Constraint-Aware Reinforcement Learning via Adaptive Action Scaling

Murad Dawood DAWOOD@CS.UNI-BONN.DE

Usama Ahmed Siddiquie S46USIDD@UNI-BONN.DE

Shahram Khorshidi KHORSHIDI@CS.UNI-BONN.DE

Maren Bennewitz MAREN@CS.UNI-BONN.DE
Humanoid Robots Lab, University of Bonn, Germany

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Safe reinforcement learning (RL) seeks to mitigate unsafe behaviors that arise from exploration during training by reducing constraint violations while maintaining task performance. Existing approaches typically rely on a single policy to jointly optimize reward and safety, which can cause instability due to conflicting objectives, or they use external safety filters that override actions and require prior system knowledge. In this paper, we propose a modular cost-aware regulator that scales the agent's actions based on predicted constraint violations, preserving exploration through smooth action modulation rather than overriding the policy. The regulator is trained to minimize constraint violations while avoiding degenerate suppression of actions. Our approach integrates seamlessly with off-policy RL methods such as SAC and TD3, and achieves state-of-the-art return-to-cost ratios on Safety Gym locomotion tasks with sparse costs, reducing constraint violations by up to 126 times while increasing returns by over an order of magnitude compared to prior methods.

1. Introduction

Reinforcement Learning (RL) has demonstrated remarkable success across a range of domains, including Atari games Mnih et al. (2015), robotics Gu et al. (2017); Wu et al. (2023), and long-horizon strategy games Silver et al. (2017); Vinyals et al. (2019). This success is significantly facilitated by exploratory behavior, which allows agents to discover effective behaviors. However, such exploratory behaviors often lead to the violation of system constraints. While such violations are tolerable in simulation and games with free resets, they pose serious risks in real-world applications Amodei et al. (2016). Violating safety constraints can lead to irreversible damage or system failure. To address this issue, Safe Reinforcement Learning (Safe RL) Garcıa and Fernández (2015) aims to minimize constraint violations during both training and deployment.

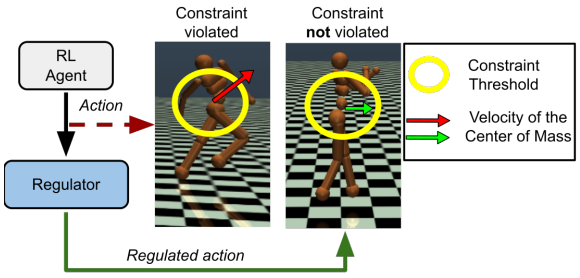

Figure 1: Overview of cost-aware action scaling. The RL agent proposes an action that would result in the center of mass (COM) exceeding the velocity threshold (left). The regulator (blue) scales the action, keeping the velocity of the COM within the safe zone while allowing progress on the task. The yellow circles highlight the velocity threshold for the COM.

Safe RL methods can be broadly categorized into two groups: *safe exploration* and *constrained RL*. Safe exploration techniques aim to prevent the agent from taking actions that violate safety constraints. These methods typically rely on prior knowledge of the system dynamics and feasible safe states to construct control barrier functions Ames et al. (2019); Dai et al. (2023); Zhang et al. (2023b), or model predictive shields Banerjee et al. (2024); Dawood et al. (2025b); Agha et al. (2024). Although effective, their applicability is limited by the need for detailed prior information about the system dynamics, an assumption that often does not hold in early learning stages or tasks where system dynamics are unknown.

Constrained RL allows the agent to learn both reward and cost signals online, without requiring knowledge of the system dynamics. The agent is trained to maximize cumulative rewards while minimizing constraint violations. Common approaches include Lagrangian-based methods Achiam et al. (2017); Tessler et al. (2018); Ray et al. (2019); Stooke et al. (2020), and budget-based methods Sootla et al. (2022b,a). However, a core limitation of these methods is the difficulty of balancing reward and cost within a single policy. Conflicting gradients can cause the agent to behave either too conservatively or unsafely, leading to instability, constraint violations, or poor performance Stooke et al. (2020); Navon et al. (2022).

In contrast to prior work, we propose a modular alternative: instead of overriding actions or jointly optimizing conflicting objectives, we scale actions based on the expected cost of future constraint violations while preserving the policy's task-directed behavior (see Fig. 1).The architecture consists of a reward-maximizing task agent and a regulator network guided by twin cost critics to conservatively estimate constraint violations. The regulator applies element-wise scaling to attenuate risky actions, enforcing safety without requiring prior knowledge of dynamics or compromising exploration. Although our approach resembles safe exploration in formulation, we do not require prior knowledge of the system dynamics, and we do not override the task agent's actions, thereby preserving both exploration and safety without external overrides.

We evaluate our approach on several dynamical systems from Safety Gymnasium Ji et al. (2023). Our method achieves the highest Return-to-Cost (RC) ratio Thananjeyan et al. (2021), reducing constraint violations by up to 126 times over recent safe RL baselines Stooke et al. (2020); Sootla et al. (2022a); Yu et al. (2022); Ganai et al. (2023); Kim et al. (2024). In summary, our contributions are: (i) We propose a modular safe RL framework that decouples reward maximization and safety enforcement via a cost-aware regulator that scales actions based on predicted violations. (ii) Our model-free approach integrates seamlessly into standard off-policy RL pipelines such as SAC Haarnoja et al. (2018) and TD3 Fujimoto et al. (2018), improving safety without compromising exploration. (iii) We achieve state-of-the-art performance on safety benchmarks, with up to 126 times fewer constraint violations and the highest RC ratios across tasks.

2. Related Work

Existing approaches fall into two categories: *safe exploration methods*, which prevent unsafe actions, and *constrained RL methods*, which embed cost objectives directly into policy optimization.

Safe Exploration Methods. Early methods, Sui et al. (2015), employed uncertainty modeling through Gaussian Processes to restrict exploration and ensure safety. Later approaches introduced safety layers Dalal et al. (2018); Sheebaelhamd et al. (2021) and predictive safety filters Banerjee et al. (2024); Dawood et al. (2025a); Agha et al. (2024); Tian et al. (2024) that prevent risky actions based on pre-trained layers or model predictive control (MPC). Control Barrier Function (CBF)-based strategies Ames et al. (2019); Dai et al. (2023); Zhang et al. (2023b) employ differentiable barriers to

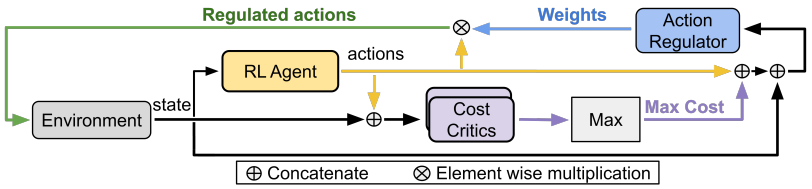

Figure 2: Overview of our modular safe RL architecture. The regulator (blue) scales actions produced by the unconstrained RL agent (yellow) based on predicted cost (purple), producing safety-aware actions (green) that are executed in the environment.

keep actions within certified safe sets. Goodall and Belardinelli (2024) extend shielding methods to continuous domains by leveraging approximate dynamics models, enabling probabilistic safety guarantees during exploration. Selim et al. (2022) leveraged model-based RL and offline collected data to develop reachability-based safety layers to ensure safe actions for navigation scenarios. Thananjeyan et al. (2021); Zhang et al. (2023a) assumes access to an offline dataset for pretraining a cost critic along with a recovery policy, which is then fixed during online learning, limiting its applicability in settings where collecting sufficient offline data is challenging or costly. **While** our method also modifies actions to maintain safety, it differs fundamentally by relying on online-learned cost predictions rather than external models or handcrafted safe sets, and by smoothly scaling actions instead of hard blocking or overwriting them, preserving the agent's exploratory behavior.

Constrained RL Approaches. Lagrangian-based algorithms Achiam et al. (2017); Ray et al. (2019); Stooke et al. (2020) optimize dual formulation balancing rewards and costs, while budgeted RL Sootla et al. (2022b,a) include the remaining cost budget in the state representation, allowing the agent to adapt its behavior based on how much cost it can afford. Risk-sensitive formulations, such as CVaR-CPO Zhang et al. (2024), constrain the conditional value-at-risk of cumulative costs, ensuring attention to costly violations. Reachability-based methods like RESPO Ganai et al. (2023) estimate the probability of reaching safe regions and optimize policies to satisfy constraints or recover when outside the feasible set. Bi-level optimization frameworks such as SRCPO Kim et al. (2024) address the nonlinearity of risk measures by optimizing over dual variables, achieving strong constraint satisfaction in continuous control tasks. Safety Editor Yu et al. (2022) trains two Soft Actor-Critic (SAC) Haarnoja et al. (2018) agents: a utility maximizer and a safety editor that modifies unsafe actions, allowing it to fully overwrite the original action when necessary. **Compared** to these methods, our approach offers a lightweight, modular alternative: instead of embedding constraints into the policy loss, relying on delicate dual updates, or training a second actor to overwrite unsafe actions, we regulate actions externally using learned cost critics. This continuous scaling preserves exploration while enabling seamless integration into standard off-policy RL pipelines.

3. Preliminaries

Markov Decision Processes. We consider a Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P(s'|s,a)$ the transition probability, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in (0,1)$ the discount factor. We assume continuous state and action spaces with $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^d$.

Constrained Markov Decision Processes. A CMDP Altman (2021) augments the MDP with a cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ that quantifies safety violations. The objective is to maximize return while keeping the expected cumulative cost below a budget χ:

$$\max_{\pi} \ \mathbb{E}_{s\sim d_0,\, a\sim\pi(\cdot|s)}[Q^{\pi}(s,a)] \quad \text{s.t. } \mathbb{E}_{s\sim d_0,\, a\sim\pi(\cdot|s)}[Q_c^{\pi}(s,a)] \leq \chi, \tag{CMDP}$$

Cost Budget. The budget χ specifies the maximum allowable expected cumulative cost and is typically treated as a human-selected threshold that reflects task-specific safety requirements Stooke et al. (2020). In this work, we assume a stricter setting by eliminating the cost budget, i.e., setting $\chi = 0$, similar to Ganai et al. (2023). This corresponds to a hard-safety regime that aims to achieve minimal constraint violations during learning.

Problem Setting. With this stricter formulation, the problem considered in this work is to learn a policy that maximizes task rewards while minimizing constraint violations under the hard-safety regime $\chi = 0$. Formally, our objective reduces to:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{s\sim d_0,\, a\sim\pi(\cdot|s)}[Q^\pi(s,a)] \quad \text{s.t. } \mathbb{E}_{s\sim d_0,\, a\sim\pi(\cdot|s)}[Q_c^\pi(s,a)] = 0. \tag{1}$$

This assumption eliminates any positive cost budget and focuses on policies that aim to achieve minimal safety violations during training and execution.

4. Methodology

We propose a modular safe reinforcement learning framework that regulates the actions of a task policy to reduce expected constraint violations without overriding agent decisions. The key idea is to scale actions based on their predicted cost, preserving exploration while inducing smoother and safer transitions in the environment.

4.1. Split Architecture for Reward and Cost Optimization

Optimizing for both task rewards and safety constraints within a single policy often leads to instability or overly conservative behavior Stooke et al. (2020). To address this, we decouple the reward and cost learning objectives across two modules: The **task policy** $\pi_\phi(a|s)$, which is trained to maximize expected rewards without incorporating safety constraints. The **regulator network** $\rho_\theta(s, a, \hat{c})$, which learns to scale the policy's actions based on cost predictions to minimize constraint violations.

4.2. Action Modulation via Regulator Scaling

In continuous control environments without stochasticity, the system evolves under deterministic transition dynamics of the form $s_{t+1} = f(s_t, a_t)$, where $s_t \in \mathcal{S}$ is the current state and $a_t \in \mathcal{A} \subset \mathbb{R}^d$ is a d-dimensional real-valued action vector, with d denoting the number of action dimensions. Since actions directly control the system evolution, high-magnitude or poorly directed actions can result in constraint violations or unstable behaviors. To mitigate this, we introduce a *regulator network* $\rho_\theta : \mathcal{S} \times \mathcal{A} \times \mathbb{R} \to (0, 1]^d$, which learns a scaling vector with an individual factor for each action dimension based on the current state, the raw action, and its predicted cost. At each step, the agent samples a raw action $a_t \sim \pi_\phi(\cdot|s_t)$, computes the cost estimate: $\hat{c}_t = \max(Q_c^1(s_t, a_t), Q_c^2(s_t, a_t))$ from a twin-critic architecture, and the regulator network outputs a scaling vector ρ_t, see Fig. 2. The final action applied to the system is:

$$\tilde{a}_t = \rho_t \odot a_t, \quad \text{where } \rho_t = \rho_\theta(s_t, a_t, \hat{c}_t), \tag{2}$$

where $\odot$ denotes element-wise multiplication; each component of the action vector is multiplied by a scaling factor between 0 (applied when high risk is predicted, resulting in large attenuation) and 1 (applied when the action is predicted to be safe, resulting in no attenuation), smoothly reducing potentially unsafe actions proportional to predicted risk. This element-wise modulation attenuates each component of the action based on its risk profile, reducing the magnitude of high-risk

components. Unlike hard safety constraints that may override agent behavior, this approach preserves the agent's exploration behavior and allows stable off-policy learning.

Assumptions. Our scaling mechanism assumes: (i) *Monotonicity*, where $Q_c(s, \rho \odot a)$ is expected to decrease as $\rho \to 0$, and (ii) *Safety of Inaction*, where zero-magnitude actions are safer than high-magnitude ones. These properties hold in many robotic domains where safety costs often scale with action magnitude, such as high torques causing actuator wear or large contact forces risking joint damage. By designing costs with this structure, scaling becomes an intuitive and effective tool for enforcing safety.

4.3. Learning Objectives and Updates

Reward Learning. We adopt a general off-policy reinforcement learning framework where the agent's actor and critic are trained using the scaled action $\tilde{a}_t$, as this is the action that is actually executed in the environment. The reward critic is updated using:

$$Q_r(s_t, \tilde{a}_t) \;\leftarrow\; r(s_t, \tilde{a}_t) + \gamma\, \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\tilde{a}_t) a_{t+1}\sim\pi(\cdot|s_{t+1})}[Q_r(s_{t+1}, \tilde{a}_{t+1})]\,, \tag{3}$$

where $\tilde{a}_{t+1} = \rho_{t+1} \odot a_{t+1}$ and $\rho_{t+1} = \rho_\theta(s_{t+1}, a_{t+1}, Q_c(s_{t+1}, a_{t+1}))$. The policy is updated to maximize the expected return under the regulated action:

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{s_t\sim\mathcal{D}, a_t\sim\pi(\cdot|s_t)}\left[-Q_r(s_t, \tilde{a}_t)\right], \tag{4}$$

ensuring that policy learning reflects the actual dynamics induced by the regulated action $\tilde{a}_t$. Our framework is algorithm-agnostic and can be integrated with any off-policy actor-critic method. For entropy-regularized algorithms such as SAC, the corresponding entropy term may be included in the actor objective. In our experiments, we demonstrate compatibility with both SAC and Twin Delayed DDPG (TD3)Fujimoto et al. (2018).

Cost Learning. The cost critic is also trained on the scaled actions using a TD-style Bellman backup Sutton et al. (1998):

$$Q_c(s_t, \tilde{a}_t) \;\leftarrow\; c(s_t, \tilde{a}_t) + \gamma\, \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,\tilde{a}_t) a_{t+1}\sim\pi(\cdot|s_{t+1})}[Q_c(s_{t+1}, \tilde{a}_{t+1})]\,, \tag{5}$$

This ensures the critic reflects the safety implications of the actual executed action $\tilde{a}_t$.

Regulator Objective. The regulator is trained to minimize the predicted cost of the executed action $\tilde{a}_t$, while avoiding degenerate solutions that collapse actions toward zero. Its loss function is given by:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{s_t\sim\mathcal{D},\, a_t\sim\pi(\cdot|s_t)}\Big[\,\beta \cdot Q_c(s_t, \tilde{a}_t) - \lambda \cdot \log \rho_\theta(s_t, a_t, \hat{c}_t)\,\Big]. \tag{6}$$

where $\beta, \lambda > 0$ are trade-off parameters. The first term encourages the regulator to scale down actions that lead to high predicted costs. However, without the second term, a trivial solution where $\rho_\theta(s, a, \hat{c}) \to 0$ would minimize this objective by collapsing all actions—halting the agent's behavior entirely. To counteract this, the second term acts as a *barrier penalty* that diverges as any element of the scaling vector approaches zero. It encourages the regulator to retain as much of the original action magnitude as possible, unless high predicted cost necessitates suppression.

Optimality Trade-Off. The regulator's training objective can be interpreted as solving a local constrained optimization problem at each state-action pair (s_t, a_t):

$$\min_{\rho\in(0,1]^d}\; \beta \cdot Q_c(s_t, \rho \odot a_t) - \lambda \cdot \log(\rho + \epsilon), \tag{7}$$

where the logarithm is applied element-wise to the scaling vector ρ, and we include ϵ to avoid instability as $\rho \to 0$, ensuring gradients remain well-defined during training.

The coefficients β and λ balance the trade-off between minimizing predicted cost and preserving action magnitude: larger λ encourages less suppression, while larger β prioritizes cost reduction. Since $Q_c(s, \rho \odot a_t)$ is typically a nonlinear function of the scaled action, the optimization problem lacks a closed-form solution but can be efficiently solved via gradient-based updates. This formulation ensures that the regulator selectively attenuates risky action dimensions while retaining as much of the agent's original behavior as possible.

Gradient Flow and Modularity. To ensure clean modularity, we detach the scaling weights $\rho_\theta(s_t, a_t, \hat{c}_t)$ from the computational graph when updating both the reward and cost critics, preventing gradients from flowing through the regulator. Similarly, the actor receives no gradients from the regulator, learning purely from task returns. Moreover, the regulator is updated independently via its own objective, ensuring that reward maximization and safety modulation remain decoupled.

This design is particularly well-suited for **off-policy reinforcement learning**, where updates are performed using transitions stored in a replay buffer, independent of the current policy. Since the regulator modulates actions *after* sampling from the policy $\pi_\phi(\cdot \mid s)$, the executed action $\tilde{a} = \rho_\theta(s, a, \hat{c}) \odot a$ differs from the originally sampled action a, and only the regulated action is stored and used for training. Off-policy methods naturally accommodate this, as policy and critic updates rely on the actual executed actions rather than the distribution used to generate them.

Implementation Details. The RL agent follows its baseline implementation without modification. The regulator and twin cost critics are feedforward neural networks with two hidden layers of 256 units and ReLU activations. The regulator outputs element-wise scaling factors $\rho \in (0, 1]^d$ via a sigmoid activation and is trained using the twin cost critics' predictions. The code is provided online[1].

5. Experiments

Our experiments are designed to achieve the following objectives: (i) compare our approach against state-of-the-art safe RL baselines across different dynamical systems, (ii) analyze the influence of the key hyperparameters (λ and β) from Eq. 7, which govern the trade-off between action preservation and cost suppression, (iii) evaluate the regulator's action-scaling mechanism through ablations such as element-wise versus scalar regulation on different systems, and (iv) study robustness under injected sensor and actuator noise, highlighting the method's potential for sim-to-real transfer.

Environments: We evaluate our method on locomotion tasks from the **Safety Gym** benchmark Ji et al. (2023), namely `Ant`, `Walker2d`, `Swimmer`, `HalfCheetah`, and `Humanoid`. In these velocity tasks, a safety cost is incurred whenever the center-of-mass speed exceeds a predefined threshold. Because the cost signal is sparse and triggered only by such threshold violations, these environments provide a challenging setting for safe RL, while naturally aligning with our regulator's goal of attenuating large actions that are most likely to induce violations.

Baselines: We compare our regulator against five state-of-the-art Safe RL baselines. **PPO-PID** Stooke et al. (2020) augments PPO with a PID-controlled Lagrangian multiplier to mitigate instabilities commonly observed in dual updates during constrained optimization. **Simmer** Sootla et al. (2022a) augments PPO with a safety state that tracks the remaining safety budget. **Safety Editor** Yu et al. (2022) uses two SAC agents, one for maximizing the reward and another for editing

1. https://github.com/HumanoidsBonn/Constraint-Aware-Adaptive-Action-Scaling

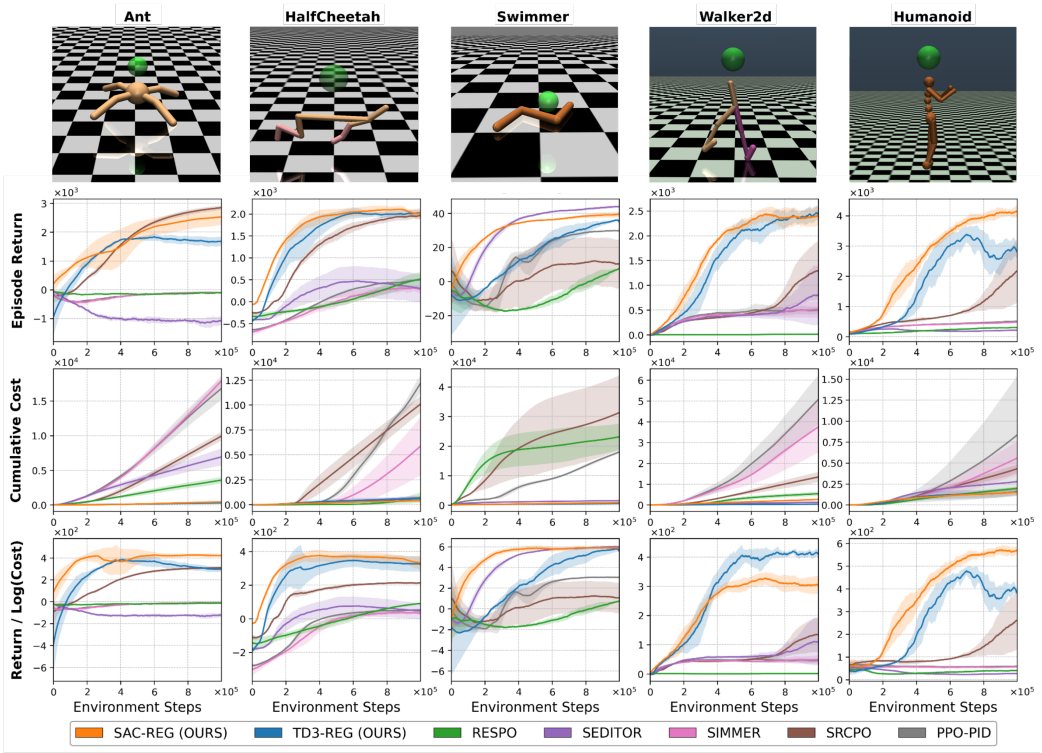

Figure 3: Performance comparison on the Safety Gymnasium locomotion environments. Each method is averaged over three independent runs; bold lines indicate the mean, and shaded areas show the standard deviation. Our methods (SAC-REG and TD3-REG) consistently achieve the best trade-off between return and cumulative constraint cost across all environments. **Top:** Episode return. **Middle:** Cumulative safety cost. **Bottom:** Return-to-Log-Cost ratio. Our methods outperform strong baselines, including PPO-PID Stooke et al. (2020), SIMMER Sootla et al. (2022a), SRCPO Kim et al. (2024), RESPO Ganai et al. (2023), and SEDITOR Yu et al. (2022). SIMMER is omitted from the `Swimmer` plot as it consistently yields negative returns, moving opposite to the target velocity.

unsafe actions. **RESPO** Ganai et al. (2023) estimates reachability sets and constrains policy to remain within safe regions. **SRCPO** Kim et al. (2024) formulates a bi-level constrained optimization using spectral risk measures to achieve a near-zero constraint violation rate while maximizing reward.

Metrics: Similar to prior safe RL studies, we report returns and cumulative costs as in Thananjeyan et al. (2021); Ganai et al. (2023); Goodall and Belardinelli (2024), and follow Thananjeyan et al. (2021) in using the return-to-cost (RC) ratio to capture the trade-off between task performance and safety. We measure (i) episodic return, (ii) cumulative cost during training, which reflects the total number of constraint violations and, in sparse-cost settings such as Safety Gym velocity tasks, implicitly captures violation frequency, and (iii) the RC ratio, defined as the total return divided by cumulative cost. For visualization, we plot the return divided by the logarithm of the accumulative cost, which improves interpretability of the safety-performance trade-off.

5.1. Comparison Against Baselines:

Safety Gym Results. Across the locomotion tasks in the Safety Gymnasium suite, our methods—**SAC-Regulator** and **TD3-Regulator**—consistently deliver strong task performance while substantially reducing safety violations. Each method was evaluated over three random seeds; bold lines in Fig.3 denote the mean return across runs, with shaded regions representing standard

Relative Return Improvement of SAC-Reg Over Baselines (↑)					
Method	Ant	HalfCheetah	Swimmer	Walker2d	Humanoid
PPO-PID Stooke et al. (2020)	27.33	3.04	0.32	3.76	7.13
SIMMER Sootla et al. (2022a)	26.14	6.09	1.21	3.59	7.60
SEditor Yu et al. (2022)	3.34	5.64	-0.10	1.95	18.52
RESPO Ganai et al. (2023)	23.97	2.89	0.17	204.40	12.73
SRCPO Kim et al. (2024)	-0.11	0.02	2.90	0.86	0.90
Relative Cost Compared to SAC-Reg (↓)					
PPO-PID Stooke et al. (2020)	39.29	28.39	21.31	19.11	5.46
SIMMER Sootla et al. (2022a)	41.88	13.70	54.22	14.14	3.64
SEditor Yu et al. (2022)	16.19	1.67	1.77	0.48	1.88
RESPO Ganai et al. (2023)	8.36	1.65	27.39	2.03	1.29
SRCPO Kim et al. (2024)	23.24	23.55	37.10	5.07	2.82

Table 1: Relative return improvement (↑) and relative cumulative cost (↓) of SAC-Reg compared to baselines across locomotion tasks. SAC-Reg consistently achieves higher returns and lower cumulative costs than prior safe RL methods across all environments.

Relative Return Improvement of TD3-Reg Over Baselines (↑)					
Method	SafetyAnt	HalfCheetah	Swimmer	Walker2d	Humanoid
PPO-PID Stooke et al. (2020)	18.49	3.05	0.17	3.87	4.51
SIMMER Sootla et al. (2022a)	17.70	6.11	1.18	3.70	4.83
SEditor Yu et al. (2022)	2.55	5.65	-0.20	2.02	12.22
RESPO Ganai et al. (2023)	16.26	2.90	3.59	209.23	8.30
SRCPO Kim et al. (2024)	-0.41	0.02	2.46	0.91	0.29
Relative Cost Compared to TD3-Reg (↓)					
PPO-PID Stooke et al. (2020)	60.14	19.77	34.06	126.18	5.20
SIMMER Sootla et al. (2022a)	64.11	9.54	86.65	93.37	3.47
SEditor Yu et al. (2022)	24.78	1.16	2.84	3.15	1.74
RESPO Ganai et al. (2023)	12.80	1.15	43.78	13.42	1.23
SRCPO Kim et al. (2024)	35.57	16.39	59.28	33.50	2.69

Table 2: Relative return improvement (↑) and relative cumulative cost (↓) of TD3-Reg compared to baselines across locomotion tasks.TD3-Reg demonstrates similar trends, outperforming baselines in return while maintaining substantially lower cumulative costs.

deviation. Compared to the baselines, our approach achieves higher or comparable episode returns, indicating that soft action scaling does not hinder exploration. Notably, **TD3-Regulator** achieves the greatest cost reductions, with up to **126×** lower cumulative cost in `Walker2d`, **64×** in `Ant`, and **86×** in `Swimmer`. Meanwhile, **SAC-Regulator** outperforms both **TD3-Regulator** and all baselines in `HalfCheetah` and `Humanoid`, achieving cost reductions of up to **28×** and **5×**, respectively. RESPO can reach comparable returns when trained for 9M steps, but only with substantially higher violations and failure to converge in `Humanoid`.

Tables 1 and 2 summarize results for both regulators against established baselines. Two metrics are reported: the *relative return improvement*,

$$\frac{\text{Return}_{\text{Ours}} - \text{Return}_{\text{Baseline}}}{|\text{Return}_{\text{Baseline}}|},$$

and the *relative cumulative cost* of each baseline normalized by our method. Positive return values indicate improved task performance, while cumulative cost ratios above 1.0 indicate higher constraint violations than ours. For example, in `Walker2d`, **SAC-Reg** outperforms RESPO with a relative return improvement of 204.4, corresponding to a 20,440% increase.

Overall, our methods deliver the lowest cumulative cost across all tasks without compromising return. Unlike approaches such as SEDITOR or RESPO, which improve safety at the expense of performance, our regulators preserve exploration and consistently achieve superior return-to-cost

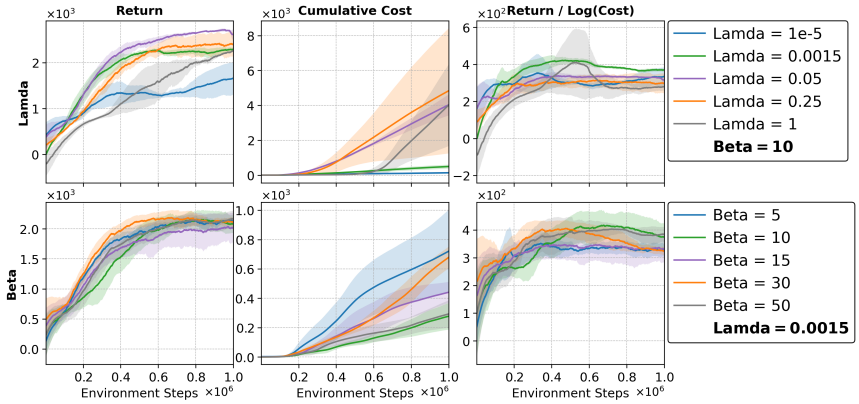

Figure 4: Ablation Study for evaluating the impact of the regulator hyperparameters λ and β on Return, Cumulative Cost, and Return-to-Cost ratio. Each curve shows the mean across three runs, and shaded regions indicate standard deviation. The top row varies λ with fixed $\beta = 10$; the bottom row varies β with fixed $\lambda = 0.0015$. As seen, smaller λ values reduce cumulative cost, with $\lambda = 0.0015$ giving the best balance between performance and safety, while $\beta = 10$ provides the most favorable trade-off overall.

trade-offs. This demonstrates the effectiveness and generality of decoupling safety regulation from reward learning.

Ablation Study on λ and β. We conduct an ablation study in the `Ant` environment to evaluate the sensitivity of our regulator framework to the hyperparameters λ and β, which control the trade-off between action retention and cost suppression. When varying λ over the range $\{1 \times 10^{-5},\ 0.0015,\ 0.05,\ 0.25,\ 1.0\}$, we find that smaller values lead to significantly lower cumulative costs. In particular, $\lambda = 0.0015$ achieves the best balance between constraint satisfaction and task performance. Higher values of λ result in larger action magnitudes and consequently higher constraint violations. Similarly, varying β over $\{5,\ 10,\ 15,\ 30,\ 50\}$ shows that $\beta = 10$ achieves the best overall safety-performance balance, minimizing constraint violations while maintaining high return. These results, as shown in Fig. 4, highlight the importance of properly tuning the regulator's loss coefficients to achieve optimal return-to-cost behavior.

Element-wise vs. Scalar Regulation: To evaluate the impact of element-wise action regulation, we conducted an ablation study comparing our full regulator with a simplified variant that uses a single scalar value to uniformly scale all action dimensions. Figure 5 presents results across all Safety Gymnasium locomotion tasks. While the scalar variant achieves comparable performance in most environments, it fails to converge in the high-dimensional `Humanoid` task. This suggests that element-wise scaling is particularly important in complex control settings, where individual action dimensions exhibit distinct risk profiles. Fine-grained modulation allows the regulator to target risky joints more precisely, improving both safety and learning stability.

Robustness and Sim-to-Real Transfer. To approximate uncertainties encountered on physical robots, we inject Gaussian noise into both observations and actions during training, modeling sensor measurement errors and actuator execution noise. Agents are trained with noise levels ($\sigma = 0, 0.025, 0.05, 0.10$), and the resulting training performance is summarized in Table 3. Across all noise settings, our regulator achieves strong returns while keeping cumulative costs bounded. Even under the highest noise level ($\sigma = 0.10$), performance remains stable, highlighting robustness to sensing and actuation imperfections and supporting the method's potential for sim-to-real transfer.

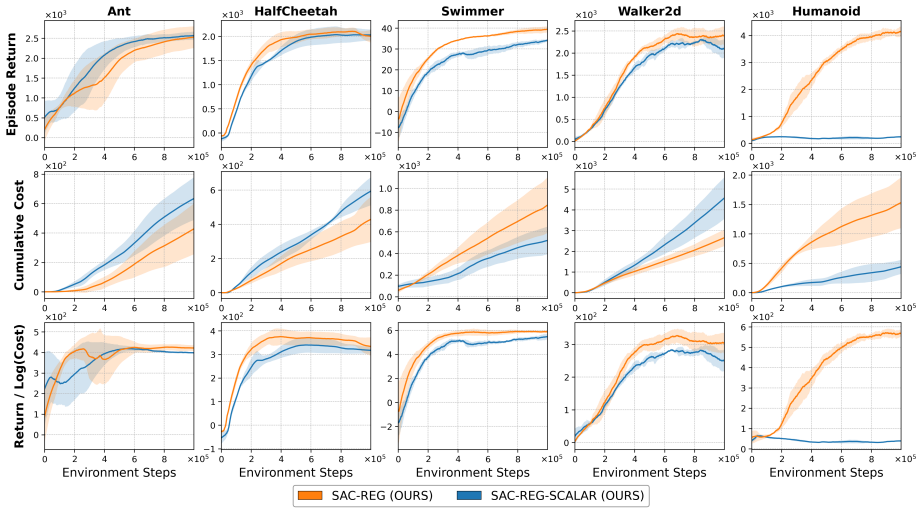

Figure 5: Comparison between element-wise and scalar action scaling. While both perform similarly in most tasks, the scalar variant fails to converge in the high-dimensional `Humanoid` environment, indicating that element-wise scaling improves safety and stability in complex control settings.

Step	Noise 0.00		Noise 0.025		Noise 0.05		Noise 0.10	
	Return	Cost	Return	Cost	Return	Cost	Return	Cost
100k	1961	6	753	0	762	28	677	1
300k	2558	79	1554	78	1614	165	1103	113
500k	2492	199	1945	195	1969	282	1381	387
700k	2533	282	2139	230	2071	339	1596	572
900k	2501	388	2104	262	2074	479	1589	767
1000k	2571	412	2016	312	2089	533	1510	849

Table 3: Training performance under different levels of injected Gaussian noise ($\sigma = 0.00, 0.025, 0.05, 0.10$) in observations and actions. Values show episode return and cumulative cost at checkpoints. Our regulator maintains bounded costs across noise levels, demonstrating robustness relevant for sim-to-real transfer.

6. Conclusion

We introduced a modular and practical framework for safe reinforcement learning that decouples reward maximization from safety enforcement through a cost-aware regulator. Instead of overriding agent actions, our method scales them smoothly based on predicted constraint violations, preserving exploration and enabling stable off-policy learning. The regulator uses twin cost critics for robust cost estimation and is trained with a loss that balances risk reduction and action preservation. Our approach is model-free and integrates seamlessly with existing off-policy RL pipelines. Empirical results on diverse benchmarks demonstrate that our method consistently achieves the highest return-to-cost ratios, reducing constraint violations by up to 126 times while maintaining or improving task performance relative to prior state-of-the-art methods. The regulator aligns with real-world safety limits such as torque bounds in manipulators, and joint load management in legged robots. Robustness experiments with injected observation and action noise further demonstrate bounded costs and stable returns under uncertainty, supporting the potential for sim-to-real transfer. A key direction for future work is to develop principled strategies for automatically tuning the regulator hyperparameters (λ and β) and to extend the approach beyond input-magnitude costs toward more general safety constraints.

Acknowledgments

M. Dawood, U. Ahmed Siddiquie, S. Khorshidi, and M. Bennwitz are with the Humanoid Robots Lab and the Center for Robotics, University of Bonn, Germany. Additionally, M. Dawood and M. Bennwitz are with the Lamarr Institute for Machine Learning and Artificial Intelligence. This work has partially been funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

Ahmed Agha, Baris Kayalibay, Atanas Mirchev, Patrick van der Smagt, and Justin Bayer. Exploring under constraints with model-based actor-critic and safety filters. In *8th Annual Conference on Robot Learning*, 2024.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European control conference (ECC)*, pages 3420–3431. Ieee, 2019.

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

Arko Banerjee, Kia Rahmani, Joydeep Biswas, and Isil Dillig. Dynamic model predictive shielding for provably safe reinforcement learning. *arXiv preprint arXiv:2405.13863*, 2024.

Bolun Dai, Rooholla Khorrambakht, Prashanth Krishnamurthy, Vinícius Gonçalves, Anthony Tzes, and Farshad Khorrami. Safe navigation and obstacle avoidance using differentiable optimization based control barrier functions. *IEEE Robotics and Automation Letters*, 8(9):5376–5383, 2023.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Murad Dawood, Sicong Pan, Nils Dengler, Siqi Zhou, Angela P Schoellig, and Maren Bennewitz. Safe multi-agent reinforcement learning for behavior-based cooperative navigation. *IEEE Robotics and Automation Letters*, 2025a.

Murad Dawood, Ahmed Shokry, and Maren Bennewitz. A dynamic safety shield for safe and efficient reinforcement learning of navigation tasks. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, Proceedings of Machine Learning Research. PMLR, 2025b.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36:69764–69797, 2023.

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

Alexander W Goodall and Francesco Belardinelli. Leveraging approximate model-based shielding for probabilistic safety guarantees in continuous environments. *arXiv preprint arXiv:2402.00816*, 2024.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

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*, pages 1861–1870. Pmlr, 2018.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36:18964–18993, 2023.

Dohyeong Kim, Taehyun Cho, Seungyub Han, Hojun Chung, Kyungjae Lee, and Songhwai Oh. Spectral-risk safe reinforcement learning with convergence guarantees. *arXiv preprint arXiv:2405.18698*, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

Mahmoud Selim, Amr Alanwar, Shreyas Kousik, Grace Gao, Marco Pavone, and Karl H Johansson. Safe reinforcement learning using black-box reachability analysis. *IEEE Robotics and Automation Letters*, 7(4):10665–10672, 2022.

Ziyad Sheebaelhamd, Konstantinos Zisis, Athina Nisioti, Dimitris Gkouletsos, Dario Pavllo, and Jonas Kohler. Safe deep reinforcement learning for multi-agent systems with continuous action spaces. *arXiv preprint arXiv:2108.03952*, 2021.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Aivar Sootla, Alexander Cowen-Rivers, Jun Wang, and Haitham Bou Ammar. Enhancing safe exploration using safety state augmentation. *Advances in Neural Information Processing Systems*, 35:34464–34477, 2022a.

Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pages 20423–20443. PMLR, 2022b.

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

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, pages 997–1005. PMLR, 2015.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3): 4915–4922, 2021.

Haozhe Tian, Homayoun Hamedmoghadam, Robert Shorten, and Pietro Ferraro. Reinforcement learning with adaptive regularization for safe control of critical systems. *arXiv preprint arXiv:2404.15199*, 2024.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft II using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

Haonan Yu, Wei Xu, and Haichao Zhang. Towards safe reinforcement learning with a safety editor policy. *Advances in Neural Information Processing Systems*, 35:2608–2621, 2022.

Qiyuan Zhang, Shu Leng, Xiaoteng Ma, Qihan Liu, Xueqian Wang, Bin Liang, Yu Liu, and Jun Yang. CVaR-constrained policy optimization for safe reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Xiao Zhang, Hai Zhang, Hongtu Zhou, Chang Huang, Di Zhang, Chen Ye, and Junqiao Zhao. Safe reinforcement learning with dead-ends avoidance and recovery. *IEEE Robotics and Automation Letters*, 9(1):491–498, 2023a.

Zhili Zhang, Songyang Han, Jiangwei Wang, and Fei Miao. Spatial-temporal-aware safe multi-agent reinforcement learning of connected autonomous vehicles in challenging scenarios. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 5574–5580, 2023b.