expert demonstrations. A simple method, behavior cloning (BC), shows promise in real-world environments [12], [13], but suffers from compounding errors [14]. Inverse reinforcement

[1] H. Kwon, J. Lee, and S. Oh are with the Interdisciplinary Program in Artificial Intelligence and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {hyeokjin.kwon, junseo.lee}@rllab.snu.ac.kr, songhwai@snu.ac.kr) [2] G. Lee and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: gunmin.lee@rllab.snu.ac.kr).

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (NO.RS-2021-II211343, Artificial Intelligence Graduate School Program [Seoul National University], 50%) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2022R1A2C2008239, General-Purpose Deep Reinforcement Learning Using Metaverse for Real World Applications, 50%). *(Corresponding author: Songhwai Oh.)*

learning (IRL) [15] offers an alternative by learning the reward function from expert demonstrations. Ho et al. [16] introduced an algorithm combining IRL and RL, allowing agents to acquire expert behaviors and estimate reward functions simultaneously, proving the convergence of training policies and discriminators. This research inspired further studies [4], [17], [18].

Additionally, there have been studies that combine imitation learning with online learning. Yiren et al. [19] experimentally demonstrated that expert demonstrations can assist agents in navigating challenging environments robustly. However, these methods still have limitations, as they do not directly address safety constraints.

B. Safe reinforcement learning

Safe reinforcement learning (safe RL) integrates safety into the learning process, requiring agents to maximize rewards while meeting constraints. This approach can be divided into Lagrangian-based and trust-region-based methods.

Lagrangian-based methods convert the safe RL problem into a dual problem. Ray et al. [11] introduced the PPO-Lagrangian algorithm which extends the traditional PPO [20] framework. Yang et al. [21] proposed worst-case soft actor-critic (WCSAC), which relaxes constrained problems with Lagrangian multipliers. However, these methods can be overly conservative in updates during early learning stages, and the use of Lagrangian multipliers can make the learning process unstable.

Trust-region-based methods extend trust region policy optimization (TRPO) [22] to solve non-convex optimization problems. Achiam et al. [8] developed constrained policy optimization (CPO), which uses a trust region to ensure policy updates stay within predefined safety limits. Kim and Oh introduced TRC and OffTRC [9], [23], assuming the discounted cost sum follows a Gaussian distribution, and derived the closed-form upper bound of conditional value at risk (CVaR). Recently, Kim et al. [24] proposed SDAC, which employs a distributional critic and gradient-integration technique to improve agent stability. Despite these advancements, challenges persist in safely training agents for complex driving environments.

III. Preliminary

A. Constrained Markov decision process

A constrained Markov decision process (CMDP) is a framework that extends the traditional Markov decision process (MDP) by incorporating an additional constraint. A CMDP is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \rho, P, R, C, \gamma \rangle$: state space $\mathcal{S}$, action space $\mathcal{A}$, initial state distribution ρ, transition probability P, reward function R, cost function C, and discount factor γ. The expected reward sum J_π can be written in the aforementioned terms as follows:

$$J_\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right], \tag{1}$$

where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. Similarly, to define constraints, the expected cost sum can be expressed as follows:

$$C_\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t) \right]. \tag{2}$$

Then the objective of safe RL can be represented as follows:

$$\underset{\pi}{\text{maximize}} \ J_\pi \ \text{s.t.} \ C_\pi \leq \frac{d}{1-\gamma}, \tag{3}$$

with the constraint threshold d.

B. Constraint reward

A constraint reward (CoR) is an additional objective term that assesses the relative distance of an agent state between two sets of state data [4]. By utilizing two disparate sets of states, denoted as S_A and S_B respectively, the agent can estimate its performance relative to these two sets of demonstrations. If the distance between the agent's state and the first set of states, S_A, is less than the distance to the other set of states, S_B, CoR value exceeds 0.5. In contrast, when the agent's state is closer to S_B than S_A, CoR is reduced to below 0.5. In the prior work [4], by defining S_A as the collection of states associated with expert performance and S_B as those corresponding to suboptimal or negative behavior, such as random policy, CoR enables the training of agents to emulate expert trajectories over undesirable ones. For the state s, CoR is defined as follows:

$$\begin{aligned} \text{CoR}(s, S_A, S_B) &= \frac{\left(1 + \frac{\Delta_A}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\Delta_A}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\Delta_B}{\alpha}\right)^{-\frac{\alpha+1}{2}}}, \\ \Delta_A &= \sqrt{\frac{1}{|S_A|} \sum_{s_a \in S_A} \|s - s_a\|_2^2}, \\ \Delta_B &= \sqrt{\frac{1}{|S_B|} \sum_{s_b \in S_B} \|s - s_b\|_2^2}, \end{aligned} \tag{4}$$

where $\|\cdot\|_2$ is the l_2 norm, and α refers to a hyperparameter used to regulate the sensitivity of CoR.

IV. Safe CoR

This work aims to combine the strengths of imitation learning (IL) and safe reinforcement learning (safe RL) by leveraging expert demonstrations. A common approach is to modify the actor's objective by adding an IL term, like the log-likelihood probability $\mathbb{E}_{(s,a)\sim D}\left[\log \pi(a|s)\right]$, where D is a dataset of expert trajectories, as in [19]. However, challenges arise when applying this framework to safe RL. A reward-focused expert can lead to constraint violations, while a safe RL-trained expert may achieve lower rewards despite optimizing constraints. Thus, relying solely on either expert type is not ideal for our framework.

One approach to address these challenges is to use both demonstrations. When safety is assured, the agent prioritizes the reward expert for higher rewards. Conversely, if the agent struggles with constraints, it emulates the safe expert. This strategy helps balance the guidance from both experts. Building upon the foundational principles outlined in the preceding sections, the constraint reward (CoR) can serve as a guidance. The constraint reward (CoR), defined as $\text{CoR}(s_\pi, S_{re}, S_{se})$, where S_{re} and S_{se} are the reward and safe experts' demonstrations, allows us to evaluate the agent's alignment with each expert. CoR increases when the agent's state aligns with the reward expert and decreases when it aligns with the safe expert. Thus, CoR can be employed as an augmented reward with the coefficient λ for objective (1), as below:

$$\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\{R(s_t,a_t)+\lambda\text{CoR}(s_t,S_{re},S_{se})\}\right]. \tag{5}$$

While the enhanced objective (5) helps the agent achieve higher rewards, it's essential to maintain constraint satisfaction. To achieve this, we can integrate CoR into the constraint optimization process, enforcing stricter constraints as CoR value increases. Finally, we redefine the safe RL problem in (3) as follows:

$$\begin{aligned}&\underset{\pi}{\text{maximize}}\;\; J_\pi+\lambda_r\cdot\text{CoR}_\pi\\&\text{s.t. } C_\pi+\lambda_c\cdot\text{CoR}_\pi\leq\frac{d}{1-\gamma},\end{aligned} \tag{6}$$

where

$$\text{CoR}_\pi:=\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\text{CoR}(s_t,S_{re},S_{se})\right]. \tag{7}$$

λ_r determines the influence of expert guidance compared to the original reward function. As its value increases, the optimization problem shifts towards the objective of IL. λ_c affects constraint satisfaction, with higher values encouraging a more conservative approach. However, excessively large values can alter the original problem. Therefore, to ensure stable training, we assign values of 0.1 to λ_r and 0.01 to λ_c.

V. Experimental Setup

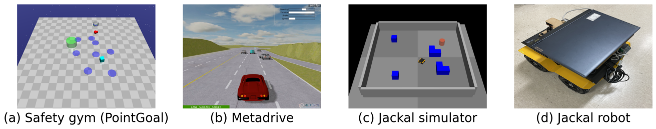
(a) Safety gym (PointGoal) (b) Metadrive (c) Jackal simulator (d) Jackal robot

Fig. 1: Experiment environments

In this section, we outline the experimental setup for deploying our framework in both simulators and real-world platform. We use two simulators, safety gym [11] and metadrive [10], along with the real-world Jackal platform [9]. Each episode in both simulators and the real-world platform lasts up to 1,000 steps. In safety gym, we compare five safe RL algorithms with our framework: PPO-L [11], WCSAC [21] (Lagrangian-based), and CPO [8], OffTRC [23], and SDAC [24] (trust-region-based). However, due to suboptimal performance of Lagrangian-based methods in the Jackal platform and metadrive, we focus on three trust-region-based methods. In metadrive, we also compare against BC-SAC [19].

A. Safety gym

The safety gym [11] environment offers tasks for testing safe RL algorithms, providing a cost function. It features two predefined robots (point and car) performing the goal and button tasks. In the goal task, the agent is required to reach a randomly assigned goal point while avoiding hazard areas. In the button task, the agent presses a button, avoiding hazard areas and moving obstacles. For both tasks, the environmental settings including the reward and cost function are the same as in [24]. The number of constraint violations (CV) is counted when the agent enters hazard zones or collides with obstacles. The description image of the environment is shown in Figure 1-(a).

B. Metadrive

The metadrive simulator, an autonomous driving platform, challenges an agent to navigate to a destination without deviating from the road, colliding with objects, or exceeding a maximum speed which we set to 30km/s for stable training. The reward function $R(s_t, a_t)$ focuses on speed and waypoint navigation, while the cost function $C(s_t, a_t)$ increases by 1 for unsafe actions like off-road incidents or collisions. Additionally, we define the score function $\Theta(s_t, a_t)$ using the coefficient $l_c(=5)$ in Eq. (8) to assess the overall driving performance. The constraint threshold d is configured at 0.02.

$$\Theta(s_t,a_t):=R(s_t,a_t)-l_cC(s_t,a_t). \tag{8}$$

Our methodology involves training experts using TRC [9] with different constraint thresholds: 0.5 for the reward expert and 0.001 for the safe expert. During testing, we evaluate the average sum of score, reward, violations, and success probabilities over 100 episodes. If an off-road incident occurs, the agent's position resets to the start, but the step count continues, ensuring continuity. The simulator's layout is shown in Figure 1-(b).

C. Jackal simulator and real-world platform

For the real-world Jackal experiment, we apply the sim-to-real method to the Jackal platform, using the pretrained agents from the Jackal simulator [9]. The Jackal simulator is an environment for training safe RL algorithms, where it utilizes the robot platform for solving the goal task of the safety gym. The state is composed of LiDAR sensor data, linear and angular velocity, as well as goal direction and distance. The overall settings of the simulator are the same as in [23]. The constraint threshold d is configured at 0.025. The description image of the environment is shown in Figure 1-(c) and (d).

VI. Results

A. Safety gym

The results from each environment are illustrated in Figure 2. Note that due to the superior performance of SDAC with the risk level of 1.0 compared to other baseline algorithms, the figure exclusively presents the results of SDAC enhanced by the safe CoR application. It suggests that the integration of

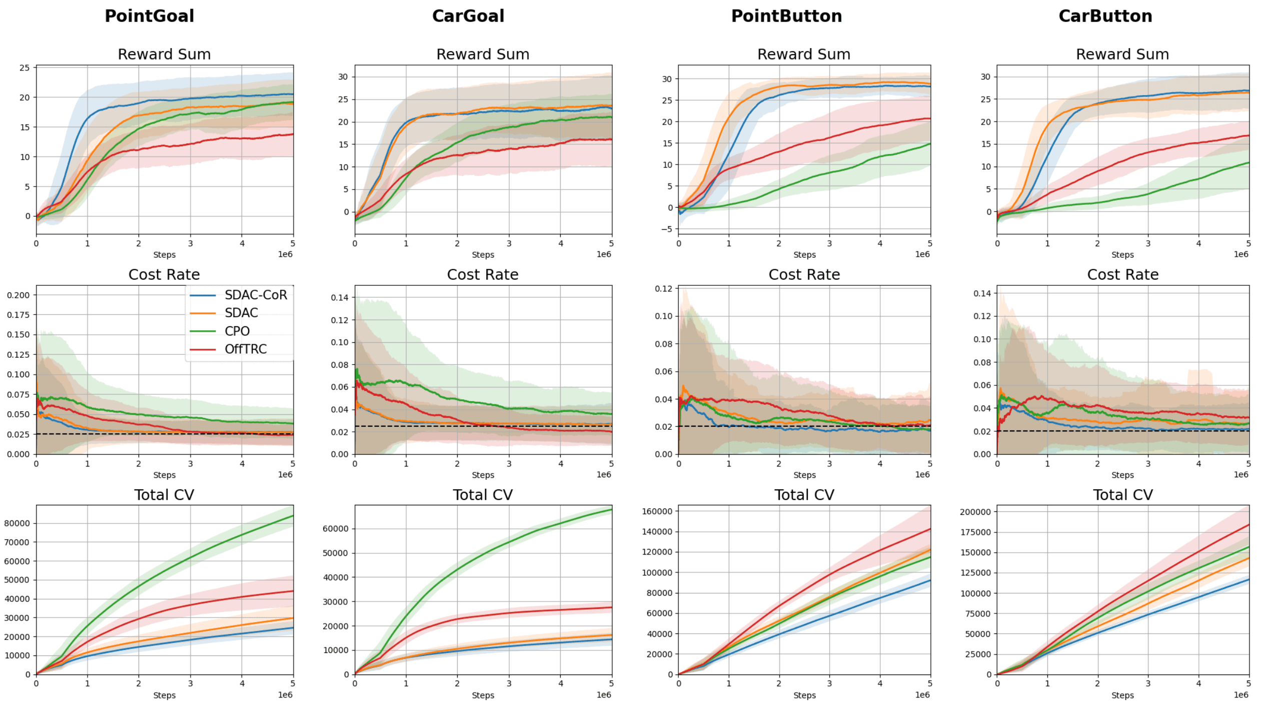

Fig. 2: Safety gym results. The cost rate refers to the average cost per step and the dashed line indicates the constraint threshold.

Algorithm	Reward	Cost(≤ 25)	CV	Total CV($\times 10^3$)
PPO-L [11]	1.55	**14.559**	**1.915**	31.315
PPO-L+CoR	**4.203**	21.168	4.23	**22.192**
WCSAC [21]	12.744	38.382	1.08	32.778
WCSAC+CoR	**14.951**	**24.903**	**0.21**	**25.407**
SDAC [24]	17.479	24.467	2.83	33.06
SDAC+CoR	**19.336**	**21.714**	**1.723**	**24.581**
CPO [8]	**21.554**	43.231	11.31	83.816
CPO+CoR	19.119	**34.951**	**5.81**	**62.776**
OffTRC [23]	**16.117**	25.253	2.507	44.016
OffTRC+CoR	14.927	**18.065**	**1.327**	**40.81**

TABLE I: PointGoal results across five seeds. For each algorithm, the best performance for each element is denoted by a bold value.

the proposed framework with SDAC yields similar outcomes to the original SDAC in terms of reward sums. Nonetheless, when evaluating the cost rate and the sum of constraint violations in the training phase (total CV), the framework significantly outperforms the baseline algorithms across all environments.

Furthermore, in order to assess the versatility and effectiveness of the proposed framework, we conducted an experiment on the PointGoal task with baseline algorithms augmented by the safe CoR. As demonstrated in Table I, the implementation of the proposed framework resulted in a beneficial impact on the overall performance. For both WCSAC and SDAC, the utilization of the framework led to a decrease in the sum of costs and constraint violations, while simultaneously enhancing the sum of rewards. Remarkably, while SDAC initially stood out as a state-of-the-art algorithm among the baselines, the proposed framework further enhanced its performance. For PPO-L, the safe CoR demonstrated improvements in both reward sum and total constraint violations, while still maintaining adherence to the constraint threshold. The decrease in reward sum observed in CPO with the safe CoR is considered reasonable, given that the original CPO algorithm faced challenges in satisfying constraints. However, with OffTRC, the proposed framework led to a decrease in the reward sum, resulting in a significantly lower cost sum compared to the constraint threshold. Considering that OffTRC operated with a risk level of CVaR as 0.25, indicative of an already conservative training approach, the implementation of the safe CoR led the agent to adhere to excessively stringent constraint.

B. Metadrive

The final result of the simulation experiment is shown in Table II. The most crucial metric is the score as it reflects the agent's effort to progress along the waypoints. Additionally, the success ratio provides an overall measurement of the agent's performance in achieving the objective. The violation comprises the sum of crash and out-of-road situations.

SAC showed minimal learning in the environment and did not reach the destination in any instance. Incorporating expert demonstrations into SAC resulted in an improved driving score. However, the agent still encountered difficulties in reaching the final goal. For safe RL algorithms, CPO exhibited superior performance. However, when applying the safe CoR, OffTRC outperforms all algorithms. The implementation of OffTRC with the framework yielded an enhancement in the driving score with a remarkable reduction in violations. Notably, the number of crashes decreased significantly. In the case of CPO, there was a 4% improvement in the score

	Algorithm	Score	Reward	Crash	Out of road	Violations	Success
Expert	TRC-0.5(Reward) [9]	564.303	861.530	25.45	0.14	25.59	0.97
	TRC-0.001(Safe)	780.252	847.405	7.61	1.53	9.14	0.62
	Human	923.557	923.557	0.0	0.0	0.0	1.0
RL	SAC [25]	76.887	113.925	0.0	7.12	7.12	0.0
	BC-SAC [19]	302.547	334.554	0.78	5.12	5.56	0.0
Safe RL	SDAC	**638.554**	**763.849**	17.1	**5.26**	22.36	**0.17**
	SDAC+CoR	632.830	753.038	**15.2**	6.72	**21.92**	0.02
	CPO	683.095	854.472	15.91	0.12	16.03	0.94
	CPO+CoR	**711.330**	**855.201**	**13.6**	**0.11**	**13.71**	**0.98**
	OffTRC	613.046	849.795	23.13	**0.19**	23.32	**0.94**
	OffTRC+CoR	**794.472**	**882.178**	**8.64**	0.24	**8.88**	0.93

TABLE II: Metadrive results across five seeds.

along with a 14.5% decrease in violations. For SDAC, the application of the framework led to lower performance. However, considering the success ratios of both versions, it appears that SDAC faced challenges during training within the simulator.

In general, safe RL agents demonstrated competent performance, and when coupled with safe CoR, they exhibited improved performance compared to the original algorithms.

C. Jackal platform

In this subsection, we conduct a comparative analysis of various safe RL algorithms and the effectiveness of the safe CoR framework applied to these algorithms, utilizing both the Jackal simulator and the real-world Jackal platform for our experiments. The outcomes from the Jackal simulator are depicted in Table III, and the results from the real-world Jackal platform are detailed in Table IV.

The data illustrates that the integration of the safe CoR framework with safe RL algorithms enhances performance, in terms of reducing constraint violations. Moreover, this integration leads to a reduction in the overall cost sum across most scenarios, indicating an improvement in agent safety. An exception is observed with the OffTRC algorithm in the real-world experiment, where a marginal increase in the cost sum is noted. This anomaly can be attributed to the inherent safety levels of the OffTRC method, which are already high, leaving minimal scope for further enhancements through the safe CoR framework.

Regarding the overall performances, the outcomes from the real-world Jackal platform indicate that the implementation of the proposed framework yields an improvement over conventional safe RL methods. Conversely, within the Jackal simulator environment, the deployment of CPO and OffTRC algorithms exhibits a marginal decline in score values. This phenomenon occurs as the agent opts to sacrifice potential score increases in favor of significantly reducing the cost sum.

In summary, the results demonstrate that the application of the framework substantially reduces constraint violations and cost sums, while concurrently enhancing the score value. Consequently, it can be deduced that the proposed framework significantly augments the robustness of safe RL algorithms against environmental variations as the framework application results are superior in the real world.

Algorithm	Reward	Cost($\leq$ 25)	CV
PPO-L	1.464	24.059	10.825
PPO-L+CoR	**2.126**	**11.513**	**5.065**
WCSAC	2.349	**9.049**	**0.0**
WCSAC+CoR	**6.777**	20.313	14.58
SDAC	4.894	26.317	6.553
SDAC+CoR	**8.256**	**23.18**	**2.387**
CPO	**7.764**	17.788	1.973
CPO+CoR	7.412	**12.94**	**1.447**
OffTRC	**12.512**	18.039	1.087
OffTRC+CoR	11.288	**14.473**	**0.287**

TABLE III: Jackal simulator results across three seeds within 1000 steps per episode.

Algorithm	Reward	Cost($\leq$ 25)	CV
SDAC	9.39	112.503	348.667
SDAC+CoR	**13.027**	**58.07**	**40.667**
CPO	8.877	36.21	19.333
CPO+CoR	**16.273**	**12.6**	**0.0**
OffTRC	21.557	**16.48**	**0.0**
OffTRC+CoR	**21.977**	17.347	**0.0**

TABLE IV: Real-world Jackal platform results across three seeds within 1000 steps per episode.

VII. Ablation Study

In our ablation study, we aim to quantitatively assess the differential impacts of CoR term on the performance and safety metrics of the navigation task using SDAC [24]. For a fair comparison, we included the results obtained when the risk level was set to 0.25. Furthermore, we assessed the influence of the proposed framework in comparison to the impact of log-likelihood probability, as in [19]. We employed TRPO [22] for the reward expert and TRC [9] for the safe expert.

Table V illustrates the outcomes of applying our proposed framework to different components of the system. Implementing the framework exclusively within the reward function yields a positive effect on the overall score but adversely affects the cost sum. Conversely, when the framework is applied solely to the cost function, we observe enhancements in the score, cost sum, and constraint violations (CV). The outcomes derived from employing the log-likelihood probability demonstrate that the exclusive integration of a single type of expert does not exhibit significantly improved performance. When employing the reward expert, SDAC-TRPO, the agent encounters challenges in meeting the constraint threshold, despite exhibiting slightly enhanced performance in reward

Algorithm	Reward	Cost(≤ 25)	CV	Total CV($\times 10^3$)
SDAC-0.25	13.832	13.551	0.034	7.768
SDAC-1.0	17.479	24.467	2.83	33.06
SDAC-TRPO	17.559	31.248	6.1	26.343
SDAC-TRC	17.277	26.341	2.58	24.618
RewCor	**21.78**	27.241	2.78	28.429
CostCor	18.597	22.393	2.374	25.65
***RewCostCor**	19.336	**21.714**	**1.723**	**24.581**

TABLE V: PointGoal results for SDAC. The expert algorithm employed for integrating the log-likelihood is indicated alongside SDAC.

maximization. However, the utilization of the safe expert, SDAC-TRC, does not demonstrate improved performance in both reward maximization and safety metrics.

The most comprehensive benefits are observed when CoR term is utilized as outlined in our proposed methodology. This approach results in optimal outcomes for safety metrics, surpassing the results of the other configurations. Although the score marginally decreases compared to its application solely in the reward function (RewCoR), this reduction is a deliberate trade-off to achieve lower constraint violations and cost sums. This strategy underscores the inherent balance between optimizing performance and enhancing safety within safe RL paradigms.

VIII. Conclusion

In this paper, the safe CoR framework is introduced as an innovative solution to the critical challenges of ensuring safety and reliability in the complex and dynamic environments encountered by autonomous agents. By ingeniously combining reward-focused and safety-oriented expert demonstrations, the safe CoR framework has significantly advanced the field of safe reinforcement learning (safe RL). Our empirical investigations across a variety of environments, including safety gym, metadrive, and the real-world Jackal platform, have demonstrated the framework's remarkable ability to enhance algorithmic performance, while concurrently reducing constraint violations. These results not only underscore the efficacy of the safe CoR framework in balancing performance with safety constraints but also highlight its capability to enhance the domain of autonomous agents and related fields. The contributions of this work, particularly the validation of our framework's superiority in real-world scenarios and its robust applicability across diverse environments, suggest the promising potential for advancements in the development of safe and reliable autonomous agents.

References

[1] A. Aksjonov and V. Kyrki, "Rule-based decision-making system for autonomous vehicles at intersections with mixed traffic environment," in *Proc. of the International Intelligent Transportation Systems Conference (ITSC), Sep*, 2021.

[2] W. Xiao, N. Mehdipour, A. Collin, A. Y. Bin-Nun, E. Frazzoli, R. D. Tebbens, and C. Belta, "Rule-based optimal control for autonomous driving," in *Proc. of the International Conference on Cyber-Physical Systems (ICCPS), May*, 2021.

[3] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] G. Lee, D. Kim, W. Oh, K. Lee, and S. Oh, "MixGAIL: Autonomous driving using demonstrations with mixed qualities," in *Proc. of the International Conference on Intelligent Robots and Systems (IROS), Oct*, 2021.

[5] J. Huang, S. Xie, J. Sun, G. Q. Ma, C. Liu, D. Lin, and B. Zhou, "Learning a decision module by imitating driver's control behaviors," in *Proc. of the Conference on Robot Learning (CoRL), Nov*, 2020.

[6] L. Wang, J. Liu, H. Shao, W. Wang, R. Chen, Y. Liu, and S. L. Waslander, "Efficient reinforcement learning for autonomous driving with parameterized skills and priors," 2023.

[7] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2021.

[8] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. of the International Conference on Machine Learning (ICML), August*, 2017.

[9] D. Kim and S. Oh, "TRC: trust region conditional value at risk for safe reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2621–2628, 2022.

[10] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3461–3475, 2022.

[11] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," vol. 7, no. 1, p. 2, 2019.

[12] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[13] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *Proc. of the International Conference on Robotics and Automation (ICRA), May*, 2018.

[14] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS), Apr*, 2011.

[15] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI), Jul*, 2008.

[16] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[17] Y. Li, J. Song, and S. Ermon, "Infogail: Interpretable imitation learning from visual demonstrations," *Advances in neural information processing systems*, vol. 30, 2017.

[18] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal Wasserstein imitation learning," in *Proc. of the International Conference on Learning Representations (ICLR), May*, 2021.

[19] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," in *Proc. of the International Conference on Intelligent Robots and Systems (IROS), Oct*, 2023.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[21] Q. Yang, T. D. Simão, S. H. Tindemans, and M. T. Spaan, "WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI), May*, 2021.

[22] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. of the International Conference on Machine Learning (ICML), July*, 2015.

[23] D. Kim and S. Oh, "Efficient off-policy safe reinforcement learning using trust region conditional value at risk," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7644–7651, 2022.

[24] D. Kim, K. Lee, and S. Oh, "Trust region-based safe distributional reinforcement learning for multiple constraints," *Advances in neural information processing systems*, vol. 36, 2024.

[25] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. of the International Conference on Machine Learning (ICML), Jul*, 2018.