EPISODIC CONTROL FOR SAFE RL WITH GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy safe reinforcement learning (RL) methods provide strong stability and constraint guarantees but are sample-inefficient, while off-policy methods improve data efficiency at the cost of stronger assumptions to obtain guarantees, creating a fundamental tension in safety-critical settings. We introduce the Safe Integration Framework (SIF), a class of admissible critic corrections that inject bias into on-policy value estimation while preserving contraction of the Bellman operator and bounding deviation from the nominal fixed point. We further show that SIF preserves constrained trust-region optimization guarantees up to an explicit slack term, and yields primal-dual convergence to a bounded neighborhood of the nominal saddle point. As an instantiation, we propose Safe Proximity-guided Episodic Control (SPEC), a lightweight retrieval-based plugin that augments on-policy cost critics with uncertainty-aware corrections. Across six maximum-difficulty Safety-Gymnasium navigation tasks spanning the Button, Goal, and Circle families, SPEC applied to TRPO-PID yields a more consistent constraint profile than all baselines while avoiding reward collapse.

1 INTRODUCTION

Modern reinforcement learning (RL) has shown promising results in applications including but not limited to robotics and controls (Levine et al., 2016; 2017; Peng et al., 2020), autonomous navigation (Jaritz et al., 2018; Yang et al., 2023; Zheng & Gu, 2025), game playing and decision-making (Mnih et al., 2015; Silver et al., 2016; Schrittwieser et al., 2020), healthcare (Yu et al., 2020), recommendation systems and online platforms (Huang et al., 2026; Kim & Kim, 2021; Li et al., 2021), natural language processing (NLP) and dialogue systems (Li et al., 2016; Brantley et al., 2025; Sun et al., 2025). Today's RL algorithms maximize reward through interaction with an environment, an assumption that is often unrealistic in certain real-world settings, where safety is critical and high-reward actions may be dangerous. Safe RL is formalized as requiring the agent to satisfy constraints on costs while maximizing reward (Altman, 1999; Dulac-Arnold et al., 2021).

On-policy local policy search methods have emerged as a strong foundation for safe RL, demonstrating competitive performance on continuous-control safety navigation and robotic locomotion tasks (Zhang et al., 2020; Yang et al., 2020; 2022; Achiam et al., 2017). While these approaches are typically less data efficient, they avoid several limitations of off-policy safe RL methods, which often rely on restrictive assumptions, such as Gaussian cost distributions or convexity in policy space, that are rarely satisfied in practice (Kim & Oh, 2022; Yang et al., 2021; Kim et al., 2023; Wu et al., 2024; Chen et al., 2022; Ding & Lavaei, 2023). In contrast, on-policy local policy search methods admit cleaner theoretical guarantees under milder conditions. They provide a principled and empirically effective framework for enforcing safety through trust-region updates and recovery/projection-based constraints (Achiam et al., 2017; Yang et al., 2020; Zhang et al., 2020; Yang et al., 2022), as well as through Lagrangian formulations based on primal–dual updates (Boyd & Vandenberghe, 2004; Ray et al., 2019; Stooke et al., 2020). This motivates a central question: can off-policy information improve the data efficiency of on-policy safe RL without compromising guarantees?

The closest line of work addressing this question is episodic control (EC), which leverages non-parametric memory for improved stability and data efficiency in RL (Pritzel et al., 2017), inspired by findings in cognitive sciences (Lengyel & Dayan, 2007). Due to their reliance on state–action value functions and their inherently off-policy nature, EC methods are typically implemented with off-policy algorithms (Pritzel et al., 2017; Hansen et al., 2018; Lin et al., 2018). While alternative approaches such as reward shaping are compatible with on-policy methods by modifying the immediate learning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

signal, they generally do not guarantee policy improvement in RL settings (Li et al., 2023). Moreover, current EC methods are largely developed for unconstrained RL, with only limited extensions to safety (e.g., Li et al. (2025)) and few works providing theoretical guarantees of policy performance or convergence (e.g., Hu et al. (2021)), both of which are essential in safe RL. Importantly, these constitute the only existing approaches in each direction, and no method achieves both simultaneously. A detailed discussion of related work is provided in Appendix A.

Despite EC's benefits, additional training information can introduce bias and violate convergence or performance guarantees. Thus, our contributions are threefold: (1) We propose the **Safe Integration Framework (SIF)**, a general theory of admissible critic corrections that preserves the policy performance guarantees of constrained trust-region methods and the local convergence structure of on-policy primal–dual dynamics under bounded critic perturbations. (2) As one instantiation of SIF, we propose **Safe Proximity-guided Episodic Control (SPEC)**, a practical retrieval-based plugin that integrates off-policy information through uncertainty-aware scaling and selective observation stacking mechanisms. (3) We empirically demonstrate that SPEC improves safety constraint satisfaction consistency on challenging navigation tasks while avoiding reward collapse.

2 PRELIMINARIES

Constrained Markov Decision Processes (CMDPs). A CMDP (Altman, 1999) is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \{c_i\}_{i=1}^m, \gamma \rangle$ extending the standard MDP (Sutton & Barto, 1998; Bellman, 1957) with cost functions $c_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and per-constraint thresholds d_i. A stationary, stochastic policy $\pi_\theta(a \mid s) \in \Pi_\theta$ induces the discounted state-visitation distribution $d^{\pi_\theta}(s) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \mid s_0, \pi_\theta)$ with $s_0 \sim \rho_0$. We write $J_r(\pi_\theta) := \mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta}[r(s,a)]$ and analogously $J_{c_i}(\pi_\theta)$. The safe-RL objective is $\max_{\pi_\theta} J_r(\pi_\theta)$ s.t. $J_{c_i}(\pi_\theta) \le d_i$ for all $i \in \{1, \dots, m\}$. Value, action-value, and advantage functions $V_r^{\pi_\theta}$, $Q_r^{\pi_\theta}$, $A_r^{\pi_\theta} := Q_r^{\pi_\theta} - V_r^{\pi_\theta}$ are defined as usual, with cost analogues $V_{c_i}^{\pi_\theta}, Q_{c_i}^{\pi_\theta}, A_{c_i}^{\pi_\theta}$. The one-step temporal-difference (TD) residual $\delta_t^{V_r^{\pi_\theta}} := r(s_t,a_t) + \gamma V_r^{\pi_\theta}(s_{t+1}) - V_r^{\pi_\theta}(s_t)$ yields the generalized advantage estimator (GAE) (Schulman et al., 2016) $A_{r,\lambda,t}^{\pi_\theta}(s_t,a_t) := \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \, \delta_{t+\ell}^{V_r^{\pi_\theta}}$, with cost analogues defined identically.

The standard Bellman operator is a γ-contraction; the λ-return induces a generalized Bellman equation

$$V_{r,\lambda}^{\pi_\theta}(s) \;=\; (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1}\, \mathbb{E}_{a_t\sim\pi_\theta,\, s_{t+1}\sim P}\left[\sum_{t=0}^{n-1} \gamma^t r(s_t,a_t) + \gamma^n V_{r,\lambda}^{\pi_\theta}(s_n) \,\middle|\, s_0 = s\right],$$

whose operator is a $\tilde{\gamma}$-contraction with $\tilde{\gamma} := \frac{\gamma(1-\lambda)}{1-\gamma\lambda} < 1$, and fixed point V^{π_θ} (Prop. C.1)[1].

Constrained trust-region bounds. Constrained trust-region policy search solves

$$\begin{aligned} \pi_\theta \;&=\; \arg\max_{\pi_{\theta_k}\in\Pi_\theta} \; \mathbb{E}_{s\sim d^{\pi_{\theta'}},\, a\sim\pi_{\theta_k}(\cdot|s)}[A_r^{\pi_{\theta'}}(s,a)] \\ \text{s.t. }\; & J_{c_i}(\pi_{\theta'}) + \tfrac{1}{1-\gamma}\,\mathbb{E}_{s\sim d^{\pi_{\theta'}},\, a\sim\pi_{\theta_k}(\cdot|s)}\big[A_{c_i}^{\pi_{\theta'}}(s,a)\big] \le d_i,\ \forall i, \\ & \bar{D}_{\mathrm{KL}}(\pi_{\theta_k} \,\|\, \pi_{\theta'}) \le \delta, \end{aligned} \tag{1}$$

where $\pi_{\theta'}$ is a reference policy, $\bar{D}_{\mathrm{KL}}(\pi_{\theta_k} \,\|\, \pi_{\theta'}) := \mathbb{E}_{s\sim d^{\pi_{\theta_k}}}[D_{\mathrm{KL}}(\pi_{\theta_k} \,\|\, \pi_{\theta'})[s]]$, and δ is the trust-region step size (Achiam et al., 2017; Yang et al., 2020). Now let $d_\lambda^{\pi_\theta}$ denote the λ-state-visitation distribution of π_θ, and let $\epsilon_{r,t}^{\pi_\theta,\pi_{\theta'}} := \sup_{s_t\in\mathcal{S}} \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t),\, s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\left|\delta_t^{V_r^{\pi_{\theta'}}}\right|\right]$ and $\epsilon_r^{\pi_\theta,\pi_{\theta'}} := \sup_{t\in\mathbb{N}} \epsilon_{r,t}^{\pi_\theta,\pi_{\theta'}}$ denote the worst-case expected absolute TD residual of $V_r^{\pi_{\theta'}}$ under actions from π_θ, with $\epsilon_{c_i}^{\pi_\theta,\pi_{\theta'}}$ defined analogously. Yang et al. (2022) formalized the policy-performance bounds under $A_{r,\lambda}^{\pi_\theta}$ as an estimator of $A_r^{\pi_\theta}$, assuming a feasible reference policy $\pi_{\theta'}$ with $J_{c_i}(\pi_{\theta'}) \le d_i$ for all i. Defining $C \;=\; \frac{2\tilde{\gamma}\left(\gamma\lambda(|\mathcal{S}|-1)+1\right)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}$, the expected-return difference between successive policies admits the lower bound

$$J_r(\pi_\theta) - J_r(\pi_{\theta'}) \;\ge\; \frac{1}{1-\tilde{\gamma}}\,\mathbb{E}_{s\sim d_\lambda^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\Big[A_{r,\lambda}^{\pi_{\theta'}}(s,a) - C\,\epsilon_r^{\pi_\theta,\pi_{\theta'}}\Big]. \tag{2}$$

[1] *Notation convention:* When the subscript r or c_i (or, later, c) is omitted, the quantity refers generically to either the reward or cost variant. A complete notation summary appears in Appendix B

Adapting Achiam et al. (2017)'s cost bound under $A^{\pi_\theta}_{r,\lambda}$ yields $J_{c_i}(\pi_\theta) \le d_i + C\,\epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i}$. Several algorithms operationalize this template—CPO (Achiam et al., 2017), PCPO (Yang et al., 2020), FOCOPS (Zhang et al., 2020), CUP (Yang et al., 2022)—differing in whether the update is first- or second-order and whether constraints are enforced via projection or via KL-coefficient hyperparameters. *Convention:* We classify hybrid methods (e.g., TRPO-Lag, TRPO-PID) as Lagrangian methods, since their safety guarantees arise from dual updates rather than per-iteration trust-region constraint enforcement.

Lagrangian methods. Replacing the constrained update with the saddle-point problem

$$\max_{\theta}\ \min_{\mu\ge 0}\ \mathcal{L}(\pi_\theta,\mu), \qquad \mathcal{L}(\pi_\theta,\mu) \;=\; J_r(\pi_\theta) - \textstyle\sum_i \mu_i\big(J_{c_i}(\pi_\theta) - d_i\big), \tag{3}$$

(Boyd & Vandenberghe, 2004; Ray et al., 2019) yields a primal–dual scheme with $\mu_i \ge 0$. Let $z := (\theta, \mu)$ and define the mean drift

$$g(z) := \begin{bmatrix} g_\theta(\theta,\mu) \\ g_\mu(\theta,\mu) \end{bmatrix} \;=\; \begin{bmatrix} \mathbb{E}_{s\sim d^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \mathcal{L}(\pi_\theta,\mu;s,a)\big] \\ -\,\mathbb{E}_{s\sim d^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\big[\nabla_\mu \mathcal{L}(\pi_\theta,\mu;s,a)\big] \end{bmatrix}.$$

The discrete recursion $z_{k+1} = z_k + \alpha_k\, g(z_k)$ converges under standard step-size assumptions to a local saddle point $z^\star$ characterized by $g(z^\star) = 0$, but *without* per-iteration monotonic bounds on policy improvement or constraint violation. PID-Lagrangian variants (Stooke et al., 2020) apply proportional–integral–derivative updates to the constraint residuals to stabilize the dual dynamics, inheriting the same asymptotic guarantees. The resulting policy gradient's advantage, $A^{\pi_\theta}_\mu(s,a)$, is given by a weighted combination of reward and cost advantages, $A^{\pi_\theta}_\mu(s,a) = A^{\pi_\theta}_r(s,a) - \sum_i \mu_i A^{\pi_\theta}_{c_i}(s,a)$, which corresponds to the advantage of the scalarized Lagrangian return. When advantages are estimated via $A^{\pi_\theta}_{r,\lambda}(s,a)$ and $A^{\pi_\theta}_{c_i,\lambda}(s,a)$, the same asymptotic guarantees are preserved, since the λ-return operator shares the same fixed point as the Bellman operator (Prop. C.1).

3 Safe integration framework (SIF)

This section develops the Safe Integration Framework (SIF), a class of admissible critic corrections for on-policy actor–critic methods that preserve policy performance guarantees in constrained trust-region methods and local convergence guarantees in primal–dual methods relying on critic-based value estimation. Modifying critic updates by incorporating additional information—such as off-policy or non-parametric signals—can, in principle, introduce bias and potentially violate convergence or safety guarantees. SIF specifies the conditions under which such corrections do not. SIF augments the standard λ-return Bellman evaluation operator $\mathcal{T}^{\pi_\theta}_\lambda$ (Prop. C.1) with a clipped additive correction:

$$\big(\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V\big)(s) \;:=\; \big(\mathcal{T}^{\pi_\theta}_{\lambda} V\big)(s) \;+\; \mathrm{clip}_\eta\big(\beta\,\Delta_V(s)\big), \tag{4}$$

such that $V^{\pi_\theta}_{\lambda,\beta}$ denotes the fixed point of $\mathcal{T}^{\pi_\theta}_{\lambda,\beta}$ when $\tilde\gamma + \beta L_\Delta < 1$, where L_Δ denotes the Lipschitz constant of the correction term $\Delta_V : \mathcal{S} \to \mathbb{R}$. Here, Δ_V is a generic correction signal that may depend on the current value function V, $\beta \ge 0$ controls the correction magnitude, and $\mathrm{clip}_\eta(\cdot)$ bounds the correction elementwise to $[-\eta, \eta]$. We treat β as a scalar throughout for simplicity; when β is state-dependent (Section 4), all results hold pointwise if $\beta(s) \le \beta_{\max}$. All proofs and additional technical results are deferred to the appendix.

The Bellman equation. The Bellman equation provides the foundation for analyzing the stability of the proposed critic correction. We consider the expected operator induced by the target in Eq. equation 4. Recall that the standard TD(λ) target corresponds, in expectation, to application of the λ-return Bellman evaluation operator $\mathcal{T}^{\pi_\theta}_\lambda$, which is a $\tilde\gamma$-contraction under the supremum norm (Proposition C.1). The correction introduced in Eq. equation 4 enters as a clipped residual term of the form $\mathrm{clip}_\eta\big(\beta\Delta_V\big)$, where Δ_V denotes a correction that may depend on the current value function.

Because clipping is 1-Lipschitz and uniformly bounded, the sensitivity of the correction term to perturbations in V is governed entirely by the dependence of Δ_V on V. In particular, we assume that Δ_V is L_Δ-Lipschitz under the supremum norm, i.e., $\|\Delta_V - \Delta_W\|_\infty \le L_\Delta \|V - W\|_\infty$, so that the overall correction is (βL_Δ)-Lipschitz in V. Under this condition, the augmented operator can

be interpreted as a perturbed Bellman evaluation operator whose deviation from $\mathcal{T}_\lambda^{\pi_\theta}$ is explicitly controlled by βL_Δ. Thus, Theorem D.2 shows that this perturbation preserves contractivity, with the contraction modulus increased by at most βL_Δ. Unlike prior EC works that establish critic convergence (Hu et al., 2021; Zhu* et al., 2020), we show that the fixed point induced by the SIF critic remains bounded to the true Bellman fixed point, with hyperparameter-controlled deviation.

Theorem 3.1 (SIF-induced deviation bound for the Bellman fixed point)**.** *Fix a stationary stochastic policy π_θ and assume $\tilde{\gamma} + \beta L_\Delta < 1$. Let V^{π_θ} denote the unique fixed point of $\mathcal{T}_\lambda^{\pi_\theta}$ (Proposition C.1), and let $V_{\lambda,\beta}^{\pi_\theta}$ denote the unique fixed point of $\mathcal{T}_{\lambda,\beta}^{\pi_\theta}$ (Theorem D.2). Then*

$$\|V_{\lambda,\beta}^{\pi_\theta} - V^{\pi_\theta}\|_\infty \le \frac{\|\mathrm{clip}_\eta\big(\beta\,\Delta_{V^{\pi_\theta}}\big)\|_\infty}{1-(\tilde{\gamma}+\beta L_\Delta)}.$$

Constrained trust-region local policy search. When combined with the SIF critic, trust-region local policy search methods retain analogous performance guarantees provided that the induced value estimation error is uniformly controlled. While contraction of the corrected Bellman operator ensures convergence to a unique fixed point, this fixed point may deviate from the true on-policy value function under perturbations. By explicitly bounding this deviation via the critic error $\|V_{\lambda,\beta}^{\pi_\theta} - V^{\pi_\theta}\|_\infty$, we ensure that the resulting advantage estimates remain sufficiently accurate. This allows the standard trust-region performance difference bounds to be extended to the SIF setting, yielding a modified bound that incorporates both the critic-induced bias and the trust-region constraint (Remark E.6).

Theorem 3.2 (SIF-induced reward performance difference lower bound)**.** *Fix a stationary reference policy $\pi_{\theta'}$ and a candidate policy π_θ. Assume $\tilde{\gamma} + \beta L_\Delta < 1$ and let $V_{\lambda,\beta}^{\pi_{\theta'}}$ denote the unique fixed point of the SIF critic operator for $\pi_{\theta'}$. Let $A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a)$ denote the λ-return reward advantage computed using $V_{r,\lambda,\beta}^{\pi_{\theta'}}$ (equivalently, using TD residuals induced by the SIF fixed point). Then the reward performance difference satisfies*

$$J_r(\pi_\theta) - J_r(\pi_{\theta'}) \;\ge\; \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d_\lambda^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\Bigg[A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a) \;-\; \frac{(1+\gamma)\eta}{(1-\gamma\lambda)\big(1-(\tilde{\gamma}+\beta L_\Delta)\big)} \\ - \frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\left(\epsilon_r^{\pi_\theta,\pi_{\theta'}} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}\right)\Bigg].$$

Corollary 3.3 (SIF-induced worst-case cost violation bound)**.** *Suppose the reference policy $\pi_{\theta'}$ is feasible for constraint i, i.e., $J_{c_i}(\pi_{\theta'}) \le d_i$. Then, under the same assumptions as Theorem 3.2, the cost of the candidate policy satisfies*

$$J_{c_i}(\pi_\theta) \;\le\; d_i \;+\; \frac{1}{1-\tilde{\gamma}}\Bigg\{\left(1+\frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\right)\left(\epsilon_{c_i}^{\pi_\theta,\pi_{\theta'}} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}\right) \\ + \frac{(1+\gamma)\eta}{(1-\gamma\lambda)\big(1-(\tilde{\gamma}+\beta L_\Delta)\big)}\Bigg\}.$$

Lagrangian methods. Lagrangian methods can be analyzed by viewing their primal–dual updates as discretizations of an underlying mean ordinary differential equation (ODE) with an exponentially stable saddle point. We assume the existence of a neighborhood $\mathcal{U}$ of the saddle point $z^\star$ such that the mean drift g is continuously differentiable on $\mathcal{U}$, satisfies $g(z^\star) = 0$, and has Jacobian $J^\star := \nabla g(z^\star)$ that is Hurwitz. This local stability condition is standard in ODE-based analyses of stochastic approximation, actor–critic, and primal–dual dynamics (Konda & Tsitsiklis, 2003; Zhang & Zheng, 2006). We further assume a bounded score function $|\nabla_\theta \log \pi_\theta(a \mid s)| \le G$, which controls policy-gradient updates, and a bounded dual variable $0 \le \mu \le M_\mu$, a standard condition in constrained policy optimization and primal–dual CMDP analyses used to control multiplier dynamics (Tessler et al., 2019; Ding et al., 2020). Collectively, these are standard local regularity and stability assumptions in analyses of actor–critic and primal–dual policy optimization (Konda & Tsitsiklis, 2003; Zhang & Zheng, 2006; Tessler et al., 2019; Ding et al., 2020) (Remark F.1), and differ in nature from the bounded-stochasticity and smooth cost-signal assumptions used in prior off-policy and episodic-control safe RL methods (Appendix A). Under these conditions, the nominal dynamics admit a locally exponentially stable equilibrium. The SIF critic introduces a uniformly bounded perturbation

to this drift, yielding a perturbed ODE whose trajectories converge to a neighborhood of $z^\star$. The radius of this neighborhood scales linearly with the critic-induced perturbation, as characterized below. Thus, SIF preserves the local stability and convergence structure of Lagrangian methods up to a controllable bias, supporting robust performance and constraint satisfaction under value estimation.

Corollary 3.4 (SIF-induced saddle-point neighborhood)**.** *Under the assumptions of Theorem F.4 and Lemma F.3, there exist constants $a_1, a_2, G > 0$ such that, for all trajectories initialized sufficiently close to $z^\star$, define, for $x \in \{r, c\}$,*

$$\|e_x\|_\infty := \frac{\eta_x}{1-(\tilde{\gamma}+\beta_x L_{\Delta,x})}, \qquad E_x := \frac{1}{1-\gamma\lambda}\left((1+\gamma)\|e_x\|_\infty + (2-\gamma\lambda)\epsilon_x^{\pi_\theta,\pi_\theta}\right).$$

Then, for all trajectories initialized sufficiently close to $z^\star$,

$$\limsup_{t\to\infty} \|z_\beta(t) - z^\star\| \le \frac{Ga_1}{a_2}\left(E_r + M_\mu E_c\right).$$

Together, these results give admissible critic corrections that preserve performance, and local convergence guarantees of on-policy constrained trust-region and primal–dual methods. Under boundedness and Lipschitz assumptions, SIF enables auxiliary information integration with controlled perturbation.

4 SAFE PROXIMITY-GUIDED EPISODIC CONTROL (SPEC)

Building on SIF, we propose **Safe Proximity-guided Episodic Control (SPEC)** as one instantiation of the framework; other methods may instantiate SIF differently as long as they satisfy its conditions. Within this design space, SPEC applies retrieval-based corrections only to the cost critic, although SIF itself permits corrections to both reward and cost critics. In many safety benchmark environments, cost signals are significantly sparser than rewards (Ji et al., 2023; Gronauer, 2022), leading to higher-variance estimates. SPEC therefore prioritizes the cost critic, using retrieval-based modulation to stabilize updates and improve constraint satisfaction, with *the goal of achieving a more consistent cost profile and reward–cost trade-off across diverse tasks rather than uniform dominance per task.* SPEC also introduces several SIF-independent mechanisms that provide initial design ideas for improving EC methods more broadly, both within SIF and beyond this particular instantiation.

Figure 1 illustrates the SPEC pipeline. SPEC maintains a non-parametric memory as a key–value lookup table and performs retrieval in a learned embedding space induced by an encoder $f_c(s)$, jointly trained with the cost critic $\hat{V}_{c,\lambda}^{\pi_\theta}(s)$. The critic operates on $f_c(s)$ rather than raw states (we write $V(s)$ for simplicity), and f_c is frozen after an initial warmup phase for stability. We also maintain a reward encoder $f_r(s)$ without retrieval-based correction, restricting retrieval to the cost pathway while preserving architectural symmetry. We present SPEC for a single scalar cost c, as is standard in safe RL with aggregated constraints. In contrast to SafeEC (Li et al., 2025) which relies on explicit state abstraction, we preserve fine-grained state information by learning representations aligned with value estimation. This coupling ensures the embedding space captures policy- and constraint-relevant structure, enabling policy-consistent and safety-aware retrieval. For scalability, we maintain a bounded replay buffer $\mathcal{D}_c$ as a rolling retriever that retains recent samples.

Given a query state s, we retrieve the k nearest neighbors using the embedding $f_c(s)$ and their associated rollout trajectories τ_c to construct a neighborhood value estimate. Specifically, we compute an n-step return estimate $V_c^{\mathrm{nb}}(s)$ by averaging multi-step returns from retrieved trajectories to form a non-parametric local value estimate in embedding space. We incorporate this estimate into the cost critic through a bounded residual correction. Let $V_{c,\lambda}^{\mathrm{targ}}(s)$ denote the sample-based λ-return critic target, distinct from the true value function $V_{c,\lambda}^{\pi_\theta}$. Define the residual $\Delta_c(s) := V_c^{\mathrm{nb}}(s) - \hat{V}_{c,\lambda}^{\pi_\theta}(s)$, which measures the discrepancy between the neighborhood estimate and the current critic prediction. We then define the adjusted target

$$V_{c,\lambda,\beta}^{\mathrm{targ}}(s) = V_{c,\lambda}^{\mathrm{targ}}(s) + \mathrm{clip}_\eta(\beta(s)\,\Delta_c(s)), \tag{5}$$

where $\beta(s)$ is a state-dependent scaling coefficient controlling the correction strength based on retrieval reliability. The cost critic is optimized toward $V_{c,\lambda,\beta}^{\mathrm{targ}}(s)$ with mean-squared error.

The neighbor value is computed via a truncated n-step return with bootstrap from the current cost critic, $\hat{V}_{c,\lambda,\beta}^{\pi_\theta}$. For each query, we average over the k retrieved neighbors to obtain V_c^{nb}. We cache V_c^{nb}

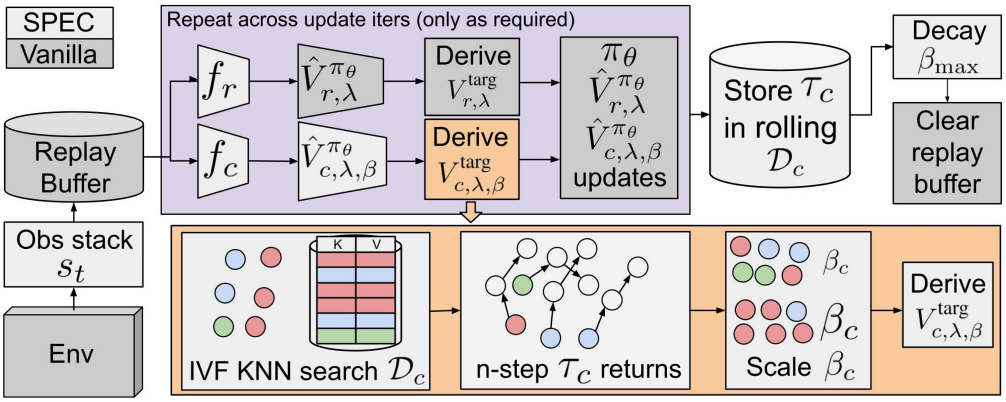


Figure 1: SPEC pipeline after warmup. Encoded observations query a rolling buffer $\mathcal{D}_c$ via IVF-based k-NN retrieval. Retrieved trajectories are incorporated into the cost critic through bounded, scaled corrections, yielding adjusted training targets.

and $\beta(s)$ once per rollout collection phase, so both remain fixed during critic optimization, avoiding repeated retrieval and matching the assumption that the neighbor target is exogenous within each update (Algorithm 1).

Remark 4.1. *Note that SPEC's L_Δ-Lipschitz condition concerns perturbations of the critic value function, not changes in the embedding space or retrieval mapping. During each critic-update phase, the policy, encoder, and retrieved neighborhood are fixed, and new experiences are added only after the critic and policy updates. Thus, $V^{\text{nb}}(s)$ is fixed during Bellman operator evaluation. For $\Delta_V(s) = V^{\text{nb}}(s) - \hat{V}(s)$, and any two critic estimates $\hat{V}_1$ and $\hat{V}_2$,*

$$|\Delta_{V_1}(s) - \Delta_{V_2}(s)| = \left|\left(V^{\text{nb}}(s) - \hat{V}_1(s)\right) - \left(V^{\text{nb}}(s) - \hat{V}_2(s)\right)\right| = \left|\hat{V}_1(s) - \hat{V}_2(s)\right|.$$

Taking the supremum over states gives

$$\|\Delta_{V_1} - \Delta_{V_2}\|_\infty \le \|\hat{V}_1 - \hat{V}_2\|_\infty,$$

so Δ_V is 1-Lipschitz with respect to the critic, or, $L_\Delta = 1$. Although k-NN retrieval may be discontinuous with respect to the embedding, this does not affect the Bellman operator analysis because the retrieved neighborhood remains fixed within each critic-update phase.

Adaptive corrections. We also propose an adaptive correction to retrieved experiences using inverted file (IVF) indices for efficient nearest-neighbor retrieval and introduce an entropy-aware weighting to account for uncertainty in local structure. An IVF index uses K-Means to coarsely quantize the embedding space and restricts nearest-neighbor search to a small set of cells for faster search (Johnson et al., 2021). For a query i, let $z_{i,1}, \ldots, z_{i,k}$ denote the IVF cluster assignments of its k retrieved neighbors. Define the empirical cell distribution $p_{i,c} = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}\{z_{i,j} = c\}$ and its entropy $H_i = -\sum_c p_{i,c} \log p_{i,c}$. The normalized score and resulting retrieval weight are

$$g^{(i)}_{\text{ent}} = 1 - \frac{H_i}{\log k}, \qquad \beta = \beta_{\max} \frac{1}{1 + \exp(-10\,(g_{\text{ent}} - 0.5))}. \tag{6}$$

This adaptively scales the retrieval residual, increasing corrections in coherent regions and suppressing them in diffuse ones. This approach accounts for local uncertainty in the embedding space, where higher entropy over IVF cluster assignments indicates retrieval from a more heterogeneous set of states. $\beta_{\max}$ is linearly decayed to zero over the course of training, gradually reducing reliance on retrieval-based corrections as the cost critic converges. Since IVF indices rely on learned clustering, we perform an encoder warmup of N_{enc} steps before freezing f_c, followed by a retriever warmup of N_{retr} steps to train $\mathcal{D}_c$. Retrieval is initiated after $N_{\text{enc}} + N_{\text{retr}}$ timesteps.

Observation stacking. Lastly, we propose observation stacking with only environment history by decomposing each observation as $o_t = (o_{\text{env},t}, o_{\text{agent},t})$, where $o_{\text{env},t}$ captures exogenous environment features (e.g., scene observations), while $o_{\text{agent},t}$ represents the agent's internal state (e.g.,

Algorithm 1 SPEC-plugin for Local Policy Search

Require: Policy π_θ, reward critic $\hat{V}_{r,\lambda}^{\pi_\theta}$, cost critic $\hat{V}_{c,\lambda,\beta}^{\pi_\theta}$, cost encoder f_c, cost index $\mathcal{D}_c$, bounds $\beta_{\max}$, η, n_{step}

Rollout π_θ with stacked observations to collect $\{(s_t, a_t, r_t, c_t, s_{t+1})\}_{t=0}^{B-1}$.
Compute standard on-policy targets $V_{r,\lambda}^{\text{targ}}$ and $V_{c,\lambda}^{\text{targ}}$.
if retrieval has started **then**
    Retrieve k cost neighbors from $\mathcal{D}_c$ using $f_c(s_t)$.
    Compute and cache $V_c^{\text{nb}}(s_t)$ and $\beta(s_t)$
end if
Update π_θ ▷ policy update follows the base algorithm
for each critic optimization iteration **do**
    Set $y_r = V_{r,\lambda}^{\text{targ}}$.
    if retrieval has started **then**
        $y_c = V_{c,\lambda}^{\text{targ}} + \text{clip}_\eta\Big(\beta(s_t)\,\Big(V_c^{\text{nb}}(s_t) - \text{stopgrad}(\hat{V}_{c,\lambda,\beta}^{\pi_\theta}(s_t))\Big)\Big)$.
    else
        $y_c = V_{c,\lambda}^{\text{targ}}$.
    end if
    Update $\hat{V}_{r,\lambda}^{\pi_\theta}$ and $\hat{V}_{c,\lambda,\beta}^{\pi_\theta}$ by MSE on y_r and y_c.
end for
Add collected experiences to $\mathcal{D}_c$ once f_c is frozen after encoder warmup.
Anneal $\beta_{\max}$ if enabled.

actuator states). For a history length H, the stacked input is $s_t = \big[o_{\text{env},t-H+1}, \ldots, o_{\text{env},t-1}, o_t\big]$, where past environment features are retained while the current observation is kept in full, with the earliest observation repeated when fewer than H past observations are available. We use environment-only history instead of full stacking to avoid repetition of o_{agent}, reducing input dimensionality while preserving the environmental dynamics we believe are most relevant for retrieval. Details on the decomposition are in Appendix H. We write s_t for the (stacked) state throughout the paper.

5 SPEC Experiments

Experiments are conducted on Safety-Gymnasium (Ji et al., 2023) Button, Circle, and Goal tasks at level 2, using the OmniSafe (Ji et al., 2024) framework (full environment details in Appendix H). We evaluate constrained policy optimization methods, including CPO (Achiam et al., 2017), PCPO (Yang et al., 2020), CUP (Yang et al., 2022), FOCOPS (Zhang et al., 2020), TRPO-Lag and PPO-Lag (Schulman et al., 2015; 2017; Ray et al., 2019), and TRPO-PID and CPPO-PID (Stooke et al., 2020). For brevity, we denote PID as P (e.g., TRPO-P), Lag as L, and use S to denote SafeEC when paired with a base algorithm (e.g., S-TRPO-P). Based on Appendix I.1, TRPO-P exhibits the most consistent empirical performance across environments, outperforming CPO and TRPO-L in terms of stability, while the remaining methods show comparatively weaker and less reliable results. We therefore instantiate SPEC and SafeEC (Li et al., 2025) on TRPO-P. For SafeEC, we use the hyperparameters reported in the original public repository and reimplement the method within OmniSafe. In the main results, we focus on CPO, TRPO-L, and TRPO-P, along with S-TRPO-P and SPEC-TRPO-P (ours), restricting attention to the strongest-performing baselines. We defer full results for all methods to Appendix I.1. For SPEC, we set $\eta = 2$ and $\beta_{\max} = 0.15$, stack observations over 3 timesteps; all hyperparameters are provided in Appendix I.2. Following standard deep-RL practice, we report IQM and standard deviation across five seeds, with IQM providing robust central-tendency estimates at modest seed counts (Agarwal et al., 2021).

Main experiments. As shown in Figure 4 and Table 1, SPEC improves the safety consistency of TRPO-P across environments. On the Goal tasks, SPEC reduces cost on PointGoal ($27.8 \rightarrow 24.4$) and CarGoal ($31.8 \rightarrow 24.9$), bringing both within the cost limit while maintaining comparable reward. On DoggoButton, SPEC matches TRPO-P's cost (24.1) while improving reward ($0.1 \rightarrow 0.2$), and on PointButton it substantially reduces cost ($32.3 \rightarrow 18.1$) while matching TRPO-P's reward (-1.3). On the Circle tasks, where TRPO-P already operates well below the cost limit, SPEC uses more of

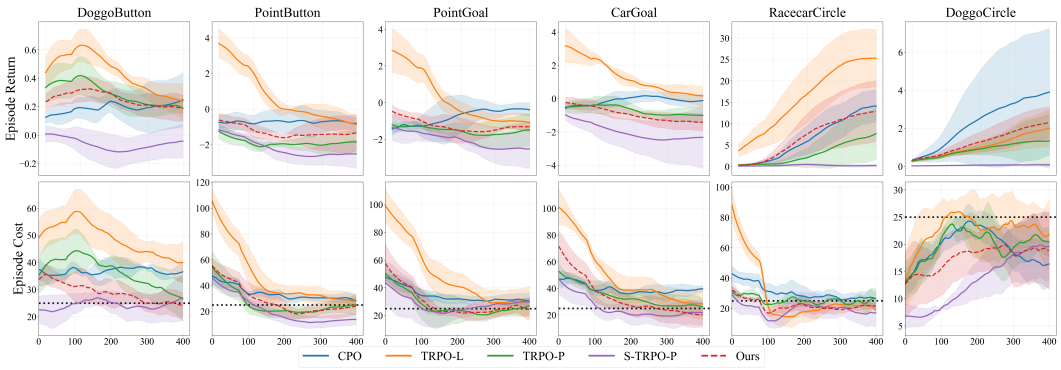


Figure 2: Learning curves on Safety-Gymnasium tasks over epochs. Lines show IQM and bands denote standard deviation across seeds. Top: episodic return; bottom: episodic cost (black dashed lines indicate cost limit = 25). Our approach (SPEC-TRPO-P) maintains predominantly safe behavior, with violations remaining closer to the safety threshold than competing methods, with strong reward.

Table 1: IQM $\pm$ STD for reward (top) and cost (bottom). Shaded cells satisfy IQM cost ≤ 25. Bold denotes the highest-reward safe method per environment (or, if none are safe, the method closest to the threshold). * denotes safe methods with reward below 5% of the highest safe reward.

Env	Metric	CPO	TRPO-L	TRPO-P	S-TRPO-P	Ours
DoggoButton	Reward	0.2 ± 0.1	0.2 ± 0.1	0.1 ± 0.1	0.0 ± 0.1	$\mathbf{0.2 \pm 0.1}$
	Cost	45.9 ± 3.9	33.3 ± 11.7	24.1 ± 6.4	28.7 ± 7.6	$\mathbf{24.1 \pm 3.9}$
PointButton	Reward	-0.4 ± 0.2	-1.0 ± 0.4	-1.3 ± 0.9	-2.5 ± 0.9	$\mathbf{-1.3 \pm 0.8}$
	Cost	30.0 ± 1.5	25.3 ± 6.8	32.3 ± 11.6	14.6 ± 4.8	$\mathbf{18.1 \pm 9.5}$
PointGoal	Reward	-0.3 ± 0.3	-1.3 ± 0.5	-1.5 ± 1.2	-2.5 ± 1.2	$\mathbf{-1.2 \pm 0.4}$
	Cost	34.0 ± 9.3	27.8 ± 8.7	27.8 ± 11.6	23.7 ± 6.4	$\mathbf{24.4 \pm 6.6}$
CarGoal	Reward	0.1 ± 0.7	0.4 ± 0.6	-1.1 ± 0.7	-2.5 ± 1.8	$\mathbf{-1.5 \pm 0.4}$
	Cost	37.2 ± 14.8	30.9 ± 4.7	31.8 ± 6.3	25.5 ± 11.2	$\mathbf{24.9 \pm 16.9}$
RacecarCircle	Reward	14.0 ± 8.2	$\mathbf{25.7 \pm 6.5}$	5.9 ± 7.9	$0.3 \pm 0.4^*$	14.8 ± 7.7
	Cost	21.9 ± 8.8	$\mathbf{12.5 \pm 6.7}$	8.5 ± 4.4	9.5 ± 7.8	23.8 ± 13.3
DoggoCircle	Reward	2.3 ± 1.1	2.1 ± 1.1	$\mathbf{4.5 \pm 3.0}$	$0.1 \pm 0.1^*$	2.4 ± 0.8
	Cost	26.5 ± 14.2	15.5 ± 6.4	$\mathbf{13.4 \pm 12.1}$	16.6 ± 6.8	18.3 ± 7.7

the available cost budget: on RacecarCircle, cost increases from 8.5 to 23.8 while reward improves from 5.9 to 14.8; on DoggoCircle, cost increases from 13.4 to 18.3 with reward decreasing from 4.5 to 2.4. Notably, in both Circle environments where SPEC does not achieve the highest safe reward, it attains the second-highest safe reward, indicating that its stronger safety consistency does not come at the cost of broadly poor task performance. Compared to S-TRPO-P, SPEC avoids the severe reward collapse observed on Circle tasks (RacecarCircle: 0.3 vs. 14.8; DoggoCircle: 0.1 vs. 2.4). Our approach also achieves a cost IQM within the cost limit across all six reported environments.

Ablation studies. Table 3 summarizes trade-offs in $\beta_{\max}$, correction scaling, and observation history. The $\beta_{\max}$ ablation shows an overall safety–performance trade-off: larger $\beta_{\max}$ values generally lead to safer performance, with costs closer to or below the constraint, sometimes at the expense of reward. This effect is not strictly monotonic, as expected in safe RL, where changing the correction strength also alters the coupled evolution of the critic, policy, visitation distribution, and retrieval memory; nevertheless, the sweep follows the broader trend toward safer behavior at larger $\beta_{\max}$. At $\beta_{\max} = 0$, SPEC reduces to TRPO-P with environment observation stacking enabled, providing a useful control for isolating the contribution of the SIF correction. Its performance also shows that environment-only stacking is not universally beneficial across algorithms, as its gains appear primarily when paired with SPEC-TRPO-P, suggesting that the observation representation and strong SIF corrections are most effective in combination. The adaptive-β ablation shows a

Table 2: $\beta_{\max}$ ablation; shading and boldface follow Table 1.

Env	Metric	$\beta_{\max} = 0$	$\beta_{\max} = 0.05$	$\beta_{\max} = 0.1$	$\beta_{\max} = 0.15$
DoggoButton	Reward	0.2 ± 0.1	0.2 ± 0.1	0.1 ± 0.1	$\mathbf{0.2 \pm 0.1}$
	Cost	27.8 ± 7.5	31.5 ± 4.3	30.6 ± 10.5	$\mathbf{24.1 \pm 3.9}$
PointButton	Reward	-1.5 ± 0.7	-1.3 ± 0.5	-0.9 ± 0.3	$\mathbf{-1.3 \pm 0.8}$
	Cost	24.4 ± 8.9	27.1 ± 5.4	27.4 ± 9.4	$\mathbf{18.1 \pm 9.5}$
PointGoal	Reward	-1.4 ± 1.4	-1.1 ± 0.8	-1.8 ± 0.4	$\mathbf{-1.2 \pm 0.4}$
	Cost	31.0 ± 13.7	26.1 ± 6.6	17.4 ± 7.0	$\mathbf{24.4 \pm 6.6}$
CarGoal	Reward	-1.6 ± 0.4	-1.1 ± 0.5	-1.5 ± 0.4	$\mathbf{-1.5 \pm 0.4}$
	Cost	25.1 ± 7.5	28.2 ± 9.6	23.5 ± 7.2	$\mathbf{24.9 \pm 16.9}$
RacecarCircle	Reward	14.1 ± 3.6	14.2 ± 8.3	12.5 ± 6.0	$\mathbf{14.8 \pm 7.7}$
	Cost	15.0 ± 10.5	22.3 ± 18.7	10.8 ± 10.5	$\mathbf{23.8 \pm 13.3}$
DoggoCircle	Reward	1.7 ± 1.1	2.3 ± 0.9	1.2 ± 0.7	$\mathbf{2.5 \pm 0.8}$
	Cost	20.0 ± 10.4	19.4 ± 10.6	18.9 ± 7.6	$\mathbf{18.3 \pm 9.6}$

Table 3: adaptive β (left) and obs-stacking (right) ablations; shading and boldface follow Table 1.

Env	Metric	non-adapt. β	adapt. β
DoggoButton	Reward	0.2 ± 0.0	$\mathbf{0.2 \pm 0.1}$
	Cost	27.0 ± 4.2	$\mathbf{24.1 \pm 3.9}$
PointButton	Reward	$\mathbf{-1.2 \pm 1.0}$	-1.3 ± 0.8
	Cost	$\mathbf{24.3 \pm 14.1}$	18.1 ± 9.5
PointGoal	Reward	-2.0 ± 0.6	$\mathbf{-1.2 \pm 0.4}$
	Cost	18.8 ± 5.8	$\mathbf{24.4 \pm 6.6}$
CarGoal	Reward	$\mathbf{-0.8 \pm 0.5}$	-1.5 ± 0.4
	Cost	$\mathbf{18.8 \pm 10.0}$	24.9 ± 16.9
RacecarCircle	Reward	15.6 ± 7.4	$\mathbf{14.8 \pm 7.7}$
	Cost	26.5 ± 18.8	$\mathbf{23.8 \pm 13.3}$
DoggoCircle	Reward	$\mathbf{2.9 \pm 1.3}$	2.5 ± 0.8
	Cost	$\mathbf{7.2 \pm 3.9}$	18.3 ± 9.6

full obs stack	env obs stack
0.1 ± 0.1	$\mathbf{0.2 \pm 0.1}$
30.4 ± 8.5	$\mathbf{24.1 \pm 3.9}$
$\mathbf{-1.2 \pm 1.4}$	-1.3 ± 0.8
$\mathbf{24.7 \pm 10.3}$	18.1 ± 9.5
-1.8 ± 0.6	$\mathbf{-1.2 \pm 0.4}$
14.8 ± 2.9	$\mathbf{24.4 \pm 6.6}$
$\mathbf{-0.9 \pm 0.4}$	-1.5 ± 0.4
$\mathbf{24.5 \pm 7.4}$	24.9 ± 16.9
9.0 ± 3.2	$\mathbf{14.8 \pm 7.7}$
13.8 ± 14.1	$\mathbf{23.8 \pm 13.3}$
1.3 ± 0.6	$\mathbf{2.5 \pm 0.8}$
27.9 ± 22.2	$\mathbf{18.3 \pm 9.6}$

similar environment-dependent trade-off: adaptive scaling yields safer behavior on DoggoButton and RacecarCircle and a higher safe reward on PointGoal, while the non-adaptive variant achieves higher reward on PointButton, CarGoal, and DoggoCircle. Finally, environment-only observation history outperforms full observation stacking on four of the six tasks while maintaining an IQM cost at or below the cost limit on all six, whereas full stacking achieves slightly higher reward on PointButton and CarGoal. These ablations show that SPEC's safest performance comes from combining correction strength, adaptive scaling, and selective temporal context rather than maximizing any single component in isolation. Additional learning curves and ablations are provided in Appendix I.3.

6 Conclusion

We introduced the Safe Integration Framework (SIF), a class of admissible critic corrections that preserve the per-iteration trust-region performance bound and the local stability of primal–dual Lagrangian dynamics in on-policy safe RL, up to an explicit, hyperparameter-controlled slack/neighborhood. As one instantiation, Safe Proximity-guided Episodic Control (SPEC) augments the cost critic with retrieval-based corrections, entropy-aware scaling, and environment-only observation stacking. On Level 2 Safety-Gymnasium tasks, SPEC applied to TRPO-PID yields a more consistent constraint-satisfaction profile than prior baselines while avoiding the severe reward-collapse failure mode exhibited by prior episodic-control approaches to safe RL. Limitations include non-uniform reward gains across tasks. Future work could explore: (i) unique SIF instantiations with different auxiliary information; and (ii) mechanisms for achieving reward dominance across tasks.

AI USE STATEMENT

In this work, we used generative AI tools to help formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, implement methods, and support quantitative data analysis and interpretation. We did not use generative AI tools for qualitative or thematic data analysis, and the remaining required disclosure tasks were either not performed with generative AI or were not applicable to this work. Additionally, we used generative AI tools to polish and improve the readability of the manuscript, assist with formatting graphs and scientific figures, summarize existing literature, and format references.

We reviewed and verified all AI-assisted work. AI-assisted writing was manually reviewed, edited, and revised by the authors to ensure that it accurately reflected our intended claims and conclusions. Mathematical claims and AI-suggested proof ingredients were checked against relevant literature and independently derived and verified by the authors. All proofs were subsequently reviewed in depth by the authors for correctness. AI-assisted implementations were manually inspected and tested for correctness. Quantitative analyses were likewise manually checked, and their interpretation was independently reviewed to ensure that the reported results and takeaways reflected the authors' own analysis. Literature summaries were verified against the original sources, and AI-assisted figures, graphs, and references were manually checked for accuracy and appropriate presentation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide an anonymized repository containing the full implementation of SIF and SPEC (§ 3–4), including integration with OmniSafe, training pipelines, and evaluation scripts:

`https://anonymous.4open.science/r/spec-saferl-A7AF`

The repository includes: (i) implementations of all baseline methods used in § 5 and Appendix I, (ii) configuration files for each Safety-Gymnasium task described in Appendix H, (iii) scripts to reproduce all figures and tables in § 5 and Appendix I, and (iv) detailed instructions for environment setup and experiment execution.

All hyperparameters and experimental settings are consistent with those reported in Appendix I.2.

For the entire benchmark suite, we used a Linux HPC cluster. All experiments were run on Linux compute nodes; our code has not been tested on non-Linux systems. For our standard parallel configuration with $\texttt{num_pool} = 30$, each run used 54 CPU cores. Under this configuration, each random seed completed in under 3.5 hours on average and required approximately 2.5 GB of RAM.

AUTHOR CONTRIBUTIONS

Landon led the project and was responsible for the majority of the work. He developed the core research direction, formulated the Safe Integration Framework (SIF), designed Safe Proximity-guided Episodic Control (SPEC), developed the theoretical analysis and proofs, implemented the proposed methods and experimental infrastructure, designed and conducted the experiments and ablation studies, analyzed and visualized the results, and wrote the manuscript and appendices. Mateus contributed through discussion and critical review of the theoretical arguments, experimental analysis, and manuscript. Ram provided research guidance and contributed to the review and refinement of the analysis, theoretical presentation, and manuscript.

ACKNOWLEDGMENTS

The computing for this project was performed using institutional high-performance computing resources. We thank the associated research computing staff for providing computational resources and support.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, Boca Raton, FL, 1999. ISBN 9780849303821.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, 6(5): 679–684, 1957.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control, 2016. URL https://arxiv.org/abs/1606.04460.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D. Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating RL for LLM reasoning with optimal advantage regression. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=T1V8BJO0iG.

Claire Chen, Shuze Liu, and Shangtong Zhang. Efficient policy evaluation with safety constraint for reinforcement learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34989–35010, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5681251fa039cf49d6d11b906eded1b3-Paper-Conference.pdf.

Fan Chen, Junyu Zhang, and Zaiwen Wen. A near-optimal primal-dual method for off-policy learning in cmdp. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10521–10532. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/444d69470b24ded080183c907b711bbf-Paper-Conference.pdf.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8378–8390. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5f7695debd8cde8db5abcb9f161b49ea-Paper.pdf.

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7396–7404, Jun. 2023. doi: 10.1609/aaai.v37i6.25900. URL https://ojs.aaai.org/index.php/AAAI/article/view/25900.

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110:2419–2468, 2021. doi: 10.1007/s10994-021-05961-4.

Sven Gronauer. Bullet-safety-gym: A framework for constrained reinforcement learning. Technical report, mediaTUM, 2022.

Steven Hansen, Alexander Pritzel, Pablo Sprechmann, Andre Barreto, and Charles Blundell. Fast deep reinforcement learning using online adjustments from the past. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f7bdb0e100275600f9e183e25d81822d-Paper.pdf.

Hao Hu, Jianing Ye, Guangxiang Zhu, Zhizhou Ren, and Chongjie Zhang. Generalizable episodic memory for deep reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4380–4390. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hu21d.html.

Xiaohui Huang, Xiaoji Cheng, Nan Jiang, Wei Liu, and Ying Yu. Hierarchical reinforcement learning with attention-driven multi-actor a2c and shared critic for fleet management. *Expert Systems with Applications*, 295:128856, 2026. doi: 10.1016/j.eswa.2025.128856.

Maximilian Jaritz, Raoul de Charette, Marin Toromanoff, Etienne Perot, and Fawzi Nashashibi. End-to-end race driving with deep reinforcement learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2070–2075, 2018. doi: 10.1109/ICRA.2018.8460934.

Jiaming Ji, Borui Zhang, Jiayi Zhou, Xianyuan Pan, Weidong Huang, Rui Sun, Yiran Geng, Yizhou Zhong, Jialu Dai, et al. Safety-gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *Journal of Machine Learning Research*, 25(285):1–6, 2024. URL http://jmlr.org/papers/v25/23-0681.html.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Dohyeong Kim and Songhwai Oh. Efficient off-policy safe reinforcement learning using trust region conditional value at risk. *IEEE Robotics and Automation Letters*, 7(3):7644–7651, 2022. doi: 10.1109/LRA.2022.3184793.

Dohyeong Kim, Kyungjae Lee, and Songhwai Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19908–19939. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3f20f2b0315c72201e23512fdbd1ee91-Paper-Conference.pdf.

Juhyeon Kim and Kihyun Kim. Optimizing large-scale fleet management on a road network using multi-agent deep reinforcement learning with graph neural network. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 990–995, 2021. doi: 10.1109/ITSC48978.2021.9565029.

Vijay R. Konda and John N. Tsitsiklis. Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. doi: 10.1137/S0363012901385691. URL https://doi.org/10.1137/S0363012901385691.

Igor Kuznetsov and Andrey Filchenkov. Solving continuous control with episodic memory. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2651–2657. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/365. URL https://doi.org/10.24963/ijcai.2021/365. Main Track.

Hung Le, Thommen Karimpanal George, Majid Abdolshah, Truyen Tran, and Svetha Venkatesh. Model-based episodic memory induces dynamic hybrid controls. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30313–30325. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fe73f687e5bc5280214e0486b273a5f9-Paper.pdf.

Máté Lengyel and Peter Dayan. Hippocampal contributions to control: The third way. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/1f4477bad7af3616c1f933a02bfabe4e-Paper.pdf.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL http://jmlr.org/papers/v17/15-522.html.

Sergey Levine, Peter Pastor Sampedro, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. 2017. URL https://arxiv.org/pdf/1603.02199.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127/.

Yan Li, Junjie He, and Yayu Gao. Intelligent traffic signal control with deep reinforcement learning at single intersection. In *Proceedings of the 2021 7th International Conference on Computing and Artificial Intelligence*, ICCAI '21, pp. 399–406, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389501. doi: 10.1145/3467707.3467767. URL https://doi.org/10.1145/3467707.3467767.

Zhuo Li, Derui Zhu, Yujing Hu, Xiaofei Xie, Lei Ma, YAN ZHENG, Yan Song, Yingfeng Chen, and Jianjun Zhao. Neural episodic control with state abstraction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=C2fsSj3ZGiU.

Zhuo Li, Derui Zhu, and Jens Grossklags. Safe reinforcement learning via episodic control. *IEEE Access*, 13:35270–35280, 2025. doi: 10.1109/ACCESS.2025.3535679.

Zichuan Lin, Tianqi Zhao, Guangwen Yang, and Lintao Zhang. Episodic memory deep q-networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2433–2439. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/337. URL https://doi.org/10.24963/ijcai.2018/337.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020. doi: 10.1609/aaai.v34i04.5932. URL https://aaai.org/papers/04940-ipo-interior-point-policy-optimization-under-constraints/.

Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13644–13668. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22b.html.

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. doi: 10.1038/nature14236.

Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020. doi: 10.15607/RSS.2020.XVI.064.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2827–2836. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pritzel17a.html.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *OpenAI*, 2019. URL https://openai.com/index/benchmarking-safe-exploration-in-deep-reinforcement-learning/.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961.

Lindsay Spoor, Álvaro Serra-Gómez, Aske Plaat, and Thomas Moerland. An empirical study of lagrangian methods in safe reinforcement learning, 2025. URL https://arxiv.org/abs/2510.17564.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

Yifan Sun, Jingyan Shen, Yibin Wang, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, and Huan Zhang. Improving data efficiency for LLM reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uwUkETPIJN.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Zifan Wu, Bo Tang, Qian Lin, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Off-policy primal-dual safe reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 11371–11390,

2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2f8b56543953d60f262fb2c4b85c50b3-Paper-Conference.pdf.

Long Yang, Jiaming Ji, Juntao Dai, Linrui Zhang, Binbin Zhou, Pengfei Li, Yaodong Yang, and Gang Pan. Constrained update projection approach to safe policy optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9111–9124. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3ba7560b4c3e66d760fbdd472cf4a5a9-Paper-Conference.pdf.

Qisong Yang, Thiago D. Simão, Simon H Tindemans, and Matthijs T. J. Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10639–10646, May 2021. doi: 10.1609/aaai.v35i12.17272. URL https://ojs.aaai.org/index.php/AAAI/article/view/17272.

Ruolin Yang, Zhuoren Li, Bo Leng, and Lu Xiong. Safe reinforcement learning for autonomous vehicles to make lane-change decisions: Constraint based on incomplete information game theory. In *2023 7th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, pp. 1–6, 2023. doi: 10.1109/CVCI59596.2023.10397427.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rke3TJrtPS.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey, 2020. URL https://arxiv.org/abs/1908.08796.

Junshan Zhang and Dong Zheng. A stochastic primal-dual algorithm for joint flow control and mac design in multi-hop wireless networks. In *2006 40th Annual Conference on Information Sciences and Systems*, pp. 339–344, 2006. doi: 10.1109/CISS.2006.286489.

Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Xueqian Wang, Bo Yuan, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3744–3750. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/520. URL https://doi.org/10.24963/ijcai.2022/520. Main Track.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15338–15349. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/af5d5ef24881f3c3049a7b9bfe74d58b-Paper.pdf.

Zhi Zheng and Shangding Gu. Safe multiagent reinforcement learning with bilevel optimization in autonomous driving. *IEEE Transactions on Artificial Intelligence*, 6(4):829–842, 2025. doi: 10.1109/TAI.2024.3497919.

Guangxiang Zhu*, Zichuan Lin*, Guangwen Yang, and Chongjie Zhang. Episodic reinforcement learning with associative memory. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxjqxBYDB.

A Related Work

Deep safe RL. Deep safe RL algorithms, particularly off-policy approaches with theoretical guarantees, typically rely on strong assumptions that are difficult to satisfy in realistic safety-critical environments. For instance, Conditional Value at Risk (CVaR)-based methods (Kim & Oh, 2022; Yang et al., 2021) optimize for the worst-case cost-return, but assume Gaussian or Gaussian-induced cost-return distributions to obtain tractable guarantees, despite safety costs often being sparse, discontinuous, and heavy-tailed. Similarly, (Kim et al., 2023) assumes differentiability and convexity of the cost-return function in policy space, which is rarely satisfied by threshold-based safety constraints. Offline safe RL methods (Chen et al., 2025) require high-quality, high-coverage datasets and are typically only guaranteed for discrete action spaces. Off-policy primal–dual approaches (Wu et al., 2024) provide convergence guarantees under specific assumptions, but are known to suffer from instability, leaving many methods to remain largely theoretical (Chen et al., 2022; Ding & Lavaei, 2023). Although the work from Liu et al. (2022) avoids some of these stronger assumptions, it has been shown to perform worse in complex environments than other approaches mentioned (Wu et al., 2024).

On-policy safe RL methods avoid many of the fragile distributional and coverage assumptions inherent to off-policy learning. However, certain approaches still suffer from challenges related to optimization stability and constraint enforcement. In particular, penalty-based methods such as Interior Point Optimization (IPO) (Liu et al., 2020) and Penalized Proximal Policy Optimization (P3O) (Zhang et al., 2022) enforce safety by augmenting the policy objective with barrier or exact penalty functions, but their effectiveness depends critically on penalty hyperparameter tuning. In IPO, increasing the log-barrier parameter t sharpens constraint enforcement at the cost of heightened gradient sensitivity near the feasibility boundary, which has been shown to induce optimization instability and even constraint violations in practice. Similarly, in P3O, the penalty coefficient κ must be sufficiently large to enforce constraints yet can destabilize learning and exacerbate approximation error when over-scaled.

Consequently, this work focuses on on-policy local policy search safe RL methods based on primal–dual and constrained trust-region optimization, which admit cleaner theoretical guarantees under comparatively milder and more interpretable assumptions. Constrained trust-region approaches enforce safety directly within each policy update by solving a locally constrained optimization problem. Early second-order methods such as Constrained Policy Optimization (CPO) (Achiam et al., 2017) provided policy performance and constraint-satisfaction bounds by optimizing a constrained trust-region, using first-order approximations of the reward and cost objectives and a second-order quadratic approximation of the KL divergence between successive policies. Building on this trust-region perspective, (Yang et al., 2020) introduced Projection-based Constrained Policy Optimization (PCPO), which enforces safety via explicit projection onto the feasible set following unconstrained trust-region updates, leveraging second-order KL geometry to provide per-update bounds on reward degradation and constraint violation. More recent first-order projection-based methods include Constrained Update Projection (CUP) (Yang et al., 2022), which improves these bounds by deriving tighter policy performance difference guarantees for generalized advantage estimation (GAE) and proposing a two-stage procedure that decouples unconstrained optimization from safety enforcement through analytically derived trust-region projections. Similarly, First Order Constrained Optimization in Policy Space (FOCOPS) (Zhang et al., 2020) performs constrained updates in a nonparametric policy space and projects them back into the parametric class using first-order optimization, yielding approximate bounds on worst-case constraint violation.

In the primal–dual setting, Lagrangian methods (Ray et al., 2019; Stooke et al., 2020; Spoor et al., 2025) enforce constraints by introducing dual variables and optimizing a saddle-point objective, leading to widely used variants such as TRPO-Lag and PPO-Lag. While these methods are simple to implement and often effective in practice, their performance depends on the stability of coupled policy and multiplier updates, which can exhibit oscillatory or even divergent behavior without careful step-size tuning. To address this, PID-based approaches (Stooke et al., 2020) augment the standard dual ascent update with proportional–integral–derivative control, incorporating not only the instantaneous constraint violation but also its accumulated history and rate of change. This yields more responsive and damped multiplier dynamics, improving transient behavior and reducing oscillations in practice, and often leads to greater robustness to hyperparameter choices while maintaining the scalability of first-order optimization. Both theoretically and empirically, on-policy local policy

search methods—both constrained trust-region and primal–dual—rely strongly on accurate critic estimation and stable advantage computation. *Our EC method introduces a controlled, bounded critic bias through clipped episodic correction, demonstrating that the policy performance guarantees of constrained trust-region methods and the local convergence structure of primal–dual dynamics can be preserved under biased value and advantage estimation, up to an explicit slack term and perturbation neighborhood, respectively.*

Episodic control. Episodic control was originally introduced in cognitive science (Lengyel & Dayan, 2007) and later adapted to reinforcement learning by (Blundell et al., 2016) as a non-parametric approach that stores state–action pairs with their observed returns and selects actions greedily via nearest-neighbor value lookup. This framework was extended in Neural Episodic Control (NEC) (Pritzel et al., 2017), which replaces raw state storage with learned embeddings and differentiable memory access to improve generalization. Closely related methods such as episodic value augmentation (EVA) (Hansen et al., 2018) similarly compute value estimates and greedy policies directly from episodic memory at decision time. *While these approaches enable fast early learning and strong data efficiency, they bypass Bellman backups, resulting in a lack of Bellman consistency or asymptotic guarantees, and making natural extensions to continuous action spaces challenging.*

Subsequent work sought to integrate episodic returns into standard value-based learning. Episodic Memory Deep Q-Networks (EMDQN) (Lin et al., 2018) incorporates episodic returns into Q-learning by jointly optimizing against the standard TD target and a memory-based TD target, while Episodic Memory Actor-Critic (EMAC) (Kuznetsov & Filchenkov, 2021) extends this idea to continuous control by introducing stored Monte Carlo returns as an auxiliary regression target for the critic. In discrete action spaces, Model-Based Episodic Control (MBEC) (Le et al., 2021) integrates trajectory-level episodic memory with model-based planning, using recalled high-return trajectories to guide action selection and value estimation, at the cost of strong sensitivity to model accuracy and a large number of interacting hyperparameters. More recently, NEC with State Abstraction (NECSA) (Li et al., 2023) improves both discrete and continuous EC by clustering states into abstract representations and constructing revised rewards over these abstractions, improving generalization and sample efficiency. *Unlike our method, these EC algorithms are designed for unconstrained RL settings. Most importantly, none of these methods provides policy performance or convergence guarantees, a property that is particularly critical in safe reinforcement learning, where uncontrolled performance degradation or constraint violations are unacceptable.*

Several methods introduce structured episodic memory with theoretical guarantees, *but only in unconstrained setting and under highly idealized assumptions.* For example, Zhu* et al. (2020) introduces Episodic Reinforcement Learning with Associative Memory (ERLAM), which propagates episodic returns through a graph-structured associative memory. However, its guarantees are limited to Bellman-style value convergence and do not extend to policy convergence or policy performance. Moreover, ERLAM is restricted to discrete action spaces and has been shown to degrade significantly in stochastic environments, where optimistic max-based propagation amplifies noise rather than useful signal Zhu* et al. (2020). Similarly, Hu et al. (2021) proposes Generalizable Episodic Memory (GEM), which uses parametric trajectory-level memory and implicit value propagation to generalize high-return experiences, together with double-estimator mechanisms to mitigate overestimation. Unlike ERLAM, GEM additionally provides guarantees of policy convergence, which is why we refer to it in the main text as one of the only episodic-memory methods with policy-level guarantees. However, these guarantees for the learned value function and induced policy rely on near-deterministic environment dynamics to ensure meaningful convergence. *In contrast, our method is for the constrained setting and avoids explicit near-determinism assumptions on the environment dynamics, and instead provides safety and performance guarantees under assumptions compatible with stochastic continuous-control settings.*

All of the above approaches are integrated with off-policy RL methods and are **only designed for unconstrained settings.** *In contrast, our method is integrated with on-policy safe RL optimization in a principled and theoretically motivated manner, allowing it to retain the policy performance guarantees characteristic of local on-policy policy search algorithms.* To our knowledge, SafeEC Li et al. (2025) is the only EC-based algorithm proposed for Safe RL. It incorporates episodic memory through discretized state abstractions and reward shaping (like NECSA (Li et al., 2023)) based on joint reward–cost statistics. However, the method doesn't have any theoretical guarantees, and the

authors explicitly reported extreme sensitivity to hyperparameters and reshaped rewards in stochastic environments, which can induce instability in policy learning Li et al. (2025). *In contrast, our method avoids explicit reward reshaping and discretized state abstractions, and instead enforces safety through principled critic updates with theoretical guarantees that remain compatible in stochastic environments.*

B NOTATION

For convenience, Table 4 summarizes the notation introduced in this paper. We include only symbols specific to the proposed framework and theoretical analysis. Standard reinforcement learning notation (e.g., states, actions, value functions) follows common conventions and is therefore omitted.

Table 4: Summary of notation specific to the proposed framework, including SIF critic correction, SPEC retrieval components, and theoretical quantities used in the analysis. Standard MDP and RL notation is omitted.

Symbol	Meaning
$\tilde{\gamma}$	Effective contraction factor, $\gamma(1-\lambda)/(1-\gamma\lambda)$
$d_{\lambda}^{\pi_\theta}$	λ-state-visitation distribution under π_θ
$\epsilon_r^{\pi_\theta,\pi_{\theta'}}, \epsilon_{c_i}^{\pi_\theta,\pi_{\theta'}}$	Worst-case expected absolute TD residual of $V_r^{\pi_{\theta'}}$ (resp. $V_{c_i}^{\pi_{\theta'}}$) under actions from π_θ
$\bar{D}_{\mathrm{KL}}$, δ	Expected trust-region KL divergence and step-size bound, $\bar{D}_{\mathrm{KL}} \le \delta$
$z = (\theta, \mu)$	Joint primal–dual variable
$g(z)$	Mean primal–dual drift field
$z^\star$	Local saddle point / equilibrium
$J^\star$	Jacobian $\nabla g(z^\star)$
G, M_μ	Bounds on the score function and dual variable on $\mathcal{U}$
a_1, a_2	Lyapunov constants from the local-stability bound (Thm. E.3)
$\mathcal{T}_{\lambda,\beta}^{\pi_\theta}$	SIF-corrected Bellman evaluation operator
$\Delta_V(s)$	Generic value-dependent critic correction, L_Δ-Lipschitz in V
$\beta, \beta(s), \beta_{\max}$	Correction strength: scalar, state-dependent, and maximum
η	Clipping threshold for the critic correction
L_Δ	Lipschitz constant of the correction signal
$V_{\lambda,\beta}^{\pi_\theta}$	Fixed point of the SIF-corrected Bellman operator
$\hat{V}_{r,\lambda}^{\pi_\theta}, \hat{V}_{c,\lambda,\beta}^{\pi_\theta}$	Parametric (learned) estimates of the reward and SIF-corrected cost critics
$A_{r,\lambda,\beta}^{\pi_{\theta'}}, A_{c_i,\lambda,\beta}^{\pi_{\theta'}}$	Reward and cost advantages under the SIF-corrected critic
e_x, E_x ($x \in \{r, c\}$)	Critic-error magnitude $\|e_x\|_\infty$ and aggregated SIF-error term
f_c, f_r	Cost and reward state encoders
$\mathcal{D}_c$	Cost-memory datastore / retriever
τ_c	Retrieved cost rollout trajectory
$V_c^{\mathrm{nb}}(s)$	Neighbor-based cost value estimate
$V_{c,\lambda,\beta}^{\mathrm{targ}}(s)$	SPEC-adjusted cost critic target
$\Delta_c(s)$	SPEC residual, $V_c^{\mathrm{nb}}(s) - \hat{V}_{c,\lambda,\beta}^{\pi_\theta}(s)$
$z_{i,j}$	IVF cluster assignment of neighbor j for query i
$p_{i,c}, H_i$	Empirical IVF-cell distribution and its entropy
$g_{\mathrm{ent}}^{(i)}$	Normalized neighborhood concentration score
H	Observation history length for stacking

C λ-RETURN BELLMAN OPERATOR

Proposition C.1 (Contraction and fixed point of the λ-return Bellman operator)**.** *Let $\pi_\theta(a \mid s)$ be a stationary stochastic policy and let $\mathcal{T}^{\pi_\theta}$ denote the Bellman expectation operator with unique fixed point V^{π_θ}, i.e., $V^{\pi_\theta} = \mathcal{T}^{\pi_\theta} V^{\pi_\theta}$. Define the λ-return Bellman operator $\mathcal{T}_\lambda^{\pi_\theta}$ as*

$$(\mathcal{T}_\lambda^{\pi_\theta} V)(s) := (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \big((\mathcal{T}^{\pi_\theta})^n V\big)(s), \quad \lambda \in [0,1).$$

Then $\mathcal{T}_\lambda^{\pi_\theta}$ is a contraction under the supremum norm with modulus

$$\tilde{\gamma} = \frac{\gamma(1-\lambda)}{1-\gamma\lambda} < 1.$$

Consequently, by the Banach fixed-point (Banach, 1922), $\mathcal{T}_\lambda^{\pi_\theta}$ admits a unique fixed point $V_\lambda^{\pi_\theta}$ and iterating $\mathcal{T}_\lambda^{\pi_\theta}$ converges to $V_\lambda^{\pi_\theta}$ from any bounded initialization. Moreover, this fixed point coincides with the true Bellman evaluation fixed point:

$$V_\lambda^{\pi_\theta} = V^{\pi_\theta}.$$

Proof. Let $V, W : \mathcal{S} \to \mathbb{R}$ be two bounded value functions.

$$\begin{aligned} \|\mathcal{T}_\lambda^{\pi_\theta} V - \mathcal{T}_\lambda^{\pi_\theta} W\|_\infty &= \left\| (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \big((\mathcal{T}^{\pi_\theta})^n V - (\mathcal{T}^{\pi_\theta})^n W\big) \right\|_\infty \\ &\le (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \|(\mathcal{T}^{\pi_\theta})^n V - (\mathcal{T}^{\pi_\theta})^n W\|_\infty. \end{aligned}$$

Since $\mathcal{T}^{\pi_\theta}$ is a γ-contraction under $\|\cdot\|_\infty$, repeated application yields

$$\|(\mathcal{T}^{\pi_\theta})^n V - (\mathcal{T}^{\pi_\theta})^n W\|_\infty \le \gamma^n \|V - W\|_\infty.$$

Substituting this bound gives

$$\|\mathcal{T}_\lambda^{\pi_\theta} V - \mathcal{T}_\lambda^{\pi_\theta} W\|_\infty \le (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \|V - W\|_\infty.$$

Evaluating the geometric series,

$$(1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n = (1-\lambda)\gamma \sum_{n=1}^{\infty} (\gamma\lambda)^{n-1} = \frac{\gamma(1-\lambda)}{1-\gamma\lambda} =: \tilde{\gamma},$$

which yields

$$\|\mathcal{T}_\lambda^{\pi_\theta} V - \mathcal{T}_\lambda^{\pi_\theta} W\|_\infty \le \tilde{\gamma}\, \|V - W\|_\infty.$$

Since $\gamma \in (0,1)$ and $\lambda \in [0,1)$, we have $\tilde{\gamma} < 1$ and therefore $\mathcal{T}_\lambda^{\pi_\theta}$ is a contraction. By the Banach fixed-point theorem, $\mathcal{T}_\lambda^{\pi_\theta}$ admits a unique fixed point, denoted $V_\lambda^{\pi_\theta}$, and the iteration $V_{k+1} = \mathcal{T}_\lambda^{\pi_\theta} V_k$ converges to $V_\lambda^{\pi_\theta}$ for any bounded V_0.

It remains to show that this fixed point coincides with the Bellman evaluation fixed point V^{π_θ}. Because V^{π_θ} is a fixed point of $\mathcal{T}^{\pi_\theta}$, it is fixed by every power $(\mathcal{T}^{\pi_\theta})^n$, i.e., $(\mathcal{T}^{\pi_\theta})^n V^{\pi_\theta} = V^{\pi_\theta}$ for all $n \ge 1$. Substituting into the definition of $\mathcal{T}_\lambda^{\pi_\theta}$ yields

$$\mathcal{T}_\lambda^{\pi_\theta} V^{\pi_\theta} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} V^{\pi_\theta} = (1-\lambda) \cdot \frac{1}{1-\lambda}\, V^{\pi_\theta} = V^{\pi_\theta}.$$

Thus V^{π_θ} is a fixed point of $\mathcal{T}_\lambda^{\pi_\theta}$. By uniqueness of the fixed point of a contraction, we conclude that $V_\lambda^{\pi_\theta} = V^{\pi_\theta}$. □

D SIF BELLMAN OPERATOR

We first establish that the SIF correction preserves the contraction structure of the underlying λ-return Bellman operator under mild regularity conditions. The analysis proceeds in three steps. First, we show that the clipped correction map is Lipschitz with respect to the correction signal. We then combine this result with the contraction of the nominal λ-return operator (Prop. C.1) to show that the integrated operator remains contractive under a bounded perturbation condition. Finally, we bound the deviation between the corrected fixed point and the nominal value function.

We begin by characterizing how the clipped correction term changes as the underlying value function varies. Since the correction depends implicitly on the critic through Δ_V, controlling its sensitivity is essential for preserving stability of the integrated Bellman operator.

Lemma D.1 (Lipschitz continuity of the clipped correction map)**.** *Let* $\mathrm{clip}_\eta : \mathbb{R} \to [-\eta, \eta]$ *denote the elementwise clipping operator, and let* $\Delta_V : \mathcal{S} \to \mathbb{R}$ *denote a correction signal that may depend on* V*. Define the operator* $\mathcal{G}_\beta : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ *by*

$$(\mathcal{G}_\beta V)(s) := \mathrm{clip}_\eta\big(\beta\, \Delta_V(s)\big), \qquad s \in \mathcal{S},$$

where $\beta \geq 0$*. Then, for any bounded* V, W*,*

$$\|\mathcal{G}_\beta V - \mathcal{G}_\beta W\|_\infty \leq \beta\, \|\Delta_V - \Delta_W\|_\infty.$$

In particular, if Δ_V *is* L_Δ*-Lipschitz in* V *under* $\|\cdot\|_\infty$*, i.e.,*

$$\|\Delta_V - \Delta_W\|_\infty \leq L_\Delta \|V - W\|_\infty,$$

then $\mathcal{G}_\beta$ *is* (βL_Δ)*-Lipschitz:*

$$\|\mathcal{G}_\beta V - \mathcal{G}_\beta W\|_\infty \leq \beta L_\Delta \|V - W\|_\infty.$$

Proof. The scalar clipping map $\mathrm{clip}_\eta(\cdot)$ is 1-Lipschitz, i.e., for all $x, y \in \mathbb{R}$,

$$|\mathrm{clip}_\eta(x) - \mathrm{clip}_\eta(y)| \leq |x - y|. \tag{7}$$

Fix any $s \in \mathcal{S}$. Applying equation 7 with $x = \beta\, \Delta_V(s)$ and $y = \beta\, \Delta_W(s)$ gives

$$\begin{aligned} |(\mathcal{G}_\beta V)(s) - (\mathcal{G}_\beta W)(s)| &= |\mathrm{clip}_\eta(\beta\, \Delta_V(s)) - \mathrm{clip}_\eta(\beta\, \Delta_W(s))| \\ &\leq \beta\, |\Delta_V(s) - \Delta_W(s)|. \end{aligned}$$

Taking the supremum over $s \in \mathcal{S}$ yields

$$\|\mathcal{G}_\beta V - \mathcal{G}_\beta W\|_\infty \leq \beta\, \|\Delta_V - \Delta_W\|_\infty.$$

If, in addition, Δ_V is L_Δ-Lipschitz in V, then

$$\|\Delta_V - \Delta_W\|_\infty \leq L_\Delta \|V - W\|_\infty,$$

and therefore

$$\|\mathcal{G}_\beta V - \mathcal{G}_\beta W\|_\infty \leq \beta L_\Delta \|V - W\|_\infty.$$

□

Lemma D.1 shows that clipping prevents the correction term from amplifying perturbations beyond a factor proportional to β. We now use this property to establish that the integrated Bellman operator inherits contraction behavior from the nominal λ-return operator, provided the correction remains sufficiently bounded in its dependence on V.

Theorem D.2 (Contraction of the integrated Bellman operator)**.** *Fix a stationary stochastic policy* π_θ *and let* $\mathcal{T}_\lambda^{\pi_\theta}$ *denote the* λ*-return Bellman operator (Proposition C.1) with contraction modulus*

$$\tilde{\gamma} = \frac{\gamma(1-\lambda)}{1-\gamma\lambda} < 1.$$

Let $\Delta_V : \mathcal{S} \to \mathbb{R}$ *denote a correction signal that may depend on* V*, and assume it is* L_Δ*-Lipschitz in* V *under* $\|\cdot\|_\infty$*, i.e.,*

$$\|\Delta_V - \Delta_W\|_\infty \leq L_\Delta \|V - W\|_\infty \qquad \forall V, W.$$

Define the integrated operator $\mathcal{T}^{\pi_\theta}_{\lambda,\beta} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ *by*

$$(\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V)(s) := (\mathcal{T}^{\pi_\theta}_{\lambda} V)(s) + \mathrm{clip}_\eta\big(\beta\, \Delta_V(s)\big), \qquad s \in \mathcal{S}. \tag{8}$$

Then, for any bounded V, W*,*

$$\|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V - \mathcal{T}^{\pi_\theta}_{\lambda,\beta} W\|_\infty \le (\tilde{\gamma} + \beta L_\Delta)\, \|V - W\|_\infty. \tag{9}$$

In particular, if $\tilde{\gamma} + \beta L_\Delta < 1$*, the operator* $\mathcal{T}^{\pi_\theta}_{\lambda,\beta}$ *is a contraction under* $\|\cdot\|_\infty$ *and admits a unique fixed point.*

Proof. Let $V, W : \mathcal{S} \to \mathbb{R}$ be bounded. Using equation 8 and the triangle inequality,

$$\begin{aligned}\|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V - \mathcal{T}^{\pi_\theta}_{\lambda,\beta} W\|_\infty \le{}& \|\mathcal{T}^{\pi_\theta}_{\lambda} V - \mathcal{T}^{\pi_\theta}_{\lambda} W\|_\infty \\ &+ \Big\|\mathrm{clip}_\eta\big(\beta\Delta_V\big) - \mathrm{clip}_\eta\big(\beta\Delta_W\big)\Big\|_\infty.\end{aligned}$$

By Proposition C.1, $\mathcal{T}^{\pi_\theta}_{\lambda}$ is a $\tilde{\gamma}$-contraction, so

$$\|\mathcal{T}^{\pi_\theta}_{\lambda} V - \mathcal{T}^{\pi_\theta}_{\lambda} W\|_\infty \le \tilde{\gamma}\, \|V - W\|_\infty.$$

By Lemma D.1,

$$\Big\|\mathrm{clip}_\eta\big(\beta\Delta_V\big) - \mathrm{clip}_\eta\big(\beta\Delta_W\big)\Big\|_\infty \le \beta\, \|\Delta_V - \Delta_W\|_\infty.$$

Using the L_Δ-Lipschitz property of Δ_V,

$$\beta\, \|\Delta_V - \Delta_W\|_\infty \le \beta L_\Delta\, \|V - W\|_\infty.$$

Combining the bounds yields equation 9. When $\tilde{\gamma} + \beta L_\Delta < 1$, contraction and uniqueness follow from the Banach fixed-point theorem. □

Having established contraction of the integrated operator, we next quantify how far its fixed point can deviate from the nominal value function. The following result shows that the correction induces only a bounded perturbation, with the deviation controlled explicitly by the clipping threshold and contraction margin.

Theorem 3.1 (restated). Fix a stationary stochastic policy π_θ and assume $\tilde{\gamma} + \beta L_\Delta < 1$. Let V^{π_θ} denote the unique fixed point of $\mathcal{T}^{\pi_\theta}_{\lambda}$ (Proposition C.1), and let $V^{\pi_\theta}_{\lambda,\beta}$ denote the unique fixed point of $\mathcal{T}^{\pi_\theta}_{\lambda,\beta}$ (Theorem D.2). Then

$$\|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty \le \frac{\|\mathrm{clip}_\eta\big(\beta\, \Delta_{V^{\pi_\theta}}\big)\|_\infty}{1 - (\tilde{\gamma} + \beta L_\Delta)}. \tag{10}$$

Proof. By definition of the fixed points,

$$V^{\pi_\theta}_{\lambda,\beta} = \mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}_{\lambda,\beta}, \qquad V^{\pi_\theta} = \mathcal{T}^{\pi_\theta}_{\lambda} V^{\pi_\theta}.$$

Subtracting,

$$\begin{aligned}V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta} &= \mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}_{\lambda,\beta} - \mathcal{T}^{\pi_\theta}_{\lambda} V^{\pi_\theta} \\ &= \big(\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}_{\lambda,\beta} - \mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}\big) + \big(\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta} - \mathcal{T}^{\pi_\theta}_{\lambda} V^{\pi_\theta}\big).\end{aligned}$$

Taking supremum norms and applying the triangle inequality yields

$$\begin{aligned}\|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty \le{}& \|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}_{\lambda,\beta} - \mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}\|_\infty \\ &+ \|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta} - \mathcal{T}^{\pi_\theta}_{\lambda} V^{\pi_\theta}\|_\infty.\end{aligned}$$

By Theorem D.2, $\mathcal{T}^{\pi_\theta}_{\lambda,\beta}$ is a $(\tilde{\gamma} + \beta L_\Delta)$-contraction, hence

$$\|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}_{\lambda,\beta} - \mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta}\|_\infty \le (\tilde{\gamma} + \beta L_\Delta) \|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty.$$

For the second term, using the definition of the operator,

$$\|\mathcal{T}^{\pi_\theta}_{\lambda,\beta} V^{\pi_\theta} - \mathcal{T}^{\pi_\theta}_{\lambda} V^{\pi_\theta}\|_\infty = \|\mathrm{clip}_\eta\big(\beta\, \Delta_{V^{\pi_\theta}}\big)\|_\infty.$$

Combining the two bounds gives

$$\|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty \le (\tilde{\gamma} + \beta L_\Delta)\|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty + \|\mathrm{clip}_\eta\big(\beta\, \Delta_{V^{\pi_\theta}}\big)\|_\infty.$$

Rearranging and using $\tilde{\gamma} + \beta L_\Delta < 1$, we obtain equation 10.

$$\|V^{\pi_\theta}_{\lambda,\beta} - V^{\pi_\theta}\|_\infty \le \frac{\|\mathrm{clip}_\eta\big(\beta\, \Delta_{V^{\pi_\theta}}\big)\|_\infty}{1 - (\tilde{\gamma} + \beta L_\Delta)}.$$

□

E SIF CONSTRAINED TRUST-REGION POLICY PERFORMANCE BOUNDS

This section lifts the fixed-point perturbation analysis to policy performance guarantees. The goal is to show that using the integrated critic does not invalidate the constrained constrained trust-region performance bounds. Instead, it introduces explicit slack terms controlled by the critic deviation $\|e\|_\infty$, and therefore by the clipping threshold η and contraction margin $1-(\tilde{\gamma}+\beta L_\Delta)$.

We first quantify how replacing the true value function with the integrated fixed point affects one-step TD residuals. This is the basic perturbation step: any value error can only enter a TD residual through the current and next state values.

Lemma E.1 (TD residual perturbation under the integration framework)**.** *Fix a stationary reference policy π_θ. Let V^{π_θ} denote the true value function and $V^{\pi_\theta}_{\lambda,\beta}$ the SPEC fixed point. Define the value error*

$$e(s) := V^{\pi_\theta}_{\lambda,\beta}(s) - V^{\pi_\theta}(s).$$

Then for any timestep t,

$$\left|\delta_t^{V^{\pi_\theta}_{\lambda,\beta}} - \delta_t^{V^{\pi_\theta}}\right| \le (1+\gamma)\,\|e\|_\infty. \tag{11}$$

Proof. By definition, for any value function V,

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t).$$

Apply this with $V = V^{\pi_\theta}_{\lambda,\beta}$ and $V = V^{\pi_\theta}$:

$$\delta_t^{V^{\pi_\theta}_{\lambda,\beta}} = r_t + \gamma V^{\pi_\theta}_{\lambda,\beta}(s_{t+1}) - V^{\pi_\theta}_{\lambda,\beta}(s_t),$$
$$\delta_t^{V^{\pi_\theta}} = r_t + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t).$$

Subtract the second equation from the first:

$$\delta_t^{V^{\pi_\theta}_{\lambda,\beta}} - \delta_t^{V^{\pi_\theta}} = \gamma\Big(V^{\pi_\theta}_{\lambda,\beta}(s_{t+1}) - V^{\pi_\theta}(s_{t+1})\Big) - \Big(V^{\pi_\theta}_{\lambda,\beta}(s_t) - V^{\pi_\theta}(s_t)\Big).$$

Using the definition $e(s) := V^{\pi_\theta}_{\lambda,\beta}(s) - V^{\pi_\theta}(s)$, this becomes

$$\delta_t^{V^{\pi_\theta}_{\lambda,\beta}} - \delta_t^{V^{\pi_\theta}} = \gamma e(s_{t+1}) - e(s_t).$$

Applying the triangle inequality,

$$|\gamma e(s_{t+1}) - e(s_t)| \le \gamma|e(s_{t+1})| + |e(s_t)| \le (1+\gamma)\|e\|_\infty.$$

□

Since GAE is a discounted sum of TD residuals, the one-step perturbation bound immediately propagates to the advantage estimate. The following lemma shows that the total advantage error is bounded by a geometric accumulation of the TD residual error.

Lemma E.2 (GAE perturbation under new value error)**.** *Let $A^{\pi_\theta}_\lambda$ denote the λ-return advantage computed with $V^{\pi_\theta}_\lambda$, and let $A^{\pi_\theta}_{\lambda,\beta}$ denote the GAE computed using $V^{\pi_\theta}_{\lambda,\beta}$. Then*

$$\left|A^{\pi_\theta}_{\lambda,\beta}(s,a) - A^{\pi_\theta}_\lambda(s,a)\right| \le \frac{1+\gamma}{1-\gamma\lambda}\,\|e\|_\infty. \tag{12}$$

Proof. Recall the GAE(γ,λ) definition

$$A^{\pi_\theta}_{\lambda,t}(s_t,a_t) = \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell\,\delta_{t+\ell}^{V^{\pi_\theta}_\lambda}, \qquad A^{\pi_\theta}_{\lambda,\beta,t}(s_t,a_t) = \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell\,\delta_{t+\ell}^{V^{\pi_\theta}_{\lambda,\beta}}.$$

Subtracting and regrouping termwise gives

$$A^{\pi_\theta}_{\lambda,\beta,t}(s_t,a_t) - A^{\pi_\theta}_{\lambda,t}(s_t,a_t) = \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell\,\big(\delta_{t+\ell}^{V^{\pi_\theta}_{\lambda,\beta}} - \delta_{t+\ell}^{V^{\pi_\theta}_\lambda}\big).$$

Taking absolute values,

$$\left|A^{\pi_\theta}_{\lambda,\beta,t}(s_t,a_t) - A^{\pi_\theta}_{\lambda,t}(s_t,a_t)\right| \le \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell \left|\delta^{V^{\pi_\theta}_{\lambda,\beta}}_{t+\ell} - \delta^{V^{\pi_\theta}_{\lambda}}_{t+\ell}\right|.$$

Now apply Lemma E.1 (Eq. equation 11) to each term: for every $\ell \ge 0$,

$$\left|\delta^{V^{\pi_\theta}_{\lambda,\beta}}_{t+\ell} - \delta^{V^{\pi_\theta}_{\lambda}}_{t+\ell}\right| \le (1+\gamma)\|e\|_\infty,$$

so

$$\begin{aligned}\left|A^{\pi_\theta}_{\lambda,\beta,t}(s_t,a_t) - A^{\pi_\theta}_{\lambda,t}(s_t,a_t)\right| &\le \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell\,(1+\gamma)\|e\|_\infty \\ &= (1+\gamma)\|e\|_\infty \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell.\end{aligned}$$

Since $\gamma \in (0,1)$ and $\lambda \in [0,1)$, we have $0 \le \gamma\lambda < 1$, and the geometric series converges:

$$\sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell = \frac{1}{1-\gamma\lambda}.$$

Substituting yields

$$\left|A^{\pi_\theta}_{\lambda,\beta,t}(s_t,a_t) - A^{\pi_\theta}_{\lambda,t}(s_t,a_t)\right| \le \frac{1+\gamma}{1-\gamma\lambda}\,\|e\|_\infty,$$

which is exactly equation 12. □

The constrained trust-region performance bound also depends on a worst-case expected absolute TD residual. We therefore next show that computing this residual with the integrated critic inflates the nominal residual by at most the same value-error-dependent slack.

Lemma E.3 (Worst-case residual inflation)**.** *Recall $\epsilon^{\pi_\theta,\pi_{\theta'}}$ denotes the worst-case expected absolute TD residual of $V^{\pi_{\theta'}}$ under actions sampled from π_θ, and let $\epsilon^{\pi_\theta,\pi_{\theta'}}_{\lambda,\beta}$ denote the same quantity computed with $V^{\pi_{\theta'}}_{\lambda,\beta}$. Then*

$$\epsilon^{\pi_\theta,\pi_{\theta'}}_{\lambda,\beta} \le \epsilon^{\pi_\theta,\pi_{\theta'}} + (1+\gamma)\|e\|_\infty. \tag{13}$$

Proof. Fix any timestep $t \in \mathbb{N}$ and state $s_t \in \mathcal{S}$. Define the TD residuals

$$\begin{aligned}\delta^{V^{\pi_{\theta'}}}_t &:= x(s_t,a_t) + \gamma V^{\pi_{\theta'}}(s_{t+1}) - V^{\pi_{\theta'}}(s_t),\\ \delta^{V^{\pi_{\theta'}}_{\lambda,\beta}}_t &:= x(s_t,a_t) + \gamma V^{\pi_{\theta'}}_{\lambda,\beta}(s_{t+1}) - V^{\pi_{\theta'}}_{\lambda,\beta}(s_t),\end{aligned}$$

where $x(s,a)$ denotes either the reward or a cost function.

Let

$$e(s_t) := V^{\pi_{\theta'}}_{\lambda,\beta}(s_t) - V^{\pi_{\theta'}}(s_t).$$

Then for any transition (s_t, a_t, s_{t+1}),

$$\begin{aligned}\delta^{V^{\pi_{\theta'}}_{\lambda,\beta}}_t &= x(s_t,a_t) + \gamma V^{\pi_{\theta'}}_{\lambda,\beta}(s_{t+1}) - V^{\pi_{\theta'}}_{\lambda,\beta}(s_t)\\ &= x(s_t,a_t) + \gamma\big(V^{\pi_{\theta'}}(s_{t+1}) + e(s_{t+1})\big) - \big(V^{\pi_{\theta'}}(s_t) + e(s_t)\big)\\ &= \underbrace{x(s_t,a_t) + \gamma V^{\pi_{\theta'}}(s_{t+1}) - V^{\pi_{\theta'}}(s_t)}_{=\,\delta^{V^{\pi_{\theta'}}}_t} + \underbrace{\gamma e(s_{t+1}) - e(s_t)}_{:=\,\Delta_{V,t}}.\end{aligned}$$

Taking absolute values and applying the triangle inequality,

$$\left|\delta^{V^{\pi_{\theta'}}_{\lambda,\beta}}_t\right| \le \left|\delta^{V^{\pi_{\theta'}}}_t\right| + |\Delta_{V,t}|.$$

Taking expectation over $a_t \sim \pi_\theta(\cdot \mid s_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$,

$$\mathbb{E}\left[\left|\delta_t^{V^{\pi_{\theta'}}_{\lambda,\beta}}\right|\right] \leq \mathbb{E}\left[\left|\delta_t^{V^{\pi_{\theta'}}}\right|\right] + \mathbb{E}[|\Delta_{V,t}|]\,.$$

For any realization,

$$|\Delta_{V,t}| = |\gamma e(s_{t+1}) - e(s_t)| \leq (1+\gamma)\|e\|_\infty,$$

so

$$\mathbb{E}[|\Delta_{V,t}|] \leq (1+\gamma)\|e\|_\infty.$$

Thus,

$$\mathbb{E}\left[\left|\delta_t^{V^{\pi_{\theta'}}_{\lambda,\beta}}\right|\right] \leq \mathbb{E}\left[\left|\delta_t^{V^{\pi_{\theta'}}}\right|\right] + (1+\gamma)\|e\|_\infty.$$

Taking the supremum over $s_t \in \mathcal{S}$ and $t \in \mathbb{N}$ and applying the definitions of $\epsilon_{\lambda,\beta}^{\pi_\theta,\pi_{\theta'}}$ and $\epsilon^{\pi_\theta,\pi_{\theta'}}$ yields

$$\epsilon_{\lambda,\beta}^{\pi_\theta,\pi_{\theta'}} \leq \epsilon^{\pi_\theta,\pi_{\theta'}} + (1+\gamma)\|e\|_\infty.$$

which proves equation 13. □

We now combine the preceding perturbation bounds with the nominal λ-return constrained trust-region performance inequality. The resulting theorem shows that the integrated critic preserves the lower bound on reward improvement, up to additional slack terms induced by the bounded critic perturbation.

Theorem 3.2 (restated). *Fix a stationary reference policy $\pi_{\theta'}$ and a candidate policy π_θ. Assume $\tilde{\gamma} + \beta L_\Delta < 1$ and let $V_{\lambda,\beta}^{\pi_{\theta'}}$ denote the unique fixed point of the integrated critic operator for $\pi_{\theta'}$. Let $A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a)$ denote the λ-return reward advantage computed using $V_{r,\lambda,\beta}^{\pi_{\theta'}}$ (equivalently, using TD residuals induced by the integrated fixed point). Then the reward performance difference satisfies*

$$J_r(\pi_\theta) - J_r(\pi_{\theta'}) \;\geq\; \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d_\lambda^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\Bigg[A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a) \;-\; \frac{(1+\gamma)\eta}{(1-\gamma\lambda)\big(1-(\tilde{\gamma}+\beta L_\Delta)\big)} \\ -\; \frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\left(\epsilon_r^{\pi_\theta,\pi_{\theta'}} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}\right)\Bigg].$$

Proof. Recall the λ-return policy performance difference bound equation 2

$$J_r(\pi_\theta) - J_r(\pi_{\theta'}) \;\geq\; \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d_\lambda^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\left[A_{r,\lambda}^{\pi_{\theta'}}(s,a) \;-\; \frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\epsilon_r^{\pi_\theta,\pi_{\theta'}}\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\right].$$

Applying the reward performance bound with the new worst-case one-step residual $\epsilon_{\lambda,\beta}^{\pi_\theta,\pi_{\theta'}}$ (the residual defined using $V_{\lambda,\beta}^{\pi_{\theta'}}$), we obtain

$$J_r(\pi_\theta) - J_r(\pi_{\theta'}) \;\geq\; \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d_\lambda^{\pi_{\theta'}},\, a\sim\pi_\theta(\cdot|s)}\left[A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a) \;-\; \frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\epsilon_{r,\beta}^{\pi_\theta,\pi_{\theta'}}\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\right]. \tag{14}$$

We now relate the integration framework's quantities to their true on-policy counterparts.

First, by Lemma E.2, which controls advantage perturbation under a bounded value shift, we have for all (s,a)

$$A_{r,\lambda}^{\pi_{\theta'}}(s,a) \;\geq\; A_{r,\lambda,\beta}^{\pi_{\theta'}}(s,a) - \frac{1+\gamma}{1-\gamma\lambda}\,\|e\|_\infty. \tag{15}$$

Next, by Lemma E.3,

$$\epsilon_{\lambda,\beta}^{\pi_\theta,\pi_{\theta'}} \;\leq\; \epsilon^{\pi_\theta,\pi_{\theta'}} + (1+\gamma)\|e\|_\infty. \tag{16}$$

Substituting equation 15 and equation 16 into equation 14 introduces slack terms inside the expectation.

Finally, Theorem 3.1 provides the uniform bound

$$\|e\|_\infty = \|V^{\pi_{\theta'}}_{\lambda,\beta} - V^{\pi_{\theta'}}\|_\infty \;\le\; \frac{\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}.$$

Substituting this bound on $\|e\|_\infty$ yields the asserted bound, completing the proof. □

The same argument applies to costs, with the direction of the performance inequality reversed. This yields an upper bound on cost increase under the integrated critic, again with explicit slack terms determined by the critic perturbation.

Corollary E.4 (Cost performance difference upper bound)**.** *Under the same conditions as Theorem 3.2, for any cost function c_i, the performance difference satisfies*

$$J_{c_i}(\pi_\theta) - J_{c_i}(\pi_{\theta'}) \;\le\; \frac{1}{1-\tilde{\gamma}}\mathbb{E}_{s\sim d^{\pi_{\theta'}}_\lambda,\, a\sim\pi_\theta(\cdot|s)}\Bigg[A^{\pi_{\theta'}}_{c_i,\lambda,\beta}(s,a) \;+\; \frac{(1+\gamma)\eta}{(1-\gamma\lambda)\big(1-(\tilde{\gamma}+\beta L_\Delta)\big)} + \frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\left(\epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}\right)\Bigg]. \tag{17}$$

Proof. The result follows directly from Theorem 3.2 by replacing the reward function with c_i and applying the corresponding upper bound form of the performance difference inequality. The same perturbation arguments (Lemmas E.2 and E.3) and fixed-point bound apply, with the contraction modulus $(\tilde{\gamma}+\beta L_\Delta)$. □

Finally, when the reference policy is feasible, the cost difference bound can be converted into an absolute bound on the candidate policy's constraint violation. This makes the trust-region implication explicit: feasibility is preserved up to a controlled additive slack.

Corollary 3.3 (restated) *Suppose the reference policy $\pi_{\theta'}$ is feasible for constraint i, i.e., $J_{c_i}(\pi_{\theta'}) \le d_i$. Then, under the same assumptions as Theorem 3.2, the cost of the candidate policy satisfies*

$$J_{c_i}(\pi_\theta) \;\le\; d_i \;+\; \frac{1}{1-\tilde{\gamma}}\Bigg\{\left(1+\frac{2\tilde{\gamma}\big(\gamma\lambda(|\mathcal{S}|-1)+1\big)\sqrt{2\delta}}{(1-\tilde{\gamma})(1-\gamma\lambda)}\right)\left(\epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)}\right) + \frac{(1+\gamma)\eta}{(1-\gamma\lambda)\big(1-(\tilde{\gamma}+\beta L_\Delta)\big)}\Bigg\}. \tag{18}$$

Proof. Starting from Corollary E.4,

$$J_{c_i}(\pi_\theta) = J_{c_i}(\pi_{\theta'}) + \big(J_{c_i}(\pi_\theta) - J_{c_i}(\pi_{\theta'})\big).$$

Using feasibility of the reference policy, $J_{c_i}(\pi_{\theta'}) \le d_i$, it suffices to upper bound the performance difference term.

Applying Corollary E.4 and bounding the expected advantage by its supremum,

$$\mathbb{E}_{s\sim d^{\pi_{\theta'}}_\lambda,\, a\sim\pi_\theta(\cdot|s)}\big[A^{\pi_{\theta'}}_{c_i,\lambda,\beta}(s,a)\big] \le \epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i,\lambda,\beta},$$

and then using the Lemma E.3,

$$\epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i,\lambda,\beta} \;\le\; \epsilon^{\pi_\theta,\pi_{\theta'}}_{c_i} + \frac{(1+\gamma)\eta}{1-(\tilde{\gamma}+\beta L_\Delta)},$$

yields the result after grouping terms. □

Remark E.5. *All theoretical results for the constrained trust region approaches apply to both unconstrained and constrained trust-region settings.*

Remark E.6. *The $|\mathcal{S}|$ dependence in Theorem 3.2 and Corollary 3.3 arises from the visitation-distribution bound used with GAE and can render the resulting policy-performance bound vacuous in continuous-state settings. This dependence is not intrinsic to the trust-region analysis itself: when the exact advantage is used, corresponding to $\lambda = 0$, the $|\mathcal{S}|$ term vanishes (Yang et al., 2022).*

F SIF LAGRANGIAN LOCAL CONVERGENCE

We now show that if the underlying primal–dual Lagrangian method admits a locally asymptotically stable saddle point in the nominal case, then replacing the exact critic by the integrated critic preserves convergence up to a neighborhood whose radius is controlled by the critic perturbation.

Recall

$$z := (\theta, \mu) \in \mathbb{R}^{d_\theta + d_\mu},$$

where d_θ is the dimension of the policy parameter vector θ and d_μ is the number of constraints (i.e., the dimension of the dual variable μ). Recall

$$g(z) := \begin{bmatrix} g_\theta(z) \\ g_\mu(z) \end{bmatrix},$$

denoting the mean primal–dual drift of the underlying Lagrangian method, where g_θ is the expected policy update and g_μ is the expected multiplier update. Let

$$g_\beta(z) := \begin{bmatrix} g_{\theta,\beta}(z) \\ g_{\mu,\beta}(z) \end{bmatrix}$$

denote the corresponding drift when the integrated critic is used.

We will use the nominal equilibrium

$$z^\star := (\theta^\star, \mu^\star),$$

which corresponds to a locally asymptotically stable saddle point of the nominal Lagrangian dynamics. Throughout this section, we work locally around an *interior* saddle point, so that no projection on the multiplier constraint is active in a sufficiently small $\mathcal{U}$ of $z^\star$ and the system is smooth.

In contrast to the trust-region performance-difference analysis, the primal–dual result operates directly on the on-policy local dynamics and therefore does not require an explicit KL mismatch term. The integrated critic instead manifests as a bounded perturbation of the nominal primal–dual drift field, enabling a local stability analysis around the nominal saddle point, which we show in this section.

This below lemma isolates the approximation error introduced by using GAE in place of the true advantage, even when the exact value function is used. The key idea is that both the GAE estimator and the true advantage can be controlled by the expected magnitude of the on-policy TD residual.

Remark F.1. *The results in §F rely on standard local regularity and stability assumptions on the underlying primal–dual dynamics mentioned in §3. The assumed neighborhood $\mathcal{U}$ is an analytical object whose radius generally cannot be computed a priori and is therefore not an implementation parameter. The bounded-score assumption requires $|\nabla_\theta \log \pi_\theta(a \mid s)|$ to remain finite over this neighborhood, which depends on the policy class and may not hold for arbitrary policies, such as distributions with unbounded scores. The Hurwitz-Jacobian assumption requires the nominal primal–dual equilibrium to be locally stable and likewise may not hold for arbitrary CMDPs or equilibria. The bounded-dual assumption requires the multiplier to remain finite within $\mathcal{U}$, as is standard in primal–dual convergence analyses. Thus, the result is a local robustness guarantee under standard conditions rather than a global guarantee for arbitrary policies or dynamics (Konda & Tsitsiklis, 2003; Zhang & Zheng, 2006; Tessler et al., 2019; Ding et al., 2020).*

Lemma F.2 (Expected worst-case GAE deviation under TD-residual error)**.** *Let $A^{\pi_\theta}_{\lambda,t}$ denote the GAE advantage, and let A^{π_θ} denote the true advantage. Let the TD residuals satisfy*

$$\epsilon_t^{\pi_\theta,\pi_\theta} := \sup_{s_t \in \mathcal{S}} \mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t),\, s_{t+1} \sim P(\cdot|s_t,a_t)} \left[\left| \delta_t^{V^{\pi_\theta}} \right| \right], \qquad \epsilon^{\pi_\theta,\pi_\theta} := \sup_{t \in \mathbb{N}} \epsilon_t^{\pi_\theta,\pi_\theta}.$$

Then

$$\sup_{s_t \in \mathcal{S}} \mathbb{E}_{a_{t:\infty} \sim \pi_\theta,\, s_{t+1:\infty} \sim P} \left[\left| A^{\pi_\theta}_{\lambda,t} - A^{\pi_\theta}(s_t, a_t) \right| \right] \le \frac{2 - \gamma\lambda}{1 - \gamma\lambda} \epsilon^{\pi_\theta,\pi_\theta}.$$

Proof. Recall that

$$A^{\pi_\theta}_{\lambda,t} = \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \delta^{V^{\pi_\theta}}_{t+\ell}.$$

Fix an arbitrary state s_t. Taking expectation over $a_{t:\infty} \sim \pi_\theta$ and $s_{t+1:\infty} \sim P$, the exact-critic GAE estimator has zero expectation under the on-policy action distribution conditional on s_t:

$$\mathbb{E}_{a_{t:\infty}\sim\pi_\theta,\, s_{t+1:\infty}\sim P}\left[A^{\pi_\theta}_{\lambda,t}\right] = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[A^{\pi_\theta}(s_t,a_t)\right] = 0.$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{a_{t:\infty}\sim\pi_\theta,\, s_{t+1:\infty}\sim P}\left[\left|A^{\pi_\theta}_{\lambda,t} - A^{\pi_\theta}(s_t,a_t)\right|\right]\\ &\le \mathbb{E}_{a_{t:\infty}\sim\pi_\theta,\, s_{t+1:\infty}\sim P}\left[\left|A^{\pi_\theta}_{\lambda,t}\right|\right] + \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[|A^{\pi_\theta}(s_t,a_t)|\right].\end{aligned}$$

For any fixed (s_t, a_t), we have

$$A^{\pi_\theta}(s_t,a_t) = \mathbb{E}_{s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\delta_t^{V^{\pi_\theta}}\right],$$

Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[|A^{\pi_\theta}(s_t,a_t)|\right] &\le \mathbb{E}_{\substack{a_t\sim\pi_\theta(\cdot|s_t),\\ s_{t+1}\sim P(\cdot|s_t,a_t)}}\left[\left|\delta_t^{V^{\pi_\theta}}\right|\right]\\ &\le \epsilon^{\pi_\theta,\pi_\theta}.\end{aligned}$$

Also,

$$\begin{aligned}&\mathbb{E}_{a_{t:\infty}\sim\pi_\theta,\, s_{t+1:\infty}\sim P}\left[\left|A^{\pi_\theta}_{\lambda,t}\right|\right]\\ &\le \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell \mathbb{E}_{a_{t:t+\ell}\sim\pi_\theta,\, s_{t+1:t+\ell+1}\sim P}\left[\left|\delta_{t+\ell}^{V^{\pi_\theta}}\right|\right]\\ &\le \sum_{\ell=0}^{\infty}(\gamma\lambda)^\ell \epsilon^{\pi_\theta,\pi_\theta} = \frac{1}{1-\gamma\lambda}\epsilon^{\pi_\theta,\pi_\theta}.\end{aligned}$$

Combining the two bounds gives

$$\mathbb{E}_{a_{t:\infty}\sim\pi_\theta,\, s_{t+1:\infty}\sim P}\left[\left|A^{\pi_\theta}_{\lambda,t} - A^{\pi_\theta}(s_t,a_t)\right|\right] \le \frac{2-\gamma\lambda}{1-\gamma\lambda}\epsilon^{\pi_\theta,\pi_\theta}.$$

Taking the supremum over s_t completes the proof. □

We next translate critic-induced advantage errors into an error bound on the primal–dual drift field. Since the policy update is a score-function expectation weighted by reward and cost advantages, bounded advantage perturbations imply a bounded perturbation of the primal drift, while the dual drift is unchanged when it uses raw cost returns.

Lemma F.3 (Primal-dual drift error relative to the true advantages)**.** *Assume*

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\| \le G \qquad \forall (s,a,\theta) \in \mathcal{U},$$

and suppose that, on $\mathcal{U}$, the multiplier satisfies

$$0 \le \mu \le M_\mu.$$

For $x \in \{r, c\}$, define the Bellman error

$$\epsilon_{x,t}^{\pi_\theta,\pi_\theta} := \sup_{s_t\in\mathcal{S}} \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t),\, s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\left|\delta_t^{V_x^{\pi_\theta}}\right|\right], \qquad \epsilon_x^{\pi_\theta,\pi_\theta} := \sup_{t\in\mathbb{N}} \epsilon_{x,t}^{\pi_\theta,\pi_\theta}.$$

Then

$$\begin{aligned}\sup_{z\in\mathcal{U}} \|g_\beta(z) - g(z)\| \le{}& \frac{G}{1-\gamma\lambda}\left((1+\gamma)\|e_r\|_\infty + (2-\gamma\lambda)\epsilon_r^{\pi_\theta,\pi_\theta}\right)\\ &+ \frac{GM_\mu}{1-\gamma\lambda}\left((1+\gamma)\|e_c\|_\infty + (2-\gamma\lambda)\epsilon_c^{\pi_\theta,\pi_\theta}\right).\end{aligned}$$

where g_β denotes the primal-dual drift obtained using the corrected GAE reward and cost advantages $A^{\pi_\theta}_{r,\lambda,\beta}$ and $A^{\pi_\theta}_{c,\lambda,\beta}$, while g denotes the primal-dual drift obtained using the true reward and cost advantages $A_r^{\pi_\theta}$ and $A_c^{\pi_\theta}$.

Proof. Fix $z = (\theta, \mu) \in \mathcal{U}$. All expectations are taken with respect to $s \sim d^{\pi_\theta}$ and $a \sim \pi_\theta(\cdot \mid s)$, unless otherwise specified. Expectations inside GAE quantities are taken over trajectories generated by π_θ.

By definition,

$$g_{\theta,\beta}(z) = \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a \mid s) \Big(A^{\pi_\theta}_{r,\lambda,\beta}(s,a) - \mu A^{\pi_\theta}_{c,\lambda,\beta}(s,a)\Big)\Big],$$

whereas

$$g_\theta(z) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) \left(A^{\pi_\theta}_r(s,a) - \mu A^{\pi_\theta}_c(s,a)\right)].$$

Therefore,

$$\begin{aligned} g_{\theta,\beta}(z) - g_\theta(z) = \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a \mid s) \Big(A^{\pi_\theta}_{r,\lambda,\beta}(s,a) - A^{\pi_\theta}_r(s,a)\Big)\Big] \\ - \mu \mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a \mid s) \Big(A^{\pi_\theta}_{c,\lambda,\beta}(s,a) - A^{\pi_\theta}_c(s,a)\Big)\Big]. \end{aligned}$$

Taking norms and using the score bound and $0 \le \mu \le M_\mu$ gives

$$\begin{aligned} \|g_{\theta,\beta}(z) - g_\theta(z)\| \le G\, \mathbb{E}\Big[\Big|A^{\pi_\theta}_{r,\lambda,\beta}(s,a) - A^{\pi_\theta}_r(s,a)\Big|\Big] \\ + G M_\mu\, \mathbb{E}\Big[\Big|A^{\pi_\theta}_{c,\lambda,\beta}(s,a) - A^{\pi_\theta}_c(s,a)\Big|\Big]. \end{aligned}$$

For each $x \in \{r, c\}$, decompose

$$A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_x(s,a) = \Big(A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_{x,\lambda}(s,a)\Big) + \Big(A^{\pi_\theta}_{x,\lambda}(s,a) - A^{\pi_\theta}_x(s,a)\Big).$$

By the triangle inequality,

$$\begin{aligned} \mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_x(s,a)\Big|\Big] \le \mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_{x,\lambda}(s,a)\Big|\Big] \\ + \mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda}(s,a) - A^{\pi_\theta}_x(s,a)\Big|\Big]. \end{aligned}$$

By Lemma E.2,

$$\mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_{x,\lambda}(s,a)\Big|\Big] \le \frac{1+\gamma}{1-\gamma\lambda}\|e_x\|_\infty.$$

By Lemma F.2,

$$\mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda}(s,a) - A^{\pi_\theta}_x(s,a)\Big|\Big] \le \frac{2-\gamma\lambda}{1-\gamma\lambda}\epsilon_x^{\pi_\theta,\pi_\theta}.$$

Thus,

$$\mathbb{E}\Big[\Big|A^{\pi_\theta}_{x,\lambda,\beta}(s,a) - A^{\pi_\theta}_x(s,a)\Big|\Big] \le \frac{1}{1-\gamma\lambda}\left((1+\gamma)\|e_x\|_\infty + (2-\gamma\lambda)\epsilon_x^{\pi_\theta,\pi_\theta}\right).$$

Applying this bound with $x = r$ and $x = c$ gives

$$\begin{aligned} \|g_{\theta,\beta}(z) - g_\theta(z)\| \le \frac{G}{1-\gamma\lambda}\left((1+\gamma)\|e_r\|_\infty + (2-\gamma\lambda)\epsilon_r^{\pi_\theta,\pi_\theta}\right) \\ + \frac{G M_\mu}{1-\gamma\lambda}\left((1+\gamma)\|e_c\|_\infty + (2-\gamma\lambda)\epsilon_c^{\pi_\theta,\pi_\theta}\right). \end{aligned} \tag{19}$$

Now consider the dual component. Since the dual update is computed using raw cost returns rather than the corrected cost advantage, the correction affects only the primal policy-gradient component. Hence,

$$g_{\mu,\beta}(z) = g_\mu(z),$$

and therefore

$$\|g_{\mu,\beta}(z) - g_\mu(z)\| = 0.$$

Taking the supremum of Eq equation 19 over $z \in \mathcal{U}$ completes the proof. □

We now show that this bounded drift perturbation only moves the asymptotic limit from the exact saddle point to a neighborhood whose radius is $O(e_{\text{sif}})$, where e_{sif} is the bound defined in Eq equation 19 (Lemma F.3).

First, in Theorem F.4 we make the assumption that there exists a neighborhood $\mathcal{U}$ of $z^\star$ such that: g is continuously differentiable on $\mathcal{U}$; $g(z^\star) = 0$; the Jacobian $J^\star := \nabla g(z^\star)$ is Hurwitz.

This is the standard local exponential stability condition for the mean primal-dual ODE $\dot{z} = g(z)$, which arises as the continuous-time limit of the discrete primal-dual updates $z_{k+1} = z_k + \alpha\, g(z_k)$ as $\alpha \to 0$. This condition ensures a uniform contraction toward $z^\star$, so that the bounded perturbations induced by the SIF critic cannot accumulate and instead yield convergence to a neighborhood whose radius scales with the perturbation magnitude. In particular, we distinguish between the nominal dynamics $\dot{z} = g(z)$ and the integration-perturbed dynamics $\dot{z}_\beta = g_\beta(z_\beta)$, which in general generate different trajectories.

Theorem F.4 (Convergence of the perturbed primal-dual dynamics to a neighborhood of the saddle point)**.** *Assume g is continuously differentiable on $\mathcal{U}$, $g(z^\star) = 0$, and the Jacobian $J^\star := \nabla g(z^\star)$ is Hurwitz.*

Now let the perturbed drift g_β satisfy

$$\sup_{z \in \mathcal{U}} \|g_\beta(z) - g(z)\| \le e_{\text{sif}}$$

for some $e_{\text{sif}} \ge 0$. Then there exist constants $a_1, a_2, r, \varepsilon_0 > 0$ such that for every $e_{\text{sif}} \le \varepsilon_0$ and every initial condition $z_\beta(0) \in B_r(z^\star)$, the solution of the perturbed dynamics

$$\dot{z}_\beta = g_\beta(z_\beta)$$

remains in $\mathcal{U}$ and satisfies

$$\|z_\beta(t) - z^\star\| \le a_1 e^{-a_2 t} \|z_\beta(0) - z^\star\| + \frac{a_1}{a_2}\, e_{\text{sif}}, \qquad t \ge 0. \tag{20}$$

In particular,

$$\limsup_{t \to \infty} \|z_\beta(t) - z^\star\| \le \frac{a_1}{a_2}\, e_{\text{sif}}. \tag{21}$$

Thus, the SIF-perturbed primal-dual dynamics converge to an $O(e_{\text{sif}})$ neighborhood of the nominal saddle point.

Proof. Since $J^\star$ is Hurwitz, the Lyapunov equation

$$(J^\star)^\top P + P J^\star = -I$$

admits a unique symmetric positive definite solution $P \succ 0$. Define the local Lyapunov function

$$W(z) := (z - z^\star)^\top P (z - z^\star).$$

Because g is continuously differentiable and $g(z^\star) = 0$, we may write a first-order Taylor expansion of g around $z^\star$:

$$g(z) = J^\star (z - z^\star) + R(z),$$

where the remainder $R(z)$ is given by

$$R(z) = \int_0^1 \left(\nabla g(z^\star + \tau(z - z^\star)) - J^\star \right) (z - z^\star)\, d\tau.$$

If ∇g is locally Lipschitz on $\mathcal{U}$, then there exists $L > 0$ such that

$$\|\nabla g(x) - \nabla g(y)\| \le L \|x - y\| \quad \text{for all } x, y \in \mathcal{U},$$

and hence

$$\|R(z)\| \le c_R \|z - z^\star\|^2, \qquad c_R := \frac{L}{2}.$$

Let

$$e(z) := g_\beta(z) - g(z), \qquad \|e(z)\| \le e_{\text{sif}}.$$

Along the perturbed trajectory $z_\beta(t)$ satisfying

$$\dot{z}_\beta = g_\beta(z_\beta) = g(z_\beta) + e(z_\beta),$$

we have

$$\begin{aligned}\dot{W}(z_\beta) &= 2(z_\beta - z^\star)^\top P\big(g(z_\beta) + e(z_\beta)\big) \\ &= 2(z_\beta - z^\star)^\top P\big(J^\star(z_\beta - z^\star)\big) + 2(z_\beta - z^\star)^\top PR(z_\beta) + 2(z_\beta - z^\star)^\top Pe(z_\beta).\end{aligned}$$

Using the Lyapunov equation,

$$2(z_\beta - z^\star)^\top PJ^\star(z_\beta - z^\star) = -\|z_\beta - z^\star\|^2.$$

Therefore,

$$\dot{W}(z_\beta) \le -\|z_\beta - z^\star\|^2 + 2\|P\|\,\|z_\beta - z^\star\|\,\|R(z_\beta)\| + 2\|P\|\,\|z_\beta - z^\star\|\,e_{\mathrm{sif}}.$$

For z_β in a sufficiently small neighborhood of $z^\star$,

$$\|R(z_\beta)\| \le c_R\|z_\beta - z^\star\|^2,$$

so shrinking the neighborhood if necessary gives

$$2\|P\|\,\|z_\beta - z^\star\|\,\|R(z_\beta)\| \le \frac{1}{2}\|z_\beta - z^\star\|^2.$$

Hence,

$$\dot{W}(z_\beta) \le -\frac{1}{2}\|z_\beta - z^\star\|^2 + 2\|P\|\,\|z_\beta - z^\star\|\,e_{\mathrm{sif}}.$$

Using Young's inequality, for any $\alpha > 0$,

$$2\|P\|\,\|z_\beta - z^\star\|\,e_{\mathrm{sif}} \le \frac{\alpha}{2}\|z_\beta - z^\star\|^2 + \frac{2\|P\|^2}{\alpha}e_{\mathrm{sif}}^2.$$

Choosing $\alpha = \frac{1}{2}$ yields

$$\dot{W}(z_\beta) \le -\frac{1}{4}\|z_\beta - z^\star\|^2 + 4\|P\|^2 e_{\mathrm{sif}}^2.$$

Since $P \succ 0$, there exist $m_P, M_P > 0$ such that

$$m_P\|z_\beta - z^\star\|^2 \le W(z_\beta) \le M_P\|z_\beta - z^\star\|^2.$$

Thus,

$$\dot{W}(z_\beta) \le -\frac{1}{4M_P}W(z_\beta) + 4\|P\|^2 e_{\mathrm{sif}}^2.$$

By Grönwall's inequality,

$$W(z_\beta(t)) \le e^{-t/(4M_P)}W(z_\beta(0)) + 16M_P\|P\|^2\,e_{\mathrm{sif}}^2.$$

Taking square roots and redefining constants gives

$$\|z_\beta(t) - z^\star\| \le a_1 e^{-a_2 t}\|z_\beta(0) - z^\star\| + \frac{a_1}{a_2}e_{\mathrm{sif}},$$

which proves equation 20. The limsup bound equation 21 follows by letting $t \to \infty$. □

Combining Theorem F.4 with Lemma E.2 immediately yields a neighborhood radius controlled by the clipping level.

Corollary 3.4 (restated) *Under the assumptions of Theorem F.4 and Lemma F.3, there exist constants $a_1, a_2, G > 0$ such that, for all trajectories initialized sufficiently close to $z^\star$, define, for $x \in \{r, c\}$,*

$$\|e_x\|_\infty := \frac{\eta_x}{1 - (\tilde{\gamma} + \beta_x L_{\Delta,x})}, \qquad E_x := \frac{1}{1 - \gamma\lambda}\left((1+\gamma)\|e_x\|_\infty + (2 - \gamma\lambda)\epsilon_x^{\pi_\theta,\pi_\theta}\right).$$

Then, for all trajectories initialized sufficiently close to $z^\star$,

$$\limsup_{t\to\infty}\|z_\beta(t) - z^\star\| \le \frac{Ga_1}{a_2}\left(E_r + M_\mu E_c\right).$$

In particular, the asymptotic neighborhood radius scales linearly with the reward and cost SIF-induced GAE errors.

Proof. By Lemma F.3,

$$\sup_{z \in \mathcal{U}} \|g_\beta(z) - g(z)\| \leq GE_r + GM_\mu E_c.$$

Thus we may take

$$e_{\text{sif}} = GE_r + GM_\mu E_c$$

in Theorem F.4. Substituting this into equation 21 gives the result. □

G ADDITIONAL SPEC DETAILS

Algorithm 2 Detailed SPEC-plugin for Local Policy Search

Require: Policy π_θ, critics $\hat{V}^{\pi_\theta}_{r,\lambda}$, $\hat{V}^{\pi_\theta}_{c,\lambda,\beta}$, encoders f_r, f_c, retrieval index $\mathcal{D}_c$, running sample count N_{tot}, warmups N_{enc}, N_{retr}, correction bound $\beta_{\max}$, clip bound η, trajectory bounds n_{step}

Sample experiences $\{(s_t, a_t, r_t, c_t, s_{t+1})\}_{t=0}^{B-1}$ by rolling out π_θ with stacked observations.
Compute standard on-policy targets $V^{\text{targ}}_{r,\lambda}$, $V^{\text{targ}}_{c,\lambda}$ and advantages $\hat{A}^{\pi_\theta}_{r,\lambda}$, $\hat{A}^{\pi_\theta}_{c,\lambda,\beta}$.
if retrieval has started **then**
    Encode queries $f_c(s_t)$ and retrieve k neighbor trajectories from $\mathcal{D}_c$.
    Compute $V^{\text{nb}}_c(s_t)$ by averaging n-step neighbor returns bootstrapped with $\hat{V}^{\pi_\theta}_{c,\lambda,\beta}$.
    Compute $\beta(s_t)$ from retrieved neighbors' IVF cluster assignments.
    Cache $V^{\text{nb}}_c(s_t)$, $\beta_c(s_t)$, and η for the current update.
end if
if natural policy gradient-based method **then**
    Update π_θ ▷ policy update follows the original algorithm
end if
for each critic optimization iteration **do**
    for each minibatch **do**
        $y_r = V^{\text{targ}}_{r,\lambda}$.
        if retrieval has started **then**
            Set $y_c = V^{\text{targ}}_{c,\lambda} + \text{clip}_\eta\Big(\beta(s_t)\,\Big(V^{\text{nb}}_c(s_t) - \text{stopgrad}(\hat{V}^{\pi_\theta}_{c,\lambda,\beta}(s_t))\Big)\Big)$.
        else
            $y_c = V^{\text{targ}}_{c,\lambda}$.
        end if
        Update $\hat{V}^{\pi_\theta}_{r,\lambda}$ and $\hat{V}^{\pi_\theta}_{c,\lambda,\beta}$ with MSE on y_r and y_c.
        if non-natural policy gradient-based method **then**
            Update π_θ ▷ policy update follows the original algorithm
        end if
    end for
end for
Increment N_{tot} by B.
if $N_{\text{tot}} \geq N_{\text{enc}}$ and encoders are not frozen **then**
    Freeze f_r and f_c.
end if
if f_c is frozen **then**
    Store the sampled trajectories as pending entries for $\mathcal{D}_c$.
end if
if $N_{\text{tot}} \geq N_{\text{enc}} + N_{\text{retr}}$ and retrieval has not started **then**
    Train the IVF index and initiate retrieval.
    Flush pending entries into $\mathcal{D}_c$.
end if
if retrieval has started **then**
    Add new pending entries to $\mathcal{D}_c$, dropping oldest entries if capacity is exceeded.
    Anneal $\beta_{\max}$ if enabled.
end if

Remark G.1. *SPEC is applicable to both natural and standard policy gradient methods. In this work, we restrict attention to TRPO-PID, which are natural gradient-based methods, but differ in how the constrained optimization problem is solved.*

H ENVIRONMENT INFORMATION

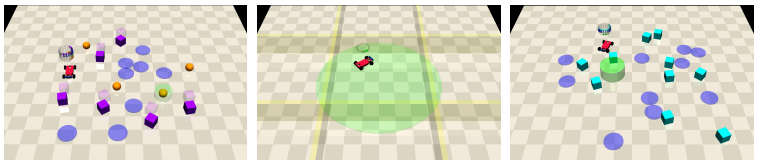
We evaluate SPEC on a diverse and challenging set of continuous control benchmarks from Safety-Gymnasium (Ji et al., 2023), which extends standard reinforcement learning tasks to constrained settings via explicit cost signals (Figure 3).

Safety-Gymnasium. Safety-Gymnasium provides a suite of goal-directed navigation tasks in which an agent must reach a target while avoiding safety violations. We consider the Goal, Button, and Circle tasks, which correspond to the most challenging Level 2 variants.

Goal is the most challenging variant of the goal-reaching task in Safety-Gymnasium, where the agent (Point or Car) must navigate to a goal location while avoiding a dense set of hazards that incur cost upon contact. The reward consists of a shaping term based on the reduction in distance to the goal across consecutive timesteps, together with a positive terminal reward for successfully reaching the goal. Observations are provided as a fixed-dimensional state vector including proprioceptive features and lidar-based sensing of the goal and surrounding obstacles (10 vases and 10 hazards). This environment is challenging due to dense safety constraints and the need to balance efficient goal-directed behavior with strict constraint satisfaction.

Button is the most challenging variant of the button-selection task, where the agent (Doggo or Point) must navigate to and press the correct goal button while avoiding surrounding hazards and gremlins. In the Level 2 setting, the environment contains 4 buttons, 6 gremlins, and 8 hazards, and incurs cost for pressing the wrong button, contacting gremlins, or entering hazardous regions. The reward combines a shaping term based on the reduction in distance to the goal button with a terminal reward for successfully pressing the correct button. Observations include lidar-based sensing over buttons, the goal, gremlins, and hazards, along with proprioceptive features. This environment is particularly challenging due to the need for target identification, obstacle avoidance, and safety-constrained exploration in a cluttered setting.

Circle departs from goal-directed navigation by requiring the agent to follow a predefined circular trajectory while avoiding hazards. The agent (Point or Car) is rewarded based on how well its velocity aligns with the tangent direction of the circle, encouraging smooth and consistent motion rather than convergence to a fixed target. In the Level 2 configuration, the environment increases difficulty by introducing boundaries that incur cost upon contact. Observations include lidar-based sensing of the surrounding environment and proprioceptive features. This environment is particularly challenging because it lacks a terminal goal, requiring sustained trajectory tracking over long horizons while maintaining strict constraint satisfaction.



(a) Button (b) Circle (c) Goal

Figure 3: Illustration of the level 2 Safety-Gymnasium environments used (Ji et al., 2023). We evaluate on a diverse set of challenging tasks spanning the Button, Goal, and Circle families, specifically DoggoButton, PointButton, PointGoal, CarGoal, RacecarCircle, and DoggoCircle.

Observation splitting. For all environments, we decompose observations into environment and agent components to provide a more structured input to our method. In particular, Safety-Gymnasium exposes observations as a concatenation of multiple components, including proprioceptive features and lidar-based sensing. We treat all lidar measurements that capture goals, hazards, obstacles, and other entities as *environment observations* because they reflect exogenous scene structure. All remaining components, such as proprioceptive states (e.g., velocities, joint states, actuator information), are treated as *agent observations*, which vary across agent morphologies (e.g., Point, Car, Doggo, and Racecar). This partition aligns naturally with the organization provided by the Safety-Gymnasium API, which already separates these modalities and allows us to selectively incorporate environment history without redundantly stacking agent-state features.

I ADDITIONAL EXPERIMENTS

This section presents the extended empirical material referenced in Section 5. Appendix I.1 reports the full baseline comparison from which we select the three strongest-performing methods used in the main results, including the omitted vanilla baselines: PCPO, CUP, FOCOPS, PPO-L, and CPPO-P. Appendix I.2 lists and motivates the SPEC hyperparameter choices. Appendix I.3 provides the learning curves and diagnostic plots underlying the ablations summarized in Table 3 of the main paper.

I.1 BASELINES

We implement SPEC and SafeEC using TRPO-P due to its reliable and consistent performance across benchmark tasks. As shown in Figure 4 and Table 5, TRPO-P exhibits stable cost behavior across environments. In the main comparison, we therefore focus on CPO, TRPO-L, and TRPO-P, as these methods demonstrate comparatively stable and competitive performance. The remaining baselines show weaker, less consistent results in Figure 4 and Table 5, and are thus excluded to maintain a focused comparison among the strongest-performing algorithms.

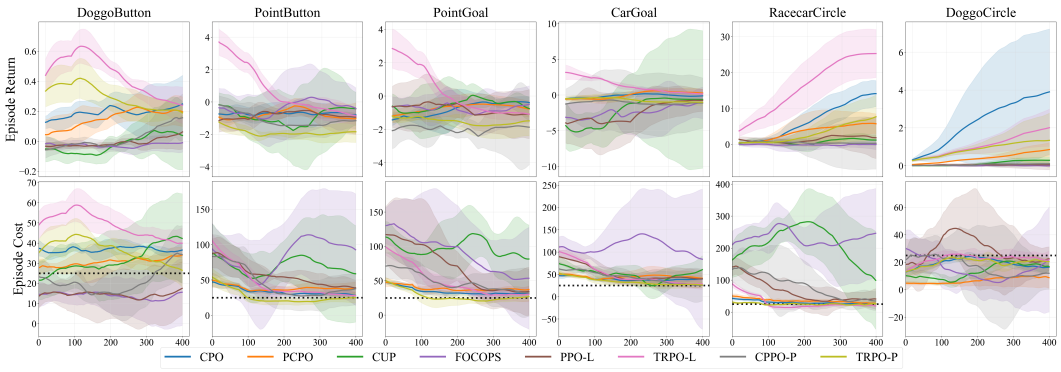


Figure 4: Learning curves for vanilla baselines. Baselines exhibit varying trade-offs between performance and constraint satisfaction across environments.

Table 5: Per-environment IQM $\pm$ std for all eight baseline methods (5 seeds). Shaded cells satisfy IQM cost ≤ 25; bold marks the highest-reward safe method per environment. TRPO-P exhibits the most consistent cost performance across environments, with relatively stable constraint satisfaction compared to other methods. Accordingly, we adopt TRPO-P as the base algorithm for implementing SPEC and SafeEC.

Env	Metric	CPO	PCPO	CUP	FOCOPS	PPO-L	TRPO-L	CPPO-P	TRPO-P
DoggoButton	Reward	0.2 ± 0.1	0.2 ± 0.1	0.1 ± 0.1	-0.0 ± 0.1	$\mathbf{0.2 \pm 0.2}$	0.2 ± 0.1	0.3 ± 0.3	0.1 ± 0.1
	Cost	45.9 ± 3.9	32.2 ± 7.8	32.3 ± 21.3	7.8 ± 5.1	$\mathbf{22.8 \pm 14.7}$	33.3 ± 11.7	61.0 ± 22.8	24.1 ± 6.4
PointButton	Reward	-0.4 ± 0.2	-0.9 ± 1.3	-0.1 ± 0.7	-1.0 ± 1.3	-0.5 ± 1.2	-1.0 ± 0.4	-0.4 ± 1.6	-1.3 ± 0.9
	Cost	30.0 ± 1.5	42.4 ± 15.8	82.2 ± 64.6	63.6 ± 75.0	51.6 ± 30.5	25.3 ± 6.8	48.0 ± 37.7	32.3 ± 11.6
PointGoal	Reward	-0.3 ± 0.3	-0.7 ± 0.5	0.0 ± 0.2	0.0 ± 0.7	-1.1 ± 1.0	-1.3 ± 0.5	$\mathbf{-1.9 \pm 2.0}$	-1.5 ± 1.2
	Cost	34.0 ± 9.3	37.6 ± 10.7	113.5 ± 44.0	115.8 ± 71.4	38.4 ± 44.9	27.8 ± 8.7	$\mathbf{20.3 \pm 5.3}$	27.8 ± 11.6
CarGoal	Reward	0.1 ± 0.7	0.3 ± 0.5	0.0 ± 10.2	0.0 ± 10.4	-0.6 ± 0.6	$\mathbf{0.4 \pm 0.6}$	-0.8 ± 1.1	-1.1 ± 0.7
	Cost	37.2 ± 14.8	42.0 ± 6.5	34.2 ± 65.4	104.0 ± 101.5	39.0 ± 12.9	$\mathbf{30.9 \pm 4.7}$	36.7 ± 21.6	31.8 ± 6.3
RacecarCircle	Reward	14.0 ± 8.2	9.0 ± 4.4	1.1 ± 1.2	-0.1 ± 1.2	1.3 ± 6.2	$\mathbf{25.7 \pm 6.5}$	0.3 ± 0.2	5.9 ± 7.9
	Cost	21.9 ± 8.8	35.2 ± 16.3	81.6 ± 147.3	272.6 ± 78.5	30.7 ± 75.8	$\mathbf{12.5 \pm 6.7}$	29.0 ± 61.7	8.5 ± 4.4
DoggoCircle	Reward	2.3 ± 1.1	1.4 ± 0.8	0.3 ± 0.2	-0.2 ± 0.3	0.1 ± 0.1	2.1 ± 1.1	0.0 ± 0.1	$\mathbf{4.5 \pm 3.0}$
	Cost	26.5 ± 14.2	16.0 ± 11.7	4.6 ± 7.4	26.8 ± 39.6	6.4 ± 24.3	15.5 ± 6.4	12.8 ± 14.6	$\mathbf{13.4 \pm 12.1}$

I.2 SPEC HYPERPARAMETERS

We set the retrieved trajectory horizon to $n_{\text{step}} = 4$ to balance immediate signal fidelity with controlled trajectory propagation. Using a short multi-step horizon allows the critic to incorporate informative near-term trajectory structure while limiting the influence of behavior farther into the future, which may be less relevant to the current state distribution. This provides additional temporal context beyond a one-step correction without allowing retrieved trajectories to propagate excessively far.

We set $\beta_{\max} = 0.15$ to allow for a strong initial correction, since β is annealed to 0 over the course of training. This choice ensures that the retrieval signal has a meaningful influence early on, before being gradually reduced as the policy stabilizes.

For η, we only require it to be sufficiently large to avoid unnecessarily clipping useful corrections. Since the correction is already scaled by β, its effective magnitude is reduced, so overly aggressive clipping is not needed. We therefore choose $\eta = 2$ as a sufficiently large threshold to bound extreme outliers without materially restricting the correction.

We leave a more systematic investigation of these hyperparameters, including the retrieval horizon, initial correction scale, and clipping threshold, to future work.

Table 6: Full set of hyperparameters used for SPEC.

Hyperparameter	**Value**
Clipping threshold η	2
Max correction weight $\beta_{\max}$	0.15
Discount factor γ	0.99
GAE parameter λ	0.95
Effective contraction $\tilde{\gamma}$	< 0.832
Retriever capacity	500,000
IVF n_{list}	1024
IVF n_{probe}	32
Neighbors k	16
Neighbor trajectory n_{step}	4
Observation stack	3 timesteps
Total training steps	2×10^6
Steps per epoch	5000
Encoder dimension	32
Encoder warmup N_{enc}	300,000
Retriever warmup N_{retr}	100,000
Total warmup N_{tot}	400,000
$\beta_{\max}$ schedule	Linear decay to 0

I.3 Additional ablations

The following subsections complement the ablation tables in the main paper (Table 2 and Table 3) with per-environment learning curves over training and additional diagnostics. We report the realized correction-strength trace β over training under the adaptive schedule (Appendix I.3.1). We additionally report encoder warmup (N_{enc}) ablations, including the corresponding table and learning curve, in Appendix I.3.4. All curves report the IQM and 1σ standard deviation across five seeds, matching the convention used in the main results.

I.3.1 ADAPTIVE β

Figure 5 compares SPEC under the adaptive, entropy-aware $\beta(s)$ schedule (Eq. equation 6) against a non-adaptive variant in which β follows the linear decay of $\beta_{\max}$ without state-dependent scaling. The adaptive variant yields slightly more consistent cost performance across environments, while the non-adaptive variant generally attains higher reward (e.g., on the Circle tasks and CarGoal). We find the resulting reward–cost trade-off favorable for the adaptive variant due to its improved consistency in maintaining near-threshold cost performance across environments.

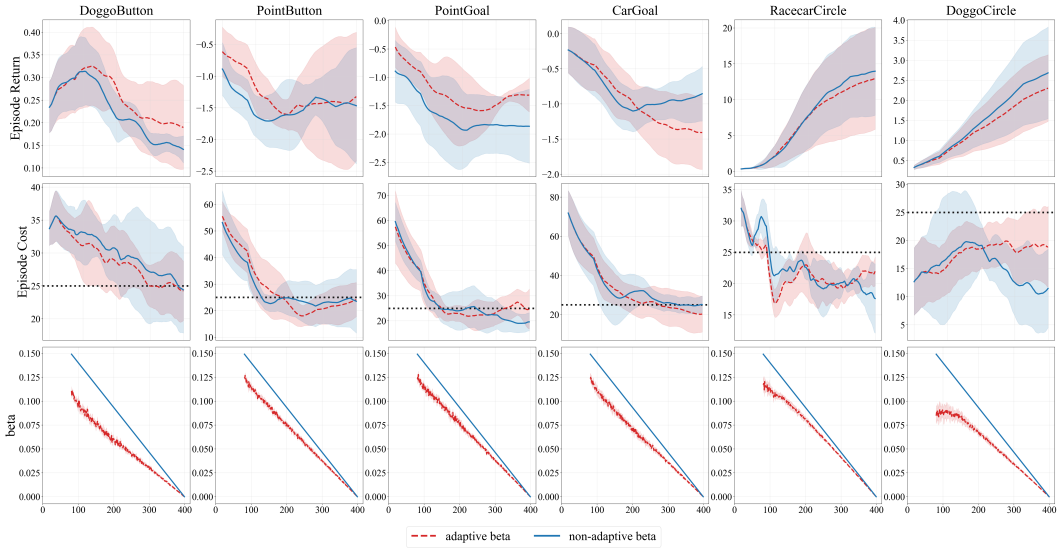


Figure 5: Learning curves and β for the adaptive β ablation.

I.3.2 OBSERVATION STACKING

Figure 6 reports learning curves for the two observation-stacking variants ablated in Table 3: *full* stacking, in which both proprioceptive (o_{agent}) and exogenous (o_{env}) components are retained across the observation history, and *environment-only* stacking, which retains only past o_{env} frames while keeping the current observation o_t in full (Section 4). Environment-only stacking yields more consistent cost performance across environments, while full stacking exhibits substantially larger constraint violations on the (e.g. Button tasks). Despite this, environment-only stacking remains competitive in reward and achieves stronger performance (e.g. Circle tasks).

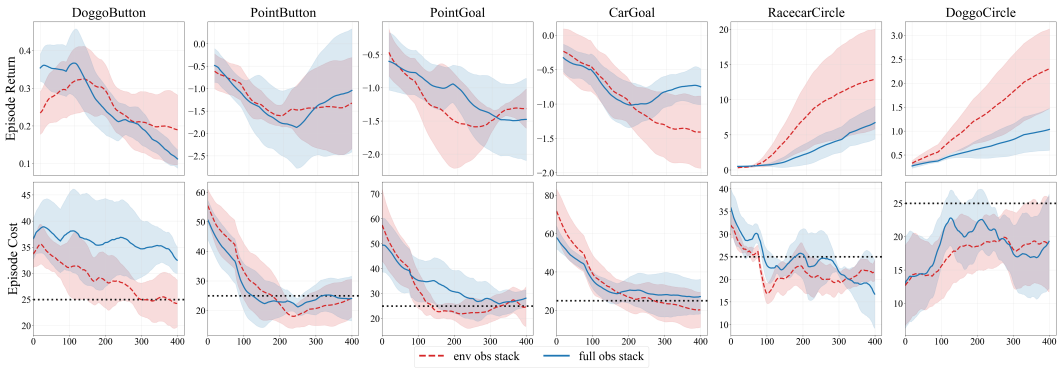


Figure 6: Learning curves for the observation stacking ablation.

I.3.3 $\beta_{\max}$

We ablate $\beta_{\max}$ because the contraction analysis specifies an admissible range for the correction magnitude, but does not determine its empirical effect on the reward–cost trade-off. In SPEC, $\beta_{\max}$ bounds the magnitude of the clipped, state-dependent correction applied to the learned cost critic, thereby controlling the influence of retrieved off-policy experience on optimization. As shown in Table 2 and Figure 7, performance across individual $\beta_{\max}$ values is not strictly monotonic within each environment. This is expected because changing $\beta_{\max}$ affects not only the magnitude of the SIF correction, but also the subsequent evolution of the critic, policy, visitation distribution, and retrieval memory. As a result, intermediate settings need not produce outcomes that interpolate smoothly between weaker and stronger correction strengths. These per-environment variations complement the broader safety–performance trend discussed in the main text.

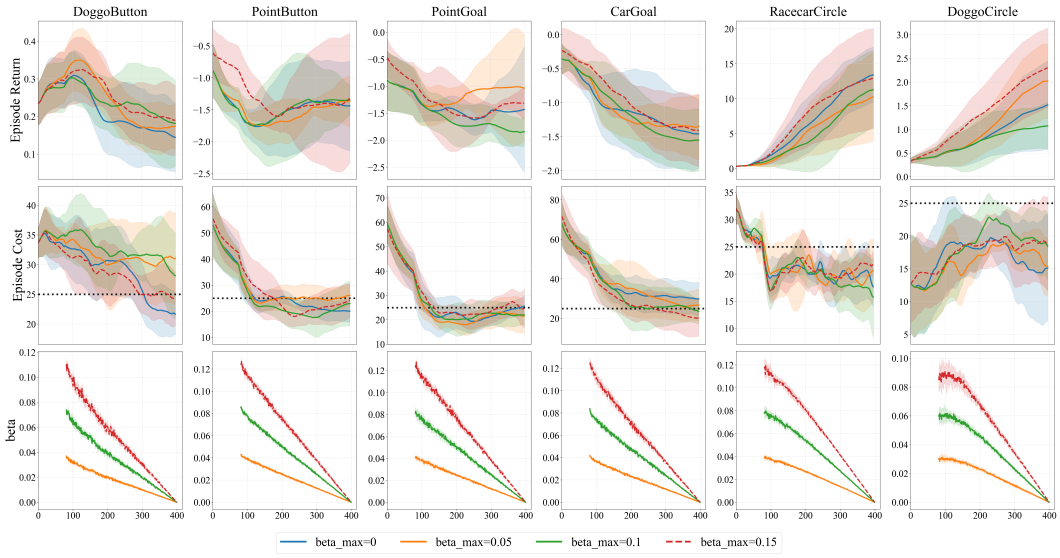


Figure 7: Learning curves for the $\beta_{\max}$ ablation.

I.3.4 N_{enc}

We ablate the encoder warmup horizon N_{enc} to characterize how representation quality affects the retrieval-based correction used by SPEC. During warmup, the encoder is trained before retrieval-based updates are introduced; increasing N_{enc} therefore gives the representation more time to organize the state space, but correspondingly leaves fewer of the fixed 2M training steps for SPEC optimization. As shown in Table 7 and Figure 8, the 100k setting remains inconsistent. The 300k setting provides the strongest overall consistency across the benchmark and is the only tested setting whose IQM cost satisfies the cost limit on every reported task, whereas extending warmup to 600k does not improve aggregate reward–cost performance and leads to constraint violations on the Button and Goal tasks.

To quantify the structure of the learned retrieval space, we additionally report the entropy-aware scaling component of β from Eq. equation 6, which we denote

$$\beta_{\text{scale}} := \frac{1}{1 + \exp(-10\,(g_{\text{ent}} - 0.5))}, \qquad \beta = \beta_{\max}\beta_{\text{scale}}.$$

As shown in Table 8, a larger β_{scale} corresponds to lower-entropy, more concentrated retrieval neighborhoods and therefore permits a stronger episodic correction. However, the improvement generally diminishes between 300k and 600k, while the corresponding reward–cost performance does not improve. These results suggest that beyond a moderate warmup horizon, additional representation refinement provides limited benefit relative to the reduction in training steps available for retrieval-based optimization; $N_{\text{enc}} = 300\text{k}$ therefore provides a practical balance between representation learning and subsequent SPEC optimization.

Table 7: N_{enc} ablation; shading and boldface follow Table 1.

Env	**Metric**	$N_{\text{enc}} = 100\text{k}$	$N_{\text{enc}} = 300\text{k}$	$N_{\text{enc}} = 600\text{k}$
DoggoButton	Reward	0.1 ± 0.1	$\mathbf{0.2 \pm 0.1}$	0.1 ± 0.1
	Cost	25.7 ± 7.6	$\mathbf{24.1 \pm 3.9}$	26.4 ± 6.2
PointButton	Reward	-2.3 ± 2.5	$\mathbf{-1.3 \pm 0.8}$	-1.8 ± 0.2
	Cost	24.9 ± 8.0	$\mathbf{18.1 \pm 9.5}$	20.0 ± 8.3
PointGoal	Reward	-1.5 ± 0.5	$\mathbf{-1.2 \pm 0.4}$	-1.2 ± 0.7
	Cost	23.4 ± 10.3	$\mathbf{24.4 \pm 6.6}$	27.9 ± 9.0
CarGoal	Reward	-0.8 ± 0.6	$\mathbf{-1.5 \pm 0.4}$	-1.3 ± 0.9
	Cost	32.4 ± 9.4	$\mathbf{24.9 \pm 16.9}$	27.9 ± 9.9
RacecarCircle	Reward	5.0 ± 8.1	$\mathbf{14.8 \pm 7.7}$	13.7 ± 2.6
	Cost	17.9 ± 10.4	$\mathbf{23.8 \pm 13.3}$	15.4 ± 16.8
DoggoCircle	Reward	1.3 ± 0.9	2.5 ± 0.8	$\mathbf{2.7 \pm 0.8}$
	Cost	24.5 ± 6.9	18.3 ± 9.6	$\mathbf{22.0 \pm 6.9}$

Table 8: Entropy-aware scaling factor β_{scale} across encoder warmup settings. Values report mean $\pm$ standard deviation.

Env	$N_{\text{enc}} = 100\text{k}$	$N_{\text{enc}} = 300\text{k}$	$N_{\text{enc}} = 600\text{k}$
DoggoButton	0.56 ± 0.05	0.65 ± 0.04	0.67 ± 0.04
PointButton	0.79 ± 0.04	0.81 ± 0.03	0.82 ± 0.03
PointGoal	0.79 ± 0.04	0.82 ± 0.03	0.82 ± 0.03
CarGoal	0.73 ± 0.04	0.78 ± 0.04	0.77 ± 0.03
RacecarCircle	0.83 ± 0.04	0.87 ± 0.04	0.90 ± 0.03
DoggoCircle	0.60 ± 0.03	0.71 ± 0.05	0.77 ± 0.06

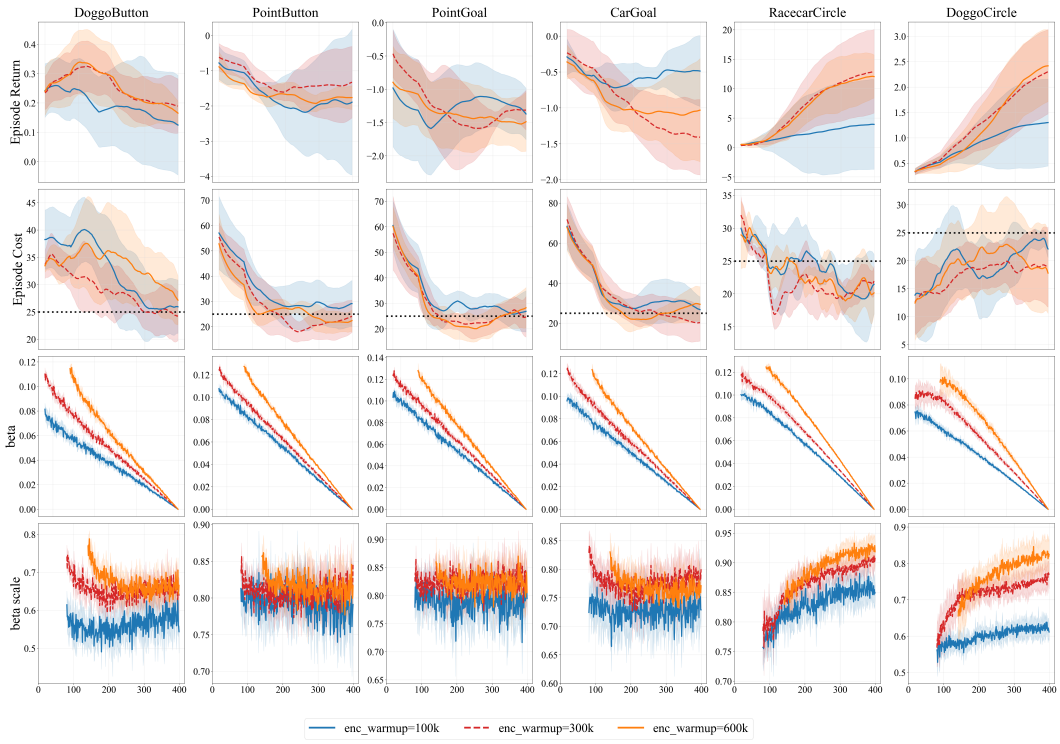


Figure 8: Learning curves for the N_{enc} ablation.