FoSeRL: Formal Sequential Robustness Certification for Reinforcement Learning Policies

Anonymous authors
Paper under double-blind review

Abstract

Even a few action perturbations can substantially degrade the performance of a deployed decision policy. Certifying the resulting return loss is challenging in stochastic environments, where returns vary even without an attack. We introduce FoSeRL, a framework for certifying deployed RL policies against precommitted, temporally sparse action attacks. The deployed policy is unchanged, with no smoothing or retraining. Certification requires a resettable simulator supporting shared randomness and independent one-step successor queries, but no analytical dynamics model. FoSeRL certifies that an attacked episode loses no more than a prescribed amount of return relative to the same episode unattacked, with at least a target probability and at a user-specified confidence level. Both runs share the initial state and randomness, so the measured loss reflects the attack, not the episode; carrying the running return gap as a state coordinate makes it the terminal value, reducing trajectory-level certification to terminal safety. Time-dependent barrier conditions on the augmented state bound the terminal failure probability: satisfied exactly, they certify every admissible precommitted attack; learned from sampled trajectories and verified on held-out data, they certify the same guarantee under a specified attack-episode setting. Across six stochastic continuous-control environments and three RL policy families (TD3, SAC, and PPO), FoSeRL certifies non-trivial cardinality–magnitude robustness frontiers, achieves substantially larger certified budgets than policy smoothing, and reveals marked robustness differences among policies with comparable nominal performance.

1 Introduction

Reinforcement learning (RL) policies have demonstrated control of physical systems, including tokamak plasmas (Degrave et al., 2022), stratospheric balloons (Bellemare et al., 2020), and racing drones (Kaufmann et al., 2023). Once such a policy is deployed, its decisions do not reach the plant directly: each command travels through communication links and actuation software before an actuator executes it. Every stage of that path is a place where an adversary with modest resources can intervene, as end-to-end attacks on production automotive control systems have shown (Koscher et al., 2010; Checkoway et al., 2011). The executed action can therefore differ from the command computed by the policy, even though the deployed policy itself remains unchanged. Thus, we study deployment-time perturbations of executed actions, with the decision policy held fixed.

The effect of such perturbations depends on their magnitude and, just as much, on their timing. On five Atari games, Lin et al. (2017) found that perturbing roughly a quarter of the time steps reduced return about as much as perturbing every step. The reason is that the closed loop is a dynamical system: a perturbed action changes the state from which all later decisions are made, so its effect persists after the perturbation itself has ended. An adversary with a limited budget can therefore concentrate it where the loop is most sensitive instead of spreading it uniformly, and the same holds when the perturbation is injected into the executed action rather than the observation (Lee et al., 2020). Two quantities govern what such an adversary can do: the number of perturbed time steps, K, and a bound b on the size of each perturbation. Certifying a deployed policy therefore means controlling their cumulative effect on return. In a stochastic environment, the nominal return already varies from episode to episode, so the baseline against which loss is measured becomes part of the certification problem. Comparing an attacked run to the policy's average return would charge that variation to the adversary. We therefore compare each attacked execution with the run the same episode would have produced unattacked, using the same initial state and the same environment randomness, and measure the loss between the two. The attack is then the only difference between

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.

them, while the randomness that shapes how an attack unfolds is preserved rather than averaged away. Our objective is to identify the attack budgets under which this loss stays within a prescribed tolerance with a target probability.

Contributions. We introduce FoSeRL, a post-hoc framework for certifying fixed deployed policies against temporally sparse action perturbations. Our contributions are threefold. ***(i)*** We formulate sequential robustness under joint cardinality–magnitude budgets and define a paired return-degradation criterion using synchronously coupled nominal and attacked executions, isolating attack-induced loss from episode variability. ***(ii)*** We lift the cumulative return gap to a terminal quantity of an augmented stochastic system and derive time-dependent barrier conditions for its tail reliability (Fig. 1). ***(iii)*** We learn neural barrier certificates from sampled rollouts and verify them using finite-sample successor bounds and episode-level conformal calibration, yielding high-confidence operational certificates under a declared attack–episode law. FoSeRL leaves the deployed policy unchanged and requires no analytical dynamics model. Across six stochastic continuous-control environments and TD3, SAC, and PPO policies, it yields non-trivial robustness frontiers and substantially larger certified budgets than the baseline.

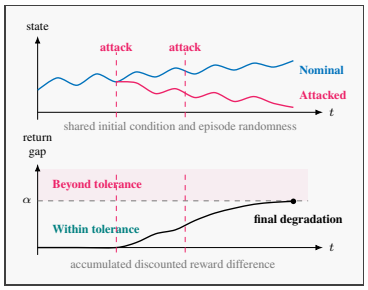


Figure 1: Matched nominal and attacked executions share episode randomness; their accumulated return gap becomes the terminal degradation certified by FoSeRL.

Related work. Temporally sparse attacks are well established empirically (Lin et al., 2017; Sun et al., 2020), including on the action channel (Lee et al., 2020), and defences against them harden a policy by retraining (Pinto et al., 2017; Tessler et al., 2019; Liang et al., 2022) without emitting a certificate. Certified robustness for RL does emit one, but of a different shape. Smoothing-based methods certify actions or cumulative reward under norm-bounded perturbations (Kumar et al., 2022; Wu et al., 2022a; Mu et al., 2024); certified state-action bounds support robust action selection (Everett et al., 2021); and CAROL learns a policy jointly with a model-based robustness certificate (Yang et al., 2024). Each bounds perturbations under a norm ball, typically guarantees an expected return, and certifies a smoothed, retrained, or robust-argmax surrogate rather than the policy as deployed. Stochastic and neural barrier certificates supply the machinery we build on (Prajna et al., 2007; Steinhardt & Tedrake, 2012; Abate et al., 2021; Anand & Zamani, 2023; Luo & Ma, 2021), but bound state-safety events rather than adversarial performance loss; the framework closest to ours certifies a supervised learner against data poisoning by modeling training as a stochastic dynamical system (Taheri & Zamani, 2026), and has no time axis for an attacker to exploit. Attacks in the literature are budgeted along two axes: how often the adversary strikes and how hard, but existing certificates fix one or the other: they either bound perturbations at every step or count corruptions without bounding their size. The cardinality–magnitude trade-off therefore appears throughout the *attack* literature and in no prior *certificate*. FoSeRL is the first to certify it, and the first to do so for the policy exactly as deployed, against perturbations of its executed actions, as a tail bound on *paired* degradation. Table 4 in Appendix F places each method on these axes.

2 Problem Formulation

All proofs and a summary of notation are provided in the appendix. We consider a finite-horizon Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0, T)$, where $\mathcal{S} \subseteq \mathbb{R}^{n_s}$ and $\mathcal{A} \subseteq \mathbb{R}^{n_a}$ are the state and action spaces, $\mathcal{A}$ is closed and convex, $P(\cdot \mid x, u)$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is measurable, $\gamma \in (0, 1]$, and $x(0) \sim \mu_0$. We use $T \in \mathbb{N}$, $\mathbb{T} := \{0, \dots, T\}$, and $\mathbb{T}' := \{0, \dots, T-1\}$. Fix a deployed decision policy π, held unchanged throughout certification, and let $u(t) \in \mathcal{A}$ denote its realized action. Its execution is specified in Assumption 2; no restriction on internal policy randomization is imposed. The state contains all variables required for the closed-loop Markov property, and conditional on the current state and executed action, the successor has law $P(\cdot \mid x(t), u(t))$. For a nominal state–action trajectory ξ, define $J(\pi; \xi) := \sum_{t \in \mathbb{T}'} \gamma^t r(x(t), u(t))$ and $J(\pi) := \mathbb{E}[J(\pi; \xi)]$. All certification statements condition on the deployed policy; randomness used to train it is not resampled during execution.

Assumption 1 (Bounded rewards)**.** The reward satisfies $|r(x, u)| \le r_{\max}$ for all $(x, u) \in \mathcal{S} \times \mathcal{A}$. Hence every realized discounted return lies in $[-S_{\max}, S_{\max}]$, where $S_{\max} := r_{\max} \sum_{t \in \mathbb{T}'} \gamma^t$.

We next formalize the deployment-time adversary acting on the policy's executed actions.

Definition 1 (Oblivious sequential attack)**.** Fix $K \in \mathbb{T}$, $b \geq 0$, and $p \in \{1, 2, \infty\}$. An attack scenario is $\upsilon = (\mathcal{T}_{\mathrm{atk}}, \Delta)$, where $\mathcal{T}_{\mathrm{atk}} \subseteq \mathbb{T}'$ contains the attacked time indices and $\Delta = \{\Delta_t\}_{t \in \mathcal{T}_{\mathrm{atk}}}$ the corresponding perturbations $\Delta_t \in \mathbb{R}^{n_a}$. Both are fixed before the initial state and episode randomness are realized. We set $\Delta_t = 0$ for $t \notin \mathcal{T}_{\mathrm{atk}}$. The admissible class $\mathcal{A}_{\mathrm{adv}}(K, b)$ consists of all such scenarios with $|\mathcal{T}_{\mathrm{atk}}| \leq K$ and $\|\Delta_t\|_p \leq b$ for every $t \in \mathcal{T}_{\mathrm{atk}}$.

To compare an attack with the nominal execution of the *same* episode, we make all fresh closed-loop randomness explicit.

Assumption 2 (Execution randomness representation)**.** There exist measurable execution maps $\pi : \mathcal{S} \times \mathcal{W} \to \mathcal{A}$ and $F : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathcal{S}$, and an i.i.d. sequence $w(t) \sim P_w$, independent of $x(0)$. Each $w(t)$ collects all fresh one-step policy and transition randomness under the declared joint law. The fixed policy generates $u(t) = \pi(x(t), w(t))$, and, given the current state and executed action $\bar{u}(t)$, $F(x(t), \bar{u}(t), w(t))$ has law $P(\cdot \mid x(t), \bar{u}(t))$. A policy without internal randomization ignores the corresponding inputs; persistent cross-step dependence is represented in the state.

The next definition uses this representation to construct matched nominal and attacked executions.

Definition 2 (Paired degradation via synchronous coupling)**.** Fix $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$ and let $\omega := (x(0), \{w(t)\}_{t \in \mathbb{T}'}) \sim \mathbb{P}_\omega$ denote the episode randomness. Both branches share $x(0)$ and the complete random-input sequence. With $\Pi_{\mathcal{A}}$ denoting Euclidean projection onto $\mathcal{A}$, set $u^0(t) := \pi(x^0(t), w(t))$ and $\bar{u}^a(t; \upsilon) := \Pi_{\mathcal{A}}\big(\pi(x^a(t), w(t)) + \Delta_t\big)$. At each step, transition and reward use the same realized action. The magnitude budget applies to Δ_t before projection; admissibility details are given in Appendix B.1. For $t \in \mathbb{T}'$, $x^0(t{+}1) = F(x^0(t), u^0(t), w(t))$ and $x^a(t{+}1) = F(x^a(t), \bar{u}^a(t; \upsilon), w(t))$, with $x^0(0) = x^a(0) = x(0)$. Let $R_{\mathrm{nom}}(\omega) := \sum_{t \in \mathbb{T}'} \gamma^t r(x^0(t), u^0(t))$ and $R_{\mathrm{att}}(\upsilon, \omega) := \sum_{t \in \mathbb{T}'} \gamma^t r(x^a(t), \bar{u}^a(t; \upsilon))$. The paired trajectory-level degradation is

$$D_{\mathrm{traj}}(\upsilon, \omega) := R_{\mathrm{nom}}(\omega) - R_{\mathrm{att}}(\upsilon, \omega). \tag{1}$$

The coupling is part of the specification: although $\mathbb{E}_\omega[D_{\mathrm{traj}}(\upsilon, \omega)] = J(\pi) - \mathbb{E}_\omega[R_{\mathrm{att}}(\upsilon, \omega)]$ depends only on the trajectory marginals, the law of D_{traj} generally does not. Under the synchronous coupling, $K = 0$ or $b = 0$ gives $D_{\mathrm{traj}} \equiv 0$ pathwise. The representation (π, F, P_w) and its shared-randomness semantics are fixed before synthesis and verification; Appendix B.2 gives details showing that the marginals alone are insufficient. We can now state the desired robustness criterion.

Definition 3 (Certified performance)**.** Fix an attack budget (K, b), $\alpha \geq 0$, and $\rho \in (0, 1]$. The fixed policy π is (α, ρ)-*certified performant* under $\mathcal{A}_{\mathrm{adv}}(K, b)$ if its paired degradation does not exceed α with probability at least ρ for every admissible attack:

$$\mathbb{P}_\omega(D_{\mathrm{traj}}(\upsilon, \omega) \leq \alpha) \geq \rho, \qquad \forall\, \upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b). \tag{2}$$

The probability is over ω, conditional on the fixed policy and execution representation; attacks are quantified universally, with no attack distribution.

We finally quantify the certified perturbation tolerance at a fixed cardinality budget.

Problem 1 (Sequential robustness certification)**.** Given the fixed policy π, prescribed levels $\alpha \geq 0$ and $\rho \in (0, 1]$, and a cardinality budget K, determine the supremal attack magnitude for which π remains (α, ρ)-certified performant under $\mathcal{A}_{\mathrm{adv}}(K, b)$. We denote this value by $b^*_{\mathrm{cert}}(K) \in [0, +\infty]$. As K varies, $K \mapsto b^*_{\mathrm{cert}}(K)$ defines the *cardinality–magnitude robustness frontier*.

3 Methodology

The criterion in Definition 3 is trajectory-level and not directly a state-space safety property. We therefore augment the synchronously coupled nominal and attacked dynamics with their cumulative discounted return gap, making the certified degradation a terminal-state quantity. This recasts sequential robustness certification as a finite-horizon terminal safety problem amenable to stochastic barrier certificates.

Definition 4 (Augmented dt-SDS)**.** Fix an attack budget (K, b) and the corresponding class $\mathcal{A}_{\mathrm{adv}}(K, b)$ of Definition 1. The augmented discrete-time stochastic dynamical system is $\mathfrak{S} := (\mathcal{X}, \mathcal{X}_0, \mathcal{A}_{\mathrm{adv}}(K, b), \mathcal{W}, P_w, f, T)$, where $\mathcal{X} := \mathcal{S} \times \mathcal{S} \times [-2S_{\max}, 2S_{\max}]$, $\mathcal{X}_0 := \{(x, x, 0) : x \in \mathrm{supp}(\mu_0)\}$, and $\hat{x}(t) := (x^0(t), x^a(t), z(t))$, initialized at $\hat{x}(0) = (x(0), x(0), 0)$ with $x(0) \sim \mu_0$. For each fixed $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$, let $u^0(t) := \pi(x^0(t), w(t))$, let $\bar{u}^a(t; \upsilon)$ be as in Definition 2, and

let $\mathcal{X}_t^{\upsilon} \subseteq \mathcal{X}$ denote the support of the augmented-state distribution at time t under υ. Whenever $f(\hat{x}, \upsilon, w, t)$ is evaluated, the actions below are recomputed from the supplied w, consistently with Definition 2. The augmented dynamics are

$$\hat{x}(t{+}1) = f(\hat{x}(t), \upsilon, w(t), t) := \begin{bmatrix} F\big(x^0(t), u^0(t), w(t)\big) \\ F(x^a(t), \bar{u}^a(t;\upsilon), w(t)) \\ z(t) + \gamma^t \big[r(x^0(t), u^0(t)) - r(x^a(t), \bar{u}^a(t;\upsilon))\big] \end{bmatrix}, \; t \in \mathbb{T}'. \tag{3}$$

Each evaluation of $f(\hat{x}, \upsilon, w, t)$ recomputes both actions and the return-gap increment using its supplied w; successor draws therefore reuse neither cached actions nor cached gap increments from the source episode. Hence $z(t{+}1)$ may remain random conditional on $(\hat{x}(t), \upsilon, t)$. Since each admissible attack is precommitted, υ parameterizes the augmented dynamics and need not be included in the state. The identity $z(T) = D_{\mathrm{traj}}(\upsilon, \omega)$ yields the following terminal safety specification.

Definition 5 (Terminal safety specification)**.** For the augmented dt-SDS $\mathfrak{S}$ of Definition 4 and $\alpha \geq 0$, define $\mathcal{X}_s^{\alpha} := \{\hat{x} \in \mathcal{X} : z \leq \alpha\}$ and $\mathcal{X}_u^{\alpha} := \mathcal{X} \setminus \mathcal{X}_s^{\alpha} = \{\hat{x} \in \mathcal{X} : z > \alpha\}$. For $\rho \in (0, 1]$, let $\Psi_{\alpha,\rho} := (\mathcal{X}_0, \mathcal{X}_s^{\alpha}, \mathcal{X}_u^{\alpha}, T, \rho)$. We say $\mathfrak{S}$ satisfies $\Psi_{\alpha,\rho}$ if $\mathbb{P}_{\omega}(\hat{x}(T) \in \mathcal{X}_s^{\alpha}) \geq \rho$ for every $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$.

This specification is equivalent to Definition 3 and imposes no intermediate safety constraint. We next derive sufficient time-dependent barrier conditions, uniformly over $\mathcal{A}_{\mathrm{adv}}(K, b)$.

Definition 6 (Barrier certificate)**.** Consider the augmented dt-SDS $\mathfrak{S}$ in Definition 4 with attack class $\mathcal{A}_{\mathrm{adv}}(K, b)$ and terminal specification $\Psi_{\alpha,\rho}$ of Definition 5. A measurable function $B : \mathcal{X} \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ is a *barrier certificate* for $\mathfrak{S}$ with respect to $\Psi_{\alpha,\rho}$ if there exist $b_1, b_2, b_3 \in \mathbb{R}_{>0}$ and $\kappa \in (0, 1]$ such that the following conditions hold for every $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$, where $\mathcal{G}_T(\kappa) := \sum_{j=0}^{T-1} \kappa^j$:

$$\begin{aligned} & B(\hat{x}, 0) \leq b_1, && \forall\, \hat{x} \in \mathcal{X}_0, \\ & B(\hat{x}, T) \geq b_2, && \forall\, \hat{x} \in \mathcal{X}_u^{\alpha}, \\ & \mathbb{E}_{w \sim P_w}[B(f(\hat{x}, \upsilon, w, t), t{+}1)] \leq \kappa B(\hat{x}, t) + b_3, && \forall\, \hat{x} \in \mathcal{X}_t^{\upsilon},\ \forall\, t \in \mathbb{T}', \\ & b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa) \leq b_2(1 - \rho). \end{aligned} \tag{4}$$

The four conditions bound the initial value, separate terminal-unsafe states, control the expected one-step drift along states reachable under each admissible attack, and calibrate the resulting bound to ρ. Together they bound terminal failure uniformly over $\mathcal{A}_{\mathrm{adv}}$. Because the same attack-independent barrier must accommodate all precommitted attack plans, these sufficient conditions can be conservative at states shared across different plans. Their conservatism is discussed in Appendix B.4.

Theorem 1 (Ideal sequential robustness guarantee)**.** *Fix $K \in \mathbb{T}$, $b, \alpha \geq 0$, and $\rho \in (0, 1)$. Under Assumptions 1 and 2, consider the augmented dt-SDS $\mathfrak{S}$ of Definition 4 with terminal safety specification $\Psi_{\alpha,\rho}$ of Definition 5. Suppose B is a barrier certificate in the sense of Definition 6, with parameters b_1, b_2, b_3, and κ. Then $\mathfrak{S}$ satisfies $\Psi_{\alpha,\rho}$; equivalently, the fixed policy π is (α, ρ)-certified performant under $\mathcal{A}_{\mathrm{adv}}(K, b)$ in the sense of Definition 3. Consequently, b belongs to the feasible set defining $b^*_{\mathrm{cert}}(K)$ in Problem 1. More precisely, for every $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$,*

$$\mathbb{P}_{\omega}(D_{\mathrm{traj}}(\upsilon, \omega) \leq \alpha) \geq 1 - \frac{b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa)}{b_2} \geq \rho. \tag{5}$$

Theorem 1 requires the barrier inequalities to hold uniformly over the specified reachable states and admissible attacks. The finite-sample procedure below instead certifies operational reliability under a declared attack law and does not establish feasibility for the universal objective in Problem 1.

4 Data-Driven Sequential Robustness Certification

We assume access to synchronously coupled rollouts and repeatable one-step queries from reachable augmented state–attack–time tuples. Each query restores both branches before action generation and draws a fresh complete $w \sim P_w$, shared within the pair and independent of the source episode. Thus, policy and transition randomness are regenerated, and the reward-gap coordinate is updated from the same realized actions. No analytical dynamics model is required. We next specify the attack–episode laws used for candidate construction and formal verification.

Definition 7 (Attack–episode laws)**.** For an attack budget (K, b), let $\mathcal{U}_{K,b}$ be a prescribed probability law supported on $\mathcal{A}_{\mathrm{adv}}(K, b)$. We model the sampled attack and episode randomness as independent, consistently with the oblivious threat model, and define the operational law $\mathbb{Q}_{K,b} := \mathcal{U}_{K,b} \otimes \mathbb{P}_{\omega}$. For synthesis at a design magnitude $b_{\mathrm{cnd}} \in \mathbb{R}_{\geq 0}$, we allow a separate law $\mathcal{U}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}}$ supported on $\mathcal{A}_{\mathrm{adv}}(K, b_{\mathrm{cnd}})$ and define $\mathbb{Q}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}} := \mathcal{U}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}} \otimes \mathbb{P}_{\omega}$.

The synthesis law is used only to construct the candidate barrier; the formal data-driven guarantee is stated under $\mathbb{Q}_{K,b}$. For the fixed policy, tolerance, horizon, and coupling, define $R_{\mathrm{op}}(K,b;\mathcal{U}_{K,b}) := \mathbb{P}_{(v,\omega)\sim\mathbb{Q}_{K,b}}(D_{\mathrm{traj}}(v,\omega)\le\alpha)$. The operational target is $R_{\mathrm{op}} \ge \rho$ for a tested specification, rather than the universal per-attack requirement of Problem 1. We reserve $b^{\mathrm{op}}_{\mathrm{cert}}$ for an accepted tested magnitude under this operational criterion, distinct from the universal radius $b^{*}_{\mathrm{cert}}(K)$. Henceforth, we consider $K>0$, $b>0$, $0\le\alpha<2S_{\max}$, and $\rho\in(0,1)$; other cases are summarized in Appendix B.5.

4.1 Certificate Synthesis

We synthesize a candidate for the conditions of Definition 6 by jointly learning a neural barrier and its associated constants from coupled trajectories. The synthesis specification consists of a cardinality budget $K \in \mathbb{T}$, a design magnitude $b_{\mathrm{cnd}} \in \mathbb{R}_{\ge 0}$, a degradation tolerance $\alpha \in \mathbb{R}_{\ge 0}$, and a target reliability $\rho \in (0,1)$. We choose a synthesis-only reserve $\eta_{\mathrm{syn}} \in (0, 1-\rho)$ and set $\rho_{\mathrm{syn}} := \rho + \eta_{\mathrm{syn}}$. This reserve provides optimization headroom only: it has no formal role, and the certification target remains ρ. Let h_ϕ be a scalar neural network with parameters ϕ. For a prescribed $B_{\max} \in \mathbb{R}_{>0}$, we parameterize $B_\phi(\hat{x},t) := B_{\max}\,\mathrm{sigmoid}(h_\phi(\hat{x},t))$, which ensures $B_\phi(\hat{x},t) \in (0, B_{\max})$ for every $(\hat{x},t)\in\mathcal{X}\times\mathbb{T}$.

Synthesis data. We draw N_s independent scenarios $(v_i,\omega_i) \overset{\text{i.i.d.}}{\sim} \mathbb{Q}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}}$, $i\in[N_s]$, and generate the corresponding trajectories $\hat{x}_i(0{:}T)$ according to Definition 4. Let $d_i := D_{\mathrm{traj}}(v_i,\omega_i)$, $\mathcal{I}_s := \{i\in[N_s] \mid d_i \le \alpha\}$, and $\mathcal{I}_u := \{i\in[N_s] \mid d_i > \alpha\}$. We organize their initial states, state–attack–time tuples, and terminal safe–unsafe partition as indexed collections that retain repeated observations:

$$I := \big(\hat{x}_i(0)\big)_{i\in[N_s]}, \quad X := \big(\hat{x}_i(t), v_i, t\big)_{i\in[N_s],t\in\mathbb{T}'}, \quad S := \big(\hat{x}_i(T)\big)_{i\in\mathcal{I}_s}, \quad U := \big(\hat{x}_i(T)\big)_{i\in\mathcal{I}_u}. \tag{6}$$

Since $z_i(T) = D_{\mathrm{traj}}(v_i,\omega_i)$, every element of I lies in $\mathcal{X}_0$, while the elements of S and U lie in $\mathcal{X}_s^\alpha$ and $\mathcal{X}_u^\alpha$, respectively. Thus, S and U collect sampled terminal safe and unsafe states; U enters the synthesis loss, while S is retained for supplementary analysis. Additional synthesis-pool conventions are given in Appendix D.4. To approximate the expectation in the third condition of Definition 6, for each indexed tuple $(\hat{x}_i(t), v_i, t)$, $i\in[N_s]$, $t\in\mathbb{T}'$, we draw M_s fresh conditionally i.i.d. random inputs $w_{itj}\sim P_w$, $j\in[M_s]$, independently of the source episode. For each tuple, define $\bar{E}_{it} := M_s^{-1}\sum_{j=1}^{M_s} B_\phi(f(\hat{x}_i(t),v_i,w_{itj},t),t{+}1)$ and replace the exact expectation during synthesis by $\bar{E}_{it}+\epsilon$, where $\epsilon\in\mathbb{R}_{\ge 0}$ is a prescribed synthesis margin. This margin tightens the sampled drift condition during synthesis only and carries no formal confidence guarantee.

Synthesis objective. We combine the sampled barrier conditions in a weighted training loss. For prescribed weights $c_\rho, c_I, c_U, c_f \in \mathbb{R}_{>0}$, we jointly optimize ϕ, $b_1, b_2, b_3 \in \mathbb{R}_{>0}$, and $\kappa\in(0,1]$, subject to $b_2 < B_{\max}$, by minimizing $\mathcal{L}/b_2$, where

$$\begin{aligned}\mathcal{L} = {} & c_\rho \mathrm{ReLU}\big(b_1\kappa^T + b_3\mathcal{G}_T(\kappa) - b_2(1-\rho_{\mathrm{syn}})\big) \;+\; c_I \sum_{i\in[N_s]} \mathrm{ReLU}(B_\phi(\hat{x}_i(0),0) - b_1) \\ & + c_U \sum_{i\in\mathcal{I}_u} \mathrm{ReLU}(b_2 - B_\phi(\hat{x}_i(T),T)) \;+\; c_f \sum_{i\in[N_s]}\sum_{t\in\mathbb{T}'} \mathrm{ReLU}\big(\bar{E}_{it}+\epsilon-\kappa B_\phi(\hat{x}_i(t),t)-b_3\big).\end{aligned} \tag{7}$$

The four terms penalize violations of the reliability, initial, unsafe-terminal, and sampled drift conditions, respectively. After training, the candidate B_{ϕ^*} and constants $b_1^*, b_2^*, b_3^*, \kappa^*$ are fixed before independent verification. Synthesis constructs a candidate only; its formal validity is established by the verification stage.

4.2 Certificate Verification

Verification converts the frozen synthesis output into a finite-sample robustness guarantee under the operational law $\mathbb{Q}_{K,b}$ of Definition 7. Fix a verification specification $K\in\mathbb{T}$, $b,\alpha\in\mathbb{R}_{\ge 0}$, and $\rho\in(0,1)$, and write $B^*_\phi := B_{\phi^*}$ with $b_1^*, b_2^*, b_3^*, \kappa^*$ from Sec. 4.1. The candidate and all verification-design choices are fixed before calibration; fresh calibration data are drawn independently of data used for candidate construction or selection. All probability statements below condition on these pre-calibration choices. We first express the barrier conditions as signed residuals.

Definition 8 (Barrier residuals)**.** For the frozen candidate and verification specification, define $\mathcal{Z}_1 := \mathcal{X}_0\times\mathcal{A}_{\mathrm{adv}}(K,b)\times\{0\}$, $\mathcal{Z}_2 := \mathcal{X}_u^\alpha\times\mathcal{A}_{\mathrm{adv}}(K,b)\times\{T\}$, and $\mathcal{Z}_3 := \{(\hat{x},v,t) : v\in\mathcal{A}_{\mathrm{adv}}(K,b),\ t\in\mathbb{T}',\ \hat{x}\in\mathcal{X}_t^v\}$, where $\mathcal{X}_t^v$ is defined in Definition 4. For $j\in\{1,2,3\}$, set $q_j = 0$ outside $\mathcal{Z}_j$. Positive residuals indicate violations of the corresponding conditions in Definition 6:

$$q_1(\hat{x},\upsilon,t) := B^*_\phi(\hat{x},0) - b^*_1, \qquad \forall\,(\hat{x},\upsilon,t)\in\mathcal{Z}_1, \tag{8}$$
$$q_2(\hat{x},\upsilon,t) := b^*_2 - B^*_\phi(\hat{x},T), \qquad \forall\,(\hat{x},\upsilon,t)\in\mathcal{Z}_2, \tag{9}$$
$$q_3(\hat{x},\upsilon,t) := \mathbb{E}_{w\sim P_w}\big[B^*_\phi(f(\hat{x},\upsilon,w,t),t{+}1)\big] - \kappa^* B^*_\phi(\hat{x},t) - b^*_3, \qquad \forall\,(\hat{x},\upsilon,t)\in\mathcal{Z}_3, \tag{10}$$
$$q_4 := b^*_1(\kappa^*)^T + b^*_3\mathcal{G}_T(\kappa^*) - b^*_2(1-\rho). \tag{11}$$

The initial and terminal residuals are directly evaluable, so $\bar{q}_1 := q_1$ and $\bar{q}_2 := q_2$. The drift residual q_3 contains an expectation under the stochastic dynamics. Fix $M\in\mathbb{N}$ and $\epsilon_M\in\mathbb{R}_{>0}$ before calibration. At each queried $(\hat{x},\upsilon,t)\in\mathcal{Z}_3$, draw $w_1,\dots,w_M \overset{\text{i.i.d.}}{\sim} P_w$ using fresh random inputs independent of the complete episode that supplies the tuple. Define

$$\overline{E}^*_M(\hat{x},\upsilon,t) := \frac{1}{M}\sum_{m=1}^{M} B^*_\phi(f(\hat{x},\upsilon,w_m,t),t{+}1) + \epsilon_M, \tag{12}$$
$$\bar{q}_3(\hat{x},\upsilon,t) := \overline{E}^*_M(\hat{x},\upsilon,t) - \kappa^* B^*_\phi(\hat{x},t) - b^*_3, \qquad (\hat{x},\upsilon,t)\in\mathcal{Z}_3, \tag{13}$$

with $\bar{q}_3 := 0$ outside $\mathcal{Z}_3$. Unlike the synthesis margin, ϵ_M has a formal finite-sample role through the following bound.

Lemma 1 (Supermartingale residual bound). *Fix $\beta_0\in(0,1)$, $\epsilon_M\in\mathbb{R}_{>0}$, and $(\hat{x},\upsilon,t)\in\mathcal{Z}_3$. Suppose $B^*_\phi(\hat{x}',t')\in(0,B_{\max})$ for every $(\hat{x}',t')\in\mathcal{X}\times\mathbb{T}$, and the successor random inputs used in equation 12 are conditionally i.i.d. from P_w at the fixed tuple. Then, with probability at least $1-\beta_0$ over the successor batch, $q_3(\hat{x},\upsilon,t)\le\bar{q}_3(\hat{x},\upsilon,t)$, provided*

$$M \ge \left\lceil \frac{B^2_{\max}}{2\epsilon^2_M}\ln\frac{1}{\beta_0}\right\rceil. \tag{14}$$

Conformal calibration. We calibrate violations at the episode level, so no independence between time steps within one episode is required.

Definition 9 (Episode-level verification score). Fix $N\in\mathbb{N}$ and let $\mathcal{D}_{\text{cal}} = (\bar{\zeta}_i)_{i=1}^N$ contain independent copies of $\bar{\zeta} = (\upsilon,\omega,W)$, where $(\upsilon,\omega)\sim\mathbb{Q}_{K,b}$ generates the augmented trajectory of Definition 4. For each episode, W contains the fresh successor batches used in equation 12 at all $t\in\mathbb{T}'$, drawn independently of the complete scenario (υ,ω). Write $\bar{q}_{1,i} := \bar{q}_1(\hat{x}_i(0),\upsilon_i,0)$ and $\bar{q}_{2,i} := \bar{q}_2(\hat{x}_i(T),\upsilon_i,T)$. The episode score combines the initial and drift residuals while retaining terminal separation as a hard condition:

$$S_i := \begin{cases} \max\big\{\bar{q}_{1,i},\ \max\limits_{t\in\mathbb{T}'}\bar{q}_3(\hat{x}_i(t),\upsilon_i,t)\big\}, & \bar{q}_{2,i}\le 0,\\ +\infty, & \bar{q}_{2,i}>0. \end{cases} \tag{15}$$

Thus an unsafe-terminal separation violation receives an infinite score and cannot be repaired by a finite calibration margin; all scores, including infinite ones, are retained.

Lemma 2 (Conformal calibration). *Fix $\beta\in(0,1]$ and let $S_{(1)}\le\cdots\le S_{(N)}$ be the order statistics of the scores in Definition 9. For each $k\in[N]$, let $\bar{v}_k$ be the smallest $v\in[0,1]$ satisfying the binomial-tail condition below. Conditional on the pre-calibration choices, with probability at least $1-\beta$ over $\mathcal{D}_{\text{cal}}$, the score of a fresh independent episode exceeds $S_{(k)}$ with probability at most $\bar{v}_k$, simultaneously for all $k\in[N]$. The statement permits ties and infinite scores:*

$$\sum_{j=k}^{N}\binom{N}{j}(1-v)^j v^{N-j} \le \frac{\beta}{N}. \tag{16}$$

Calibration and acceptance. Let $\mathcal{K}_{\text{fin}} := \{k\in[N] : S_{(k)} < +\infty\}$ and, for each finite rank, set $r_k := \max\{0, S_{(k)}\}$. The margin r_k relaxes the initial and drift bounds to $b^*_1 + r_k$ and $b^*_3 + r_k$, while retaining b^*_2. Lemma 2 bounds the out-of-sample probability of violating these empirical episode conditions by $\bar{v}_k$. Together with Lemma 1, the first-exact-violation argument transfers this to the exact adjusted conditions with penalty $\bar{v}_k/(1-\beta_0)$, without a union bound over time; the good-prefix barrier argument then bounds terminal failure. When $\mathcal{K}_{\text{fin}}\neq\varnothing$, select the rank that maximizes the resulting end-to-end reliability lower bound:

$$k^* \in \arg\max_{k\in\mathcal{K}_{\text{fin}}}\left\{1 - \frac{(b^*_1+r_k)(\kappa^*)^T + (b^*_3+r_k)\mathcal{G}_T(\kappa^*)}{b^*_2} - \frac{\bar{v}_k}{1-\beta_0}\right\}. \tag{17}$$

Set $r^* := r_{k^*}$, $\widetilde{b}^*_1 := b^*_1 + r^*$, and $\widetilde{b}^*_3 := b^*_3 + r^*$. Simultaneous validity over all ranks permits this data-dependent selection without an additional confidence penalty. The final verification residual is

$$\bar{q}_4 := \widetilde{b}^*_1(\kappa^*)^T + \widetilde{b}^*_3\mathcal{G}_T(\kappa^*) - b^*_2\left(1-\rho-\frac{\bar{v}_{k^*}}{1-\beta_0}\right). \tag{18}$$

We accept exactly when $\bar{q}_4\le 0$. If $\mathcal{K}_{\text{fin}}=\varnothing$, we set $\bar{q}_4 := +\infty$ and verification is inconclusive.

Theorem 2 (FoSeRL certificate). *Fix $K \in \mathbb{T}$, $b, \alpha \in \mathbb{R}_{\geq 0}$, $\rho \in (0,1)$, $\beta \in (0,1]$, and $\beta_0 \in (0,1)$. Under Assumptions 1 and 2, consider the augmented system of Definition 4 with operational law $\mathbb{Q}_{K,b}$ from Definition 7. Let $B^*_\phi : \mathcal{X} \times \mathbb{T} \to (0, B_{\max})$ be a measurable frozen candidate with $b^*_1, b^*_2, b^*_3 \in \mathbb{R}_{>0}$, $\kappa^* \in (0,1]$, and $b^*_2 < B_{\max}$. Choose $N, M \in \mathbb{N}$ and $\epsilon_M \in \mathbb{R}_{>0}$ satisfying equation 14, and fix the candidate, constants, operational law, and all verification-design choices before drawing $\mathcal{D}_{\mathrm{cal}}$ as in Definition 9, using fresh i.i.d data. Construct $\bar{q}_4$ by equation 17–equation 18, with $\bar{q}_4 := +\infty$ if no finite rank exists. Then, conditional on all pre-calibration choices, acceptance certifies the operational target with confidence at least $1-\beta$ over calibration:*

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cal}}}\left[\bar{q}_4 \leq 0 \implies \mathbb{P}_{(v,\omega)\sim\mathbb{Q}_{K,b}}(D_{\mathrm{traj}}(v,\omega) \leq \alpha) \geq \rho\right] \geq 1-\beta. \tag{19}$$

Algorithm 1 summarizes the workflow: synthesize and freeze a candidate, calibrate on independent episodes, and accept when $\bar{q}_4 \leq 0$; otherwise, verification is inconclusive and does not establish non-robustness. Confidence is per frozen specification; adaptive and collection-level reporting are addressed in Appendix B.7. The rationale for using a barrier witness rather than direct terminal-event certification is discussed in Sec. 6 and Appendix E.4.

5 Experiments

We instantiate the two-stage procedure in Algorithm 1 on six continuous-control environments: Inverted Pendulum (IP), Adaptive Cruise Control (ACC), Unicycle, Hopper, HalfCheetah, and Ant. The first three are custom control tasks, with IP and ACC following classical settings (Åström & Furuta, 2000; Ames et al., 2014); Hopper, HalfCheetah, and Ant use Gymnasium/MuJoCo (Towers et al., 2024; Todorov et al., 2012). Across these environments, we certify frozen policies trained with TD3 (Fujimoto et al., 2018), SAC (Haarnoja et al., 2018), and PPO (Schulman et al., 2017). Each policy is trained before certification and remains fixed throughout synthesis and verification. Complete environment and policy settings, attack generators and norm budgets, horizons, attack grids, and framework parameters are summarized in Table 2 in Appendix D.5.

Algorithm 1: FoSeRL certification

Input: fixed policy π; design magnitude b_{cnd}; tested specification (K, b, α, ρ); synthesis law $\mathbb{Q}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}}$; operational law $\mathbb{Q}_{K,b}$; sample sizes (N_s, M_s, N, M); confidence parameters (β, β_0).

1. **Certificate synthesis.** Draw N_s coupled episodes from $\mathbb{Q}^{\mathrm{syn}}_{K,b_{\mathrm{cnd}}}$ and the required M_s successor samples. Train B_ϕ and (b_1, b_2, b_3, κ) via equation 7, then freeze the candidate.
2. **Certificate verification.** With the supplied M, before drawing calibration data, fix ϵ_M satisfying equation 14, together with all remaining verification-design choices. Draw N independent calibration episodes under $\mathbb{Q}_{K,b}$ and use fresh successor batches at the queried tuples, with successor randomness independent of the source episodes. Compute the scores S_i and $(r_k, \bar{v}_k)$, select k^* by equation 17, and construct $\bar{q}_4$ using equation 18.

Return.
If no finite rank exists or $\bar{q}_4 > 0$, return *inconclusive*; o.w. certify $\mathbb{P}_{(v,\omega)\sim\mathbb{Q}_{K,b}}(D_{\mathrm{traj}} \leq \alpha) \geq \rho$ with confidence at least $1-\beta$.

Results. We evaluate FoSeRL on all six case studies using Algorithm 1. For each tested (K, b, α) specification, the candidate barrier is synthesized and frozen before independent verification on fresh calibration episodes. All formal results target $\rho = 0.90$ with $\beta = 0.001$, and each reported point carries a separate per-specification confidence guarantee. Figure 2 first provides a high-level view of what an accepted FoSeRL budget means in closed-loop behavior. In the auxiliary $T = 500$ Ant visualization, SAC and PPO are shown under the paired nominal execution, their accepted FoSeRL budgets, and larger uncertified attacks, all at $K/T = 10\%$. In the shown accepted-budget rollouts, $z(T)/\alpha = D_{\mathrm{traj}}/\alpha < 1$, whereas the larger uncertified attacks finish above the terminal tolerance and visibly collapse. This experiment is used only for qualitative visualization; the quantitative Ant frontiers and comparisons reported below use the main $T = 1000$ campaign. We report $\alpha_{\%} := 100\,\alpha/S_{\mathrm{env}}$, where $S_{\mathrm{env}} > 0$ is a fixed environment-level return scale; synthesis and verification use the corresponding raw threshold α. Complete benchmark, policy, attack, and certification settings are reported in Table 2, with additional environment details in Appendix D.3. Across environments and policy families, FoSeRL yields non-trivial cardinality–magnitude robustness frontiers.



Figure 2: Illustrative Ant rollouts for SAC and PPO in a separate $T = 500$ experiment. Each row shows a nominal rollout, an accepted FoSeRL budget, and a larger uncertified attack at $K/T = 10\%$. The lower trace shows the running gap $z(t)/\alpha$; at T, $z(T) = D_{\mathrm{traj}}$, and 1 marks the terminal tolerance. Videos for this and additional case studies are available at this link.

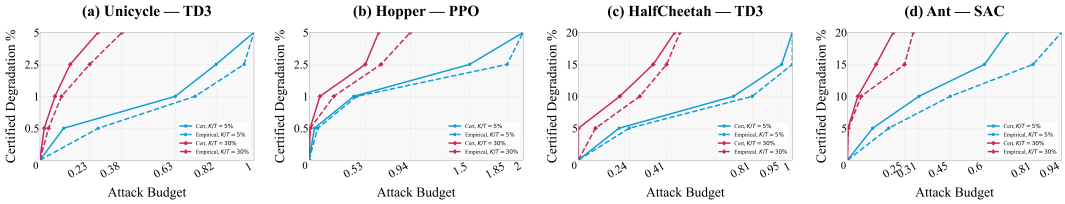


Figure 3: Certified degradation tolerance versus attack budget across different settings. Larger tolerances expand the certified budget, whereas denser temporal attacks generally contract the robustness frontier. Markers denote individually verified specifications and connecting lines are visual guides only. ($\rho = 0.90$, $\beta = 0.001$)

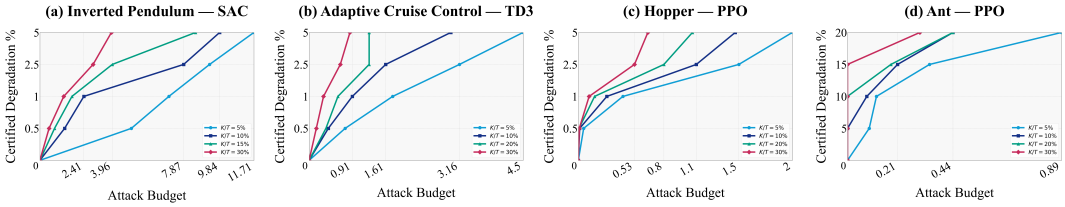


Figure 4: Certified versus empirical attack-budget profiles across degradation tolerances. Solid curves show FoSeRL-certified budgets and dashed curves show the corresponding empirical design references at matched K/T values. The empirical–certified gap varies across environments and policies. ($\rho = 0.90$, $\beta = 0.001$)

As shown in Fig. 3, larger degradation tolerances expand the certified magnitude budget, while denser attacks contract it. For example, at $\alpha_{\%} = 5\%$, the Unicycle–TD3 budget decreases from about 1.0 at $K/T = 5\%$ to about 0.3 at $K/T = 30\%$; for Ant–SAC at $\alpha_{\%} = 20\%$, it decreases from roughly 0.7 to 0.23. Thus, robustness to isolated perturbations does not directly translate to robustness under more frequent attacks. Figure 4 shows that the certified frontiers generally preserve the structure of their empirical design references, but with a setting-dependent gap: some regimes remain close, whereas others lose a substantial fraction of the empirical budget after independent verification. Hence the verification conservatism is not a fixed penalty, but depends on the policy, attack density, and degradation tolerance. Curves connect individually verified specifications and do not certify interpolation. Representative certificates are reported in Table 1, with complete profiles in Appendix D.7.

In addition, Table 1 compares FoSeRL with direct terminal-event certification and policy smoothing (PS) (Kumar et al., 2022; Mu et al., 2024). Direct certification uses exactly the same benchmark attack and norm specification as FoSeRL and can yield larger scalar budgets by bypassing the sufficient barrier conditions, without requiring barrier synthesis or conditional successor queries. Nevertheless, FoSeRL retains roughly 71–99% of the direct certified budget in the representative settings, while additionally returning a state–time barrier witness and residual diagnostics that expose how the certificate is limited along the closed-loop evolution. No existing external formal baseline matches all elements of our setting simultaneously (detailed comparison provided in Appendix F). We therefore adapt PS, the closest formal alternative, to the same action channel, paired degradation event, and reliability and confidence targets, while retaining its native ℓ_2 perturbation geometry. The comparison should therefore be interpreted as an aligned external baseline rather than a fully matched threat-model comparison. Under this alignment, FoSeRL certifies orders-of-magnitude larger nonzero budgets in the comparable settings, while PS is unable to certify a nonzero budget in several cases. Detailed protocols, applicability conditions, and complete results are given in Appendix D.8.

Table 1: Representative formal certificates at selected $(K/T, \alpha_{\%})$ settings. Direct certification, policy smoothing (PS), and FoSeRL are compared by b_{cert}, ρ_{cert}, and runtime (RT, seconds). Direct uses the corresponding benchmark attack and norm specification, whereas PS is evaluated under its native ℓ_2 perturbation setting. All results use $\rho = 0.90$ and $\beta = 0.001$. FoSeRL remains non-trivial where PS fails (Failures are denoted by "–").

Env.	Policy	K/T	$\alpha_{\%}$	Direct			PS			FoSeRL		
				b_{cert}	ρ_{cert}	RT	b_{cert}	ρ_{cert}	RT	b_{cert}	ρ_{cert}	RT
IP	SAC	8%	5%	4	0.92	14.1	0.0132	0.9	11.2	3.00	0.92	10.5
	SAC	13.5%	1%	0.82	0.91	8.6	–	0.9	–	0.65	0.92	12.2
	SAC	20%	5%	2.84	0.91	9.2	0.0083	0.9	11.2	2.50	0.92	15.5
Hopper	TD3	10%	2.5%	1.57	0.91	28.9	0.0014	0.9	357.3	1.32	0.93	96.4
	SAC	10%	2.5%	1.31	0.94	48.5	0.0023	0.9	438.4	1.12	0.94	119.6
	PPO	10%	2.5%	1.29	0.93	31.1	–	0.9	–	1.17	0.93	54.3
HCH	TD3	30%	20%	0.63	0.92	32.1	0.0003	0.9	345.3	0.45	0.92	139.5
	SAC	30%	20%	0.26	0.94	47.7	0.0013	0.9	321.3	0.22	0.94	134.1
	PPO	30%	20%	0.92	0.96	33.4	–	0.9	–	0.91	0.92	145
Ant	TD3	20%	20%	0.63	0.92	62.1	–	–	–	0.49	0.91	94.7
	SAC	20%	20%	0.39	0.9	81.1	–	–	–	0.31	0.91	138.6
	PPO	20%	20%	0.55	0.91	52.9	–	–	–	0.44	0.90	95.6

6 DISCUSSION

FoSeRL provides post-hoc guarantees for the cumulative performance loss of an unchanged policy under temporally sparse action attacks. Its paired-state construction makes return degradation a terminal quantity, connecting sequential robustness to stochastic barrier analysis and independent finite-sample verification. Experiments across six continuous-control environments and three RL policy families demonstrate non-trivial cardinality–magnitude robustness profiles and larger certified budgets than the evaluated smoothing baseline at the reported settings.

Robustness at the episode level. The joint budget (K, b) distinguishes how many actions can be perturbed from the strength of each perturbation, while cumulative degradation captures consequences that persist after the attack ends. Pairing each attacked execution with its nominal counterpart under shared initial conditions and random inputs provides a consistent reference for this loss. It avoids variability introduced by an independently randomized comparison without removing stochastic variation in the attack's consequences, and gives exactly zero degradation when no attack occurs. The resulting tail criterion addresses excessive episode-level loss, rather than only average performance or instantaneous action changes. Coupling is therefore part of the robustness specification itself, not merely a device for improving estimation.

Policy- and environment-agnostic certification. FoSeRL separates certification from how the policy is obtained. Under the stated execution and sampling assumptions, its argument does not depend on a particular RL algorithm, policy architecture, policy differentiability, or analytical family of dynamics. Learned and analytically specified policies, including deployment-time randomization, fit the same formulation. Relative to the robust-training, policy-smoothing, and model-based approaches reviewed in Appendix F, its distinguishing feature is that certification requires neither a robustness-specific policy modification nor a separately learned dynamics surrogate. Instead, coupled rollouts and fresh conditional successors support certification of the *policy as deployed*. This makes the framework applicable when the original training pipeline is unavailable and complementary to robust policy learning: a policy can be assessed regardless of whether robustness was considered during its training.

The value of a dynamical certificate. Direct terminal testing and FoSeRL address the same operational event but provide different certifying objects. A one-sided binomial bound on independent paired episodes can certify the terminal probability without barrier synthesis or successor queries. FoSeRL additionally provides a state–time witness linking initial conditions, expected one-step evolution, and terminal separation. Its residuals and reliability decomposition identify which certificate conditions and sampled episode phases limit witness quality, supporting targeted refinement. These are diagnostics, not certified conditional failure probabilities at arbitrary states. Direct testing can yield tighter budgets at lower cost by bypassing the sufficient barrier conditions. The trade-off is therefore scalar tightness and computation versus dynamical structure, not formal versus empirical validity. Appendix E.4 develops this distinction.

Formal assurance from finite samples. Separating synthesis from verification allows flexible neural candidate construction while grounding validity in an independent statistical test. Successor approximation uncertainty and out-of-sample violations are incorporated into the end-to-end guarantee, rather than inferred from training loss or empirical success rates. Episode-level calibration accommodates temporal dependence within each trajectory, and the empirical-to-exact transfer avoids an additional union-bound penalty over time. Uniform satisfaction of the ideal barrier conditions covers every admissible precommitted attack; the data-driven procedure certifies the declared attack–episode law (Theorem 2). This is a trade-off in attack coverage that enables certification from finite generative data without global verification of the closed-loop inequalities, while retaining a formal guarantee at the prescribed confidence level.

Extensibility and future directions. The common pipeline across different dynamics, horizons, action dimensions, and policy families supports FoSeRL's portability. Its modular design offers complementary improvements: better witness synthesis can reduce representational conservatism, while additional successor samples trade computation for smaller approximation margins. New specifications can be added incrementally using fresh calibration and the confidence-control protocols of Appendix B.7, without predeclaring the entire future experimental family. Alongside richer synthesis and more efficient verification, a natural extension is certificate-aware policy improvement, using witness diagnostics to guide refinement while independently re-certifying each updated policy.

Ethics Statement

Our work provides a formal framework for assessing the robustness of deployed decision policies against adversarial action perturbations. Although such analyses could, in principle, be misused to identify vulnerable operating regimes, we believe that exposing and quantifying these vulnerabilities is important for the safe deployment of learned control policies. The societal benefits of advancing robustness certification therefore outweigh the potential risks, and we do not anticipate any immediate misuse arising from the methods or experiments presented in this work. No human-subject or personal data are used.

Reproducibility Statement

We have taken considerable care to ensure the reproducibility of our results. The complete assumptions, proofs, and certification procedure are provided in the main paper and appendix. Detailed descriptions of the environments, policies, attack models and norm budgets, certificate architectures, and synthesis and verification sample budgets are reported in Sec. 5 and Appendix D. Qualitative rollout videos for the reported case studies are provided through the anonymous link in Fig. 2. The complete codebase, along with the configuration files for every experiment, is available at this link.

AI Use Statement

This paper was entirely written by the authors. Large language models (LLMs), were used solely for final-stage language editing and polishing, without contributing to the scientific content or experimental results.

References

Alessandro Abate, Daniele Ahmed, Alec Edwards, Mirco Giacobbe, and Andrea Peruffo. FOSSIL: A software tool for the formal synthesis of Lyapunov functions and barrier certificates using neural networks. In *International Conference on Hybrid Systems: Computation and Control (HSCC)*, pp. 24:1–24:11, 2021.

Aaron D. Ames, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *IEEE Conference on Decision and Control (CDC)*, pp. 6271–6278, 2014.

Mahathi Anand and Majid Zamani. Formally verified neural network control barrier certificates for unknown systems. *IFAC-PapersOnLine*, 56(2):2431–2436, 2023.

Karl J. Åström and Katsuhisa Furuta. Swinging up a pendulum by energy control. *Automatica*, 36 (2):287–295, 2000.

Marc G. Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Stephen Checkoway, Damon McCoy, Brian Kantor, Danny Anderson, Hovav Shacham, Stefan Savage, Karl Koscher, Alexei Czeskis, Franziska Roesner, and Tadayoshi Kohno. Comprehensive experimental analyses of automotive attack surfaces. In *USENIX Security Symposium*, 2011.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Michael Everett, Björn Lütjens, and Jonathan P. How. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, pp. 1587–1596, 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, 2019.

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

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017. ICLR Workshop.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Karl Koscher, Alexei Czeskis, Franziska Roesner, Shwetak Patel, Tadayoshi Kohno, Stephen Checkoway, Damon McCoy, Brian Kantor, Danny Anderson, Hovav Shacham, and Stefan Savage. Experimental security analysis of a modern automobile. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 447–462, 2010.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning (ICML)*, 2019.

Xian Yeow Lee, Sambit Ghadai, Kai Liang Tan, Chinmay Hegde, and Soumik Sarkar. Spatiotemporally constrained action space attacks on deep reinforcement learning agents. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 4577–4584, 2020.

Axel Legay, Benoît Delahaye, and Saddek Bensalem. Statistical model checking: An overview. In *International Conference on Runtime Verification (RV)*, volume 6418 of *LNCS*, pp. 122–135, 2010.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. No archival venue; arXiv is the canonical form.

Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 22547–22561, 2022.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3756–3762, 2017.

Yuping Luo and Tengyu Ma. Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 25621–25632, 2021.

Ronghui Mu, Leandro Soriano Marcolino, Tianle Zhang, Yanghao Zhang, Xiaowei Huang, and Wenjie Ruan. Reward certification for policy smoothed reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 21429–21437, 2024. doi: 10.1609/aaai.v38i19.30139. Merged from duplicate key mu2024certified.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 2817–2826, 2017.

Stephen Prajna, Ali Jadbabaie, and George J. Pappas. A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE Transactions on Automatic Control*, 52(8): 1415–1428, 2007.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning (ICML)*, 2000.

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

Jacob Steinhardt and Russ Tedrake. Finite-time regional verification of stochastic non-linear systems. *The International Journal of Robotics Research*, 31(7):901–923, 2012.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Sara Taheri and Majid Zamani. Probabilistic robustness certificates against adversarial attacks. In *International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=itrFag3Z9v.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning (ICML)*, pp. 6215–6224, 2019.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, 2015.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations (ICLR)*, 2022a.

Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. COPA: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations (ICLR)*, 2022b.

Chenxi Yang, Greg Anderson, and Swarat Chaudhuri. Certifiably robust reinforcement learning through model-based abstract interpretation. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX OVERVIEW

A Notation 14

B Additional Formulation and Verification Details 15
- B.1 Attack admissibility and action projection . . . 15
- B.2 Coupling semantics . . . 15
- B.3 Monotonicity of the ideal robustness frontier . . . 16
- B.5 Analytical boundary cases and zero tolerance . . . 16
- B.4 Conservatism of the barrier witness . . . 16
- B.6 Verification planning and sample cost . . . 17
- B.7 Tested specifications and collection-level confidence . . . 17

C Proofs 17
- C.1 Augmented-state identity . . . 18
- C.2 Proof of Theorem 1 . . . 18
- C.3 Proof of Lemma 1 . . . 19
- C.4 Proof of Lemma 2 . . . 19
- C.5 Proof of Theorem 2 . . . 21

D Additional Experimental Details and Results 23
- D.1 Benchmark formulation . . . 23
- D.2 Policy families . . . 24
- D.3 Case studies . . . 24
- D.4 Certificate parameterization . . . 26
- D.5 Fixed experimental settings . . . 26
- D.6 Certificate construction and reporting workflow . . . 26
- D.7 Complete FoSeRL certification results . . . 29
- D.8 Comparison with Policy Smoothing . . . 29

E Additional Discussion 33
- E.1 From training-time robustness to long-horizon certification . . . 34
- E.2 Episode-level calibration and statistical validity . . . 34
- E.3 Reliability, confidence, and violation levels . . . 34
- E.4 Barrier witnesses versus direct terminal certification . . . 34
- E.5 Generality, simulator access, and attack coverage . . . 35
- E.6 Reading robustness frontiers and selecting specifications . . . 35
- E.7 Performance robustness versus pathwise safety . . . 35
- E.8 Certificate refinement and policy improvement . . . 36
- E.9 Limitations and outlook . . . 36

F Positioning, Baseline Selection, and Extended Comparisons 36

A NOTATION

We denote the sets of real, nonnegative real, and strictly positive real numbers by $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{>0}$, respectively. The set of positive integers is denoted by $\mathbb{N}$, and we define $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$. For integers $a, b \in \mathbb{N}_0$ with $a \leq b$, we write $\mathbb{N}_{a:b} := \{a, a+1, \ldots, b\}$, and for $N \in \mathbb{N}$ we use $[N] := \{1, \ldots, N\}$. We define the time-index sets $\mathbb{T} := \{0, \ldots, T\}$ and $\mathbb{T}' := \{0, \ldots, T-1\}$. For a scalar $x \in \mathbb{R}$, $|x|$ and $\lceil x \rceil$ denote its absolute value and ceiling, respectively. The d-dimensional Euclidean space is denoted by $\mathbb{R}^d$, and for $x \in \mathbb{R}^d$ and $p \in [1, \infty]$, $\|x\|_p$ denotes the ℓ_p norm, with $\|x\|_\infty := \max_{i \in [d]} |x_i|$. For a finite set A, $|A|$ denotes its cardinality. For a probability measure μ, $\mathrm{supp}(\mu)$ denotes its support, while $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ denote expectation and probability with respect to the underlying probability space; conditional quantities are written in the standard form $\mathbb{E}[\cdot \mid \cdot]$ and $\mathbb{P}(\cdot \mid \cdot)$. The product of probability measures is denoted by $\otimes$, and $X_i \overset{\text{i.i.d.}}{\sim} P$ indicates that the random variables X_i are independent and identically distributed according to P. For an event $\mathcal{E}$, $\mathbf{1}_{\mathcal{E}}$ denotes its indicator. We use $\mathrm{ReLU}(x) := \max\{0, x\}$ and $\mathrm{sigmoid}(x) := (1 + e^{-x})^{-1}$.

B Additional Formulation Details

This section supplements Secs. 2–4 with attack admissibility, coupling semantics, ideal-frontier monotonicity, and boundary cases of the certification framework.

B.1 Attack admissibility and action projection

The class $\mathcal{A}_{\mathrm{adv}}(K, b)$ in Definition 1 constrains the precommitted schedule and perturbation vectors, not the state-dependent command after projection. Along the attacked branch, the executed action is $\bar{u}^a(t; \upsilon) = \Pi_{\mathcal{A}}(\pi(x^a(t), w(t)) + \Delta_t)$, where $\Delta_t = 0$ outside $\mathcal{T}_{\mathrm{atk}}$. Thus an admissible attack is specified before the episode begins, while the executed action remains in $\mathcal{A}$ for every realized state. We do not test attack membership by requiring $\pi(x^a(t), w(t)) + \Delta_t \in \mathcal{A}$ along the realized trajectory: such a membership test would depend on that trajectory rather than only on $(\mathcal{T}_{\mathrm{atk}}, \Delta)$. Instead, projection is part of the attacked closed-loop dynamics, and the budget b always refers to $\|\Delta_t\|_p$ before projection. The norm used in a projection bound must be specified. For a closed convex action set, Euclidean projection is nonexpansive in the Euclidean norm. For each realized $u = \pi(x, w) \in \mathcal{A}$,

$$\|\Pi_{\mathcal{A}}(u + \delta) - u\|_2 = \|\Pi_{\mathcal{A}}(u + \delta) - \Pi_{\mathcal{A}}(u)\|_2 \le \|\delta\|_2. \tag{20}$$

For a box action set, projection is coordinatewise clipping and $|[\Pi_{\mathcal{A}}(u + \delta)]_j - u_j| \le |\delta_j|$ for every coordinate j. Consequently, the same displacement bound holds in each of the ℓ_1, ℓ_2, and ℓ_∞ norms for boxes. We do not assert this all-norm property for arbitrary closed convex sets. The attack class and its certification criterion remain defined by the pre-projection budget and the specified projected dynamics.

B.2 Coupling semantics

The tail criterion in Definition 3 depends on the joint coupling of the nominal and attacked executions, not only on their trajectory marginals. To make this distinction explicit, fix an attack υ and let C be any coupling with the prescribed nominal and attacked marginals. Linearity of expectation gives

$$\mathbb{E}_{\mathsf{C}}[R_{\mathrm{nom}} - R_{\mathrm{att}}] = J(\pi) - \mathbb{E}_{\omega}[R_{\mathrm{att}}(\upsilon, \omega)]\,. \tag{21}$$

Thus the mean degradation is coupling-invariant, whereas $\mathbb{P}_{\mathsf{C}}(R_{\mathrm{nom}} - R_{\mathrm{att}} \le \alpha)$ generally is not. A coupling must therefore be fixed to define the paired tail event.

Definition 2 uses the synchronous coupling induced by the declared execution representation (π, F, P_w). Both branches share $x(0)$ and the complete random-input sequence, including deployment-time policy randomization, while evaluating the policy at their own states. Hence the certification statement is relative to this fixed shared-randomness semantics, chosen before synthesis and verification; the marginal transition kernel and policy action law alone do not determine the paired tail distribution.

The synchronous coupling also gives an exact nominal anchor. If $K = 0$ or $b = 0$ and $x^0(t) = x^a(t)$, the shared $w(t)$ produces the same policy action in both branches. Projection leaves that admissible action unchanged, and the same F and $w(t)$ then produce equal next states and rewards. Induction from the common initial state yields $D_{\mathrm{traj}}(\upsilon, \omega) \equiv 0$ for every ω. This is an analytical property of the declared coupling, not an empirical calibration result.

The need to specify the coupling cannot be removed by specifying only the marginals. For example, take $T = 2$, $\gamma = 1$, $x(0) = 0$, $\pi = 0$, $\mathcal{S} = \{0, 1\}$, $\mathcal{A} = [0, 1]$, $r(x, u) = x$, and fair Bernoulli noise w. Let $F_1(x, u, w) = w$, while $F_2(x, u, w) = w$ for $u < 1/2$ and $F_2(x, u, w) = 1 - w$ otherwise. Both realize the same Bernoulli transition kernel. A time-zero perturbation from zero to one gives $D_{\mathrm{traj}} = 0$ under F_1 but $D_{\mathrm{traj}} = 2w(0) - 1$ under F_2. At $\alpha = 0$, the corresponding paired reliabilities are one and one-half despite identical trajectory marginals.

B.3 Monotonicity of the ideal robustness frontier

Fix the policy, horizon, attack norm, coupling, and the pair (α, ρ). The attack classes are nested:

$$\mathcal{A}_{\mathrm{adv}}(K, b) \subseteq \mathcal{A}_{\mathrm{adv}}(K, b'), \qquad 0 \le b \le b', \tag{22}$$

$$\mathcal{A}_{\mathrm{adv}}(K, b) \subseteq \mathcal{A}_{\mathrm{adv}}(K', b), \qquad 0 \le K \le K' \le T. \tag{23}$$

Since Definition 3 quantifies over every admissible attack, satisfaction for a larger class implies satisfaction for each contained class. Let $\mathcal{B}_K := \{b \in \mathbb{R}_{\ge 0} : \text{equation 2 holds}\}$ denote the set of magnitudes satisfying the ideal criterion. By 22, this set is downward closed: if $b' \in \mathcal{B}_K$, then $b \in \mathcal{B}_K$ for every $0 \le b \le b'$. It is nonempty because the zero-attack property implies $0 \in \mathcal{B}_K$. Moreover, 23 gives $\mathcal{B}_{K'} \subseteq \mathcal{B}_K$ when $K \le K'$, and therefore

$$b^*_{\mathrm{cert}}(K') \le b^*_{\mathrm{cert}}(K) \qquad \text{for } K \le K'. \tag{24}$$

This monotonicity concerns the ideal per-attack objective and does not require a sampled-prefix construction. The distributional guarantee of Theorem 2 has a different scope. Under $\mathbb{Q}_{K,b} = \mathcal{U}_{K,b} \otimes \mathbb{P}_\omega$, changing K or b may change the attack sampling law; set inclusion alone therefore does not imply monotonicity of the distributional success probability. In particular, a certificate at one (K, b) does not automatically certify other budgets under different operational laws. This is a distinction between the ideal and distributional criteria, not merely a finite-sample effect.

B.4 Conservatism of the barrier witness

The conditions of Definition 6 are sufficient but not necessary. Conservatism arises both from requiring one common barrier across admissible attacks and from the pointwise initial-set bound.

Common witness across attacks. Let $T=2$, $K=b=\gamma=1$, $\mathcal{S}=\mathcal{A}=[-1,1]$, $x(0)=0$, and $\pi=0$. Take $F(x,u,w)=w$ with fair $w\in\{-1,1\}$ and $r(x,u)=-(u^2+xu)/2$, and set $\alpha=\rho=1/2$. A time-zero perturbation u gives $D_{\mathrm{traj}}=u^2/2\le 1/2$. A time-one perturbation gives $(u^2+su)/2$ at $x(1)=s\in\{-1,1\}$; for each precommitted u, at most one sign exceeds $1/2$. Thus every admissible precommitted attack satisfies the target.

At the shared time-one states $(s,s,0)$, however, a common barrier must also accommodate the respective attacks $u=s$, which give terminal gap one. Hence $\kappa B((s,s,0),1)+b_3\ge b_2$ for both signs, and the initial drift inequality implies $b_1\kappa^2+b_3(1+\kappa)\ge b_2$. The resulting ideal barrier bound is nonpositive even though the universal robustness property holds. This is conservatism of the common witness, not a counterexample to Theorem 1.

Pointwise initial condition. The proof of Theorem 1 only requires $\mathbb{E}_{x\sim\mu_0}[B((x,x,0),0)]\le b_1$. Definition 6 instead imposes the stronger pointwise condition $B(\hat{x},0)\le b_1$ on $\mathcal{X}_0$. We retain this form because the verifier directly calibrates the corresponding residual q_1; no separate empirical test of the initial expectation is introduced.

Unit reliability. For $T\ge 1$ and $b_1,b_3\in\mathbb{R}_{>0}$, $b_1\kappa^T+b_3\mathcal{G}_T(\kappa)>0$. Hence the sufficient bound in Theorem 1 cannot certify $\rho=1$, even when the underlying robustness property holds. This limitation concerns the certificate conditions, not the property itself.

B.5 Analytical boundary cases and zero tolerance

Some specifications are resolved analytically before invoking the data-driven procedure. If $K=0$ or $b=0$, synchronous coupling gives $D_{\mathrm{traj}}(\upsilon,\omega)\equiv 0$ pathwise, so Definition 3 holds for every $\alpha\ge 0$ and $\rho\le 1$. If $\alpha\ge 2S_{\max}$, bounded rewards imply $D_{\mathrm{traj}}(\upsilon,\omega)\le 2S_{\max}\le\alpha$ for every admissible attack and episode. These are analytical certificates, not finite-sample acceptances.

Zero-tolerance boundary. Assume $S_{\max}>0$ and that B is continuous in z. At $\alpha=0$, global terminal separation for all $z>0$ implies $B((x,x,0),T)\ge b_2$ by continuity. Under zero attack, the actual terminal state has this form; hence the ideal recursion cannot yield a positive reliability bound through such a globally separating continuous candidate. The analytical zero-attack certificate above bypasses this restriction. The data-driven theorem instead controls violations under the operational episode law and does not require global terminal separation.

B.6 VERIFICATION PLANNING AND SAMPLE COST

For one fixed tested specification, formal verification uses N paired calibration episodes and NTM augmented successor queries when all T time indices require Monte Carlo successor evaluation; each augmented query advances both branches. A complete synthesis pass uses $N_s T M_s$ successor queries, with additional cost if successor batches are regenerated during optimization.

For a fixed M, the choice $\epsilon_M = B_{\max}\sqrt{\log(1/\beta_0)/(2M)}$ saturates the sufficient condition in equation 14. The conformal calibration also imposes a finite-N statistical floor. Since the smallest rank bound is $\bar{v}_N = 1 - (\beta/N)^{1/N}$, define $c_{\mathrm{stat,min}} := [1 - (\beta/N)^{1/N}]/(1 - \beta_0)$. Because the normalized barrier contribution is strictly positive, acceptance requires $c_{\mathrm{stat,min}} < 1 - \rho$. This condition can be checked before calibration and expensive successor sampling; failure means that N must be increased or no certificate can be obtained with that verification design. For a planned positive dynamical contribution d, the stronger planning condition is $c_{\mathrm{stat,min}} \le 1 - \rho - d$.

Avoiding a union bound over the horizon does not eliminate horizon dependence. For rank k, the calibration margin contributes $r_k[(\kappa^*)^T + \mathcal{G}_T(\kappa^*)]/b_2^*$ to the normalized barrier term, reducing to $(T+1)r_k/b_2^*$ when $\kappa^* = 1$.

B.7 SCOPE ACROSS TESTED SPECIFICATIONS AND COLLECTION-LEVEL CONFIDENCE

Theorem 2 gives a separate confidence statement for each frozen tested specification, including its policy, coupling, attack budget, tolerance, operational law, candidate, and sampling design. New specifications may therefore be added incrementally using fresh calibration data; the complete future experimental family need not be fixed in advance.

Simultaneous validity over conformal ranks within one specification does not imply simultaneous validity across different specifications. Consequently, a displayed collection of individually certified points carries no automatic family-wise guarantee, interpolation guarantee, or downward-closure claim when the operational law changes. Likewise, selecting the largest accepted budget after repeated formal tests does not preserve the same nominal confidence for that selected output. For a single selected-budget claim at a stated confidence, candidate selection may use design or previous search data, but the selected specification must then be frozen and assessed using fresh final calibration data. Repeating that final test until acceptance requires additional confidence control.

For a simultaneous claim over a finite predeclared collection of J specifications, fix all candidates and designs before calibration and choose $\beta_j > 0$ with $\sum_{j=1}^{J} \beta_j \le \beta_{\mathrm{total}}$. Applying Theorem 2 to specification j with $\beta = \beta_j$ and taking a union bound yields probability at least $1 - \beta_{\mathrm{total}}$ that every accepted specification satisfies its own operational target. Independence across specifications is not required, provided each test satisfies its own calibration conditions.

The entire future family also need not be predeclared. For sequential testing, one may fix a summable allocation such as $\beta_j = \beta_{\mathrm{total}}/[j(j+1)]$. Before attempt j, freeze its specification using only previously available information and draw fresh calibration and successor samples conditionally independently of that information. The conditional false-acceptance probability at attempt j is then at most β_j; summing these bounds gives the same $1 - \beta_{\mathrm{total}}$ collection-level guarantee over all attempts, including under stopping. Failed attempts still consume their assigned confidence budget. These stronger protocols do not retrospectively apply to runs performed at other confidence levels and do not certify untested or interpolated budgets.

C PROOFS

This appendix establishes the theoretical results used in the main text. We first establish the augmented-state identity and prove the ideal sequential robustness guarantee. We then prove the two statistical ingredients of the primary data-driven verification method: the finite-sample supermartingale approximation and the conformal calibration guarantee. These results are combined directly with the stochastic barrier argument in the proof of the final data-driven FoSeRL certificate.

C.1 AUGMENTED-STATE IDENTITY

By Definition 4, the third component of the augmented state satisfies $z(0) = 0$ and, for each $t \in \mathbb{T}'$,

$$z(t+1) - z(t) = \gamma^t \left[r(x^0(t), u^0(t)) - r(x^a(t), \bar{u}^a(t; \upsilon)) \right]. \tag{25}$$

Summing 25 over $t = 0, \ldots, T-1$ gives

$$\begin{aligned} z(T) &= \sum_{t=0}^{T-1} \gamma^t \left[r(x^0(t), u^0(t)) - r(x^a(t), \bar{u}^a(t; \upsilon)) \right] \\ &= R_{\mathrm{nom}}(\omega) - R_{\mathrm{att}}(\upsilon, \omega) = D_{\mathrm{traj}}(\upsilon, \omega). \end{aligned} \tag{26}$$

Hence the terminal augmented coordinate is exactly the paired trajectory-level degradation. For completeness, fix any $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$. Since the nominal branch depends only on ω and the fixed policy π,

$$\mathbb{E}_\omega[R_{\mathrm{nom}}(\omega)] = J(\pi). \tag{27}$$

Taking expectation in 26 over ω therefore gives

$$\mathbb{E}_\omega[D_{\mathrm{traj}}(\upsilon, \omega)] = J(\pi) - \mathbb{E}_\omega[R_{\mathrm{att}}(\upsilon, \omega)], \tag{28}$$

which is the mean identity stated after Definition 2.

C.2 PROOF OF THEOREM 1

Proof. Let B satisfy Definition 6 with constants $b_1, b_2, b_3 \in \mathbb{R}_{>0}$ and $\kappa \in (0, 1]$. Fix an arbitrary $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$. All probabilities and expectations below are over the episode randomness ω, conditional on the fixed policy and execution representation. Define $B_t := B(\hat{x}(t), t)$ and the filtration

$$\mathcal{F}_t := \sigma\big(x(0), w(0), \ldots, w(t-1)\big), \qquad t \in \mathbb{T}. \tag{29}$$

For fixed υ, the augmented state $\hat{x}(t)$ and hence B_t are $\mathcal{F}_t$-measurable. Because $\hat{x}(0) \in \mathcal{X}_0$ almost surely, the initial barrier condition gives

$$\mathbb{E}_\omega[B_0] \le b_1. \tag{30}$$

Since $\hat{x}(t) \in \mathcal{X}_t^\upsilon$ almost surely and $w(t)$ is independent of $\mathcal{F}_t$ with law P_w by Assumption 2, the third condition in Definition 6 and the tower property give

$$\begin{aligned} \mathbb{E}_\omega[B_{t+1}] &= \mathbb{E}_\omega[\mathbb{E}_\omega[B_{t+1} \mid \mathcal{F}_t]] \\ &\le \kappa \mathbb{E}_\omega[B_t] + b_3, \qquad t \in \mathbb{T}'. \end{aligned} \tag{31}$$

The conditional barrier inequality implies that the process $\kappa^{T-t} B_t + b_3 \sum_{j=0}^{T-t-1} \kappa^j$, $t \in \mathbb{T}$, is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}_{t \in \mathbb{T}}$, where an empty sum is zero. Iterating 31 and using 30 yields

$$\mathbb{E}_\omega[B_T] \le b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa). \tag{32}$$

By 26, the terminal unsafe event is

$$\{\hat{x}(T) \in \mathcal{X}_u^\alpha\} = \{D_{\mathrm{traj}}(\upsilon, \omega) > \alpha\}. \tag{33}$$

On this event the terminal separation condition gives $B_T \ge b_2$. Since $B_T \ge 0$, Markov's inequality and 32 imply

$$\mathbb{P}_\omega\left(D_{\mathrm{traj}}(\upsilon, \omega) > \alpha\right) \le \frac{b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa)}{b_2}. \tag{34}$$

Taking complements gives

$$\mathbb{P}_\omega\left(D_{\mathrm{traj}}(\upsilon, \omega) \le \alpha\right) \ge 1 - \frac{b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa)}{b_2}. \tag{35}$$

Finally, the scalar condition in Definition 6 implies

$$1 - \frac{b_1 \kappa^T + b_3 \mathcal{G}_T(\kappa)}{b_2} \ge \rho. \tag{36}$$

Because υ was arbitrary, the bound holds for every $\upsilon \in \mathcal{A}_{\mathrm{adv}}(K, b)$, proving 5 and completing the proof. □

C.3 Proof of Lemma 1

Proof. Fix $(\hat{x}, \upsilon, t) \in \mathcal{Z}_3$. The candidate and verification choices are fixed; the only randomness considered here is that of the fresh successor samples. Let $Y_m := B^*_\phi(f(\hat{x}, \upsilon, w_m, t), t+1)$, $\widehat{\mu}_M := M^{-1}\sum_{m=1}^{M} Y_m$, and $\mu := \mathbb{E}_w[Y_m \mid \hat{x}, \upsilon]$. By the assumed output bound, $Y_m \in (0, B_{\max})$ almost surely, and $Y_1, \dots, Y_M$ are conditionally i.i.d. given $(\hat{x}, \upsilon, t)$. Define the approximation event $E_{\text{Approx}} := \{q_3(\hat{x}, \upsilon, t) \le \bar{q}_3(\hat{x}, \upsilon, t)\}$. At the fixed tuple, $\overline{E}^*_M = \widehat{\mu}_M + \epsilon_M$ by equation 12, while the other terms in q_3 and $\bar{q}_3$ coincide. Therefore, $\neg E_{\text{Approx}}$ is exactly the event $\mu - \widehat{\mu}_M > \epsilon_M$. Applying the one-sided Hoeffding inequality to the conditionally i.i.d. variables $Y_m \in (0, B_{\max})$, and then using 14, gives

$$\begin{aligned}\mathbb{P}_W(\neg E_{\text{Approx}} \,|\, \hat{x}, \upsilon, t) &= \mathbb{P}_W(\mu - \widehat{\mu}_M > \epsilon_M \,|\, \hat{x}, \upsilon, t) \\ &\le \exp\left(-\frac{2M(\epsilon_M)^2}{(B_{\max})^2}\right) \le \beta_0.\end{aligned} \tag{37}$$

Taking complements therefore yields

$$\mathbb{P}_W(E_{\text{Approx}} \,|\, \hat{x}, \upsilon, t) = \mathbb{P}_W(q_3(\hat{x}, \upsilon, t) \le \bar{q}_3(\hat{x}, \upsilon, t) \,|\, \hat{x}, \upsilon, t) \ge 1 - \beta_0. \tag{38}$$

Hence the corrected empirical supermartingale residual upper-bounds the exact supermartingale residual at the fixed state–attack–time tuple with the confidence stated in Lemma 1. □

C.4 Proof of Lemma 2

Proof. Condition on the synthesis data and on all quantities fixed before calibration. Under this conditioning, the score construction in 15 is fixed, and the calibration scores $S_1, \dots, S_N$ are i.i.d. with the same distribution as the score $S(\bar{\zeta})$ of a fresh independent augmented episode $\bar{\zeta}$. Let

$$F(s) := \mathbb{P}_{\bar{\zeta}}\big(S(\bar{\zeta}) \le s\big), \qquad s \in \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}, \tag{39}$$

denote the population distribution function of the score. The score distribution may contain atoms, including an atom at $+\infty$ induced by hard unsafe-terminal violations; no continuity or tie-breaking assumption is required. For each rank $k \in [N]$, define the conditional exceedance probability associated with the k-th order statistic by

$$V_k(\mathcal{D}_{\text{cal}}) := \mathbb{P}_{\bar{\zeta}}\big(S(\bar{\zeta}) > S_{(k)} \,\big|\, \mathcal{D}_{\text{cal}}\big) = 1 - F\big(S_{(k)}\big). \tag{40}$$

We first control V_k for a fixed rank k and a fixed $v \in [0, 1)$. Define the population $(1-v)$-quantile

$$c_v := \inf\big\{s \in \overline{\mathbb{R}} : F(s) \ge 1 - v\big\}, \tag{41}$$

and let

$$p_v^- := \mathbb{P}_{\bar{\zeta}}\big(S(\bar{\zeta}) < c_v\big). \tag{42}$$

By the definition of c_v, for every $s < c_v$,

$$F(s) < 1 - v. \tag{43}$$

Therefore, by continuity of probability from below,

$$p_v^- = \mathbb{P}_{\bar{\zeta}}\big(S(\bar{\zeta}) < c_v\big) = \sup_{s < c_v} F(s) \le 1 - v. \tag{44}$$

Suppose now that $V_k(\mathcal{D}_{\text{cal}}) > v$. By 40, this implies

$$F\big(S_{(k)}\big) < 1 - v. \tag{45}$$

By the definition of c_v in 41, 45 implies

$$S_{(k)} < c_v. \tag{46}$$

Hence at least k of the N calibration scores are strictly smaller than c_v. Defining

$$N_v := \sum_{i=1}^{N} \mathbf{1}_{\{S_i < c_v\}}, \tag{47}$$

we therefore have the event inclusion

$$\{V_k(\mathcal{D}_{\text{cal}}) > v\} \subseteq \{N_v \geq k\}. \tag{48}$$

Since c_v depends only on the fixed population score distribution, the indicators in 47 are i.i.d. Bernoulli random variables with success probability p_v^-. Therefore,

$$N_v \sim \text{Binomial}\left(N, p_v^-\right). \tag{49}$$

Using 48, the binomial law in 49, the bound $p_v^- \leq 1 - v$ from 44, and monotonicity of the binomial upper tail in its success probability, we obtain

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_{\text{cal}}}(V_k(\mathcal{D}_{\text{cal}}) > v) &\leq \mathbb{P}_{\mathcal{D}_{\text{cal}}}(N_v \geq k) \\ &= \sum_{j=k}^{N} \binom{N}{j} (p_v^-)^j (1 - p_v^-)^{N-j} \\ &\leq \sum_{j=k}^{N} \binom{N}{j} (1 - v)^j v^{N-j}. \end{aligned} \tag{50}$$

We now set $v = \bar{v}_k$, where $\bar{v}_k$ is defined by 16. By construction,

$$\sum_{j=k}^{N} \binom{N}{j} (1 - \bar{v}_k)^j \bar{v}_k^{N-j} \leq \frac{\beta}{N}. \tag{51}$$

Combining 50 and 51 gives, for every fixed $k \in [N]$,

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}(V_k(\mathcal{D}_{\text{cal}}) > \bar{v}_k) \leq \frac{\beta}{N}. \tag{52}$$

Define the rank-wise validity events

$$E_k := \{V_k(\mathcal{D}_{\text{cal}}) \leq \bar{v}_k\}, \qquad k \in [N]. \tag{53}$$

Equation 52 implies

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}(\neg E_k) \leq \frac{\beta}{N}, \qquad k \in [N]. \tag{54}$$

Applying a union bound over all N ranks yields

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(\bigcap_{k=1}^{N} E_k\right) &= 1 - \mathbb{P}_{\mathcal{D}_{\text{cal}}}\left(\bigcup_{k=1}^{N} \neg E_k\right) \\ &\geq 1 - \sum_{k=1}^{N} \mathbb{P}_{\mathcal{D}_{\text{cal}}}(\neg E_k) \\ &\geq 1 - \beta. \end{aligned} \tag{55}$$

Using the definition of V_k in 40, the event in 55 is exactly

$$\left\{\forall k \in [N] : \mathbb{P}_{\bar{\zeta}}\left(S(\bar{\zeta}) > S_{(k)} \,\middle|\, \mathcal{D}_{\text{cal}}\right) \leq \bar{v}_k\right\}. \tag{56}$$

Combining equation 55 and equation 56 yields

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left[\forall k \in [N] : \quad \mathbb{P}_{\bar{\zeta}}\left(S(\bar{\zeta}) > S_{(k)} \mid \mathcal{D}_{\text{cal}}\right) \leq \bar{v}_k\right] \geq 1 - \beta. \tag{57}$$

This proves Lemma 2. In particular, the guarantee holds simultaneously for all ranks on the same calibration-validity event. Therefore, a rank may subsequently be selected as a function of the calibration data, including through 17, without introducing any additional confidence penalty. □

C.5 Proof of Theorem 2

Proof. Condition on all pre-calibration choices, including the data used for construction or selection. Under this conditioning, the candidate, constants, operational law, and verification design are fixed, and the calibration episodes are independent copies drawn as in Definition 9. Define the simultaneous conformal-validity event

$$E_{\mathrm{Conf}} := \left\{\forall k \in [N] : \mathbb{P}_{\bar{\zeta}}\left(S(\bar{\zeta}) > S_{(k)} \,\middle|\, \mathcal{D}_{\mathrm{cal}}\right) \le \bar{v}_k\right\}. \tag{58}$$

By Lemma 2,

$$\mathbb{P}_{\mathcal{D}_{\mathrm{cal}}}(E_{\mathrm{Conf}}) \ge 1 - \beta. \tag{59}$$

Because the bounds in 58 hold simultaneously for every rank $k \in [N]$, the data-dependent selection of k^* in 17 does not incur an additional confidence penalty. Fix an arbitrary realization of $\mathcal{D}_{\mathrm{cal}}$ belonging to E_{Conf} for which $\bar{q}_4 \le 0$. By the empty-rank convention in Sec. 4.2, this implies $\mathcal{K}_{\mathrm{fin}} \neq \varnothing$. Hence k^*, r^*, $\widetilde{b}_1^*$, $\widetilde{b}_3^*$, and $\bar{v}_{k^*}$ are well defined and fixed for the remainder of the argument. All probabilities below involving a fresh episode are evaluated for this fixed calibration realization; we suppress this conditioning for notational simplicity. For a fresh augmented episode $\bar{\zeta} = (\upsilon, \omega, W)$, define the selected empirical barrier event

$$E_{M,*} := \left\{\bar{q}_1(\hat{x}(0), \upsilon, 0) \le r^*,\ \bar{q}_2(\hat{x}(T), \upsilon, T) \le 0,\ \max_{t \in \mathbb{T}'} \bar{q}_3(\hat{x}(t), \upsilon, t) \le r^*\right\}. \tag{60}$$

Since $k^* \in \mathcal{K}_{\mathrm{fin}}$,

$$S_{(k^*)} < +\infty, \qquad r^* = \max\left\{0, S_{(k^*)}\right\}. \tag{61}$$

If $\neg E_{M,*}$ occurs, then either $\bar{q}_2(\hat{x}(T), \upsilon, T) > 0$, in which case $S(\bar{\zeta}) = +\infty$, or one of the calibrated residuals exceeds r^*. In the latter case,

$$S(\bar{\zeta}) > r^* \ge S_{(k^*)}. \tag{62}$$

Therefore,

$$\neg E_{M,*} \subseteq \left\{S(\bar{\zeta}) > S_{(k^*)}\right\}. \tag{63}$$

Since the fixed calibration realization belongs to E_{Conf}, equations 58 and 63 imply

$$V_{M,*} := \mathbb{P}_{\bar{\zeta}}(\neg E_{M,*}) \le \bar{v}_{k^*}. \tag{64}$$

We next transfer this empirical violation probability to the exact state-dependent barrier conditions. For the corresponding episode $\zeta = (\upsilon, \omega) \sim \mathbb{Q}_{K,b}$, define the exact adjusted barrier event

$$E_{\mathrm{Barrier},*} := \{q_1(\hat{x}(0), \upsilon, 0) \le r^*,\ q_2(\hat{x}(T), \upsilon, T) \le 0,\ q_3(\hat{x}(t), \upsilon, t) \le r^*\ \forall t \in \mathbb{T}'\}. \tag{65}$$

Thus, on $E_{\mathrm{Barrier},*}$, the initial and supermartingale conditions hold with the adjusted constants

$$\widetilde{b}_1^* = b_1^* + r^*, \qquad \widetilde{b}_3^* = b_3^* + r^*, \tag{66}$$

while the unsafe-terminal condition retains the original threshold b_2^*. Let

$$V_* := \mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\mathrm{Barrier},*}). \tag{67}$$

If $V_* = 0$, the empirical-to-exact risk bound established below is immediate. Assume therefore that $V_* > 0$. On an episode in $\neg E_{\mathrm{Barrier},*}$, let $E_{\mathrm{Det},*}$ denote the event that the finite-M test detects at least one exact adjusted barrier violation. If $q_1(\hat{x}(0), \upsilon, 0) > r^*$ or $q_2(\hat{x}(T), \upsilon, T) > 0$, the violation is detected deterministically because

$$\bar{q}_1 = q_1, \qquad \bar{q}_2 = q_2. \tag{68}$$

Otherwise, an exact supermartingale violation must occur. On this remaining event, define the first such time by

$$\tau := \min\{t \in \mathbb{T}' : q_3(\hat{x}(t), \upsilon, t) > r^*\}. \tag{69}$$

Conditional on the complete episode ζ, the index τ and the tuple $(\hat{x}(\tau), \upsilon, \tau)$ are fixed, whereas the successor batch associated with that tuple is fresh and conditionally i.i.d. By Lemma 1,

$$\mathbb{P}_W(q_3(\hat{x}(\tau), \upsilon, \tau) \le \bar{q}_3(\hat{x}(\tau), \upsilon, \tau) \,|\, \zeta) \ge 1 - \beta_0. \tag{70}$$

Since the exact residual at time τ already satisfies $q_3(\hat{x}(\tau), \upsilon, \tau) > r^*$, the event in 70 implies

$$\bar{q}_3(\hat{x}(\tau), \upsilon, \tau) > r^*, \tag{71}$$

and therefore produces an empirical violation. Consequently, whether the exact violation occurs in q_1, q_2, or q_3, its conditional probability of being detected is at least $1-\beta_0$. Hence

$$\mathbb{P}_{\bar{\zeta}}(E_{\mathrm{Det},*} \,|\, \neg E_{\mathrm{Barrier},*}) \geq 1-\beta_0. \tag{72}$$

Moreover, every detected exact violation is also an empirical violation, so

$$E_{\mathrm{Det},*} \cap \neg E_{\mathrm{Barrier},*} \subseteq \neg E_{M,*}. \tag{73}$$

Using 72 and 73,

$$\begin{aligned} V_{M,*} &\geq \mathbb{P}_{\bar{\zeta}}(E_{\mathrm{Det},*} \cap \neg E_{\mathrm{Barrier},*}), && (74)\\ &= \mathbb{P}_{\bar{\zeta}}(E_{\mathrm{Det},*} \,|\, \neg E_{\mathrm{Barrier},*})\, V_*, && (75)\\ &\geq (1-\beta_0) V_*. && (76) \end{aligned}$$

Combining 64 and 76 yields

$$V_* \leq \frac{V_{M,*}}{1-\beta_0} \leq \frac{\bar{v}_{k^*}}{1-\beta_0}. \tag{77}$$

The argument uses only the first exact supermartingale violation after conditioning on the complete episode. Lemma 1 is therefore invoked once per exact violating episode, and no union bound over the horizon is required. It remains to relate satisfaction of the exact adjusted barrier conditions to the trajectory-level degradation event. Define the trajectory-safety event

$$E_{\mathrm{Safe}} := \{D_{\mathrm{traj}}(\upsilon, \omega) \leq \alpha\}, \tag{78}$$

and let

$$B_t := B^*_\phi(\hat{x}(t), t). \tag{79}$$

With

$$\mathcal{F}_t := \sigma\big(\upsilon, x(0), w(0), \ldots, w(t-1)\big), \qquad t \in \mathbb{T}, \tag{80}$$

define the good-prefix events

$$G_0 := \{q_1(\hat{x}(0), \upsilon, 0) \leq r^*\}, \qquad G_{t+1} := G_t \cap \{q_3(\hat{x}(t), \upsilon, t) \leq r^*\}, \quad t \in \mathbb{T}'. \tag{81}$$

By Assumption 2 and the product law in Definition 7, $w(t)$ is independent of $\mathcal{F}_t$ with law P_w. Thus the expectation in equation 10 satisfies, along the trajectory,

$$\mathbb{E}[B_{t+1} \mid \mathcal{F}_t] = \mathbb{E}_{w \sim P_w}\left[B^*_\phi(f(\hat{x}(t), \upsilon, w, t), t+1)\right]. \tag{82}$$

On G_{t+1}, the exact adjusted supermartingale condition therefore gives

$$\mathbb{E}[B_{t+1} \,|\, \mathcal{F}_t] \leq \kappa^* B_t + \widetilde{b}^*_3. \tag{83}$$

Moreover, G_{t+1} is $\mathcal{F}_t$-measurable, $G_{t+1} \subseteq G_t$, and $B_t \geq 0$. Thus, by the tower property,

$$\begin{aligned} \mathbb{E}\big[B_{t+1}\mathbf{1}_{G_{t+1}}\big] &= \mathbb{E}\big[\mathbf{1}_{G_{t+1}}\mathbb{E}[B_{t+1} \,|\, \mathcal{F}_t]\big] \\ &\leq \kappa^* \mathbb{E}\big[B_t \mathbf{1}_{G_{t+1}}\big] + \widetilde{b}^*_3 \mathbb{P}(G_{t+1}) \\ &\leq \kappa^* \mathbb{E}[B_t \mathbf{1}_{G_t}] + \widetilde{b}^*_3. \end{aligned} \tag{84}$$

On G_0, the adjusted initial condition gives

$$B_0 \leq b^*_1 + r^* = \widetilde{b}^*_1. \tag{85}$$

Iterating 84 from 85 yields

$$\mathbb{E}[B_T \mathbf{1}_{G_T}] \leq \widetilde{b}^*_1 (\kappa^*)^T + \widetilde{b}^*_3 \mathcal{G}_T(\kappa^*). \tag{86}$$

On $\neg E_{\mathrm{Safe}} \cap E_{\mathrm{Barrier},*}$, all good-prefix conditions hold. In addition, $\neg E_{\mathrm{Safe}}$ implies $\hat{x}(T) \in \mathcal{X}^\alpha_u$. Since $q_2(\hat{x}(T), \upsilon, T) \leq 0$ on $E_{\mathrm{Barrier},*}$, the unsafe-terminal residual 9 gives

$$B_T \geq b^*_2. \tag{87}$$

Consequently,

$$\begin{aligned} b^*_2 \mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\mathrm{Safe}} \cap E_{\mathrm{Barrier},*}) &\leq \mathbb{E}\big[B_T \mathbf{1}_{\{\neg E_{\mathrm{Safe}} \cap E_{\mathrm{Barrier},*}\}}\big] \\ &\leq \mathbb{E}[B_T \mathbf{1}_{G_T}] \\ &\leq \widetilde{b}^*_1 (\kappa^*)^T + \widetilde{b}^*_3 \mathcal{G}_T(\kappa^*). \end{aligned} \tag{88}$$

Define

$$\delta_B^* := \frac{\widetilde{b}_1^*(\kappa^*)^T + \widetilde{b}_3^*\mathcal{G}_T(\kappa^*)}{b_2^*}. \tag{89}$$

Then 88 implies

$$\mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\text{Safe}} \cap E_{\text{Barrier},*}) \le \delta_B^*. \tag{90}$$

We now decompose trajectory-level failure according to whether the exact adjusted barrier conditions hold. Using 77 and 90,

$$\begin{aligned}\mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\text{Safe}}) &= \mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\text{Safe}} \cap E_{\text{Barrier},*}) \\ &\quad + \mathbb{P}_{\mathbb{Q}_{K,b}}(\neg E_{\text{Safe}} \cap \neg E_{\text{Barrier},*}) \\ &\le \delta_B^* + V_* \\ &\le \frac{\widetilde{b}_1^*(\kappa^*)^T + \widetilde{b}_3^*\mathcal{G}_T(\kappa^*)}{b_2^*} + \frac{\bar{v}_{k^*}}{1-\beta_0}.\end{aligned} \tag{91}$$

By the definition of $\bar{q}_4$ in 18 and the fact that $b_2^* > 0$, the condition $\bar{q}_4 \le 0$ is equivalent to

$$\frac{\widetilde{b}_1^*(\kappa^*)^T + \widetilde{b}_3^*\mathcal{G}_T(\kappa^*)}{b_2^*} + \frac{\bar{v}_{k^*}}{1-\beta_0} \le 1-\rho. \tag{92}$$

Combining 91 and 92 therefore gives

$$\bar{q}_4 \le 0 \implies \mathbb{P}_{(\upsilon,\omega)\sim\mathbb{Q}_{K,b}}(\neg E_{\text{Safe}}) \le 1-\rho. \tag{93}$$

Taking complements,

$$\bar{q}_4 \le 0 \implies \mathbb{P}_{(\upsilon,\omega)\sim\mathbb{Q}_{K,b}}(E_{\text{Safe}}) \ge \rho. \tag{94}$$

Since $E_{\text{Safe}} = \{D_{\text{traj}}(\upsilon,\omega) \le \alpha\}$, this is equivalent to

$$\bar{q}_4 \le 0 \implies \mathbb{P}_{(\upsilon,\omega)\sim\mathbb{Q}_{K,b}}(D_{\text{traj}}(\upsilon,\omega) \le \alpha) \ge \rho. \tag{95}$$

We have established 95 for every calibration realization belonging to E_{Conf}. Since 59 gives $\mathbb{P}_{\mathcal{D}_{\text{cal}}}(E_{\text{Conf}}) \ge 1-\beta$, it follows that

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left[\bar{q}_4 \le 0 \implies \mathbb{P}_{(\upsilon,\omega)\sim\mathbb{Q}_{K,b}}(D_{\text{traj}}(\upsilon,\omega) \le \alpha) \ge \rho\right] \ge 1-\beta. \tag{96}$$

Equation 96 is exactly 19, completing the proof. □

D Additional Experimental Details and Results

This appendix gives the complete experimental specification underlying Sec. 5. We first state the benchmark instantiation used in our experiments, then summarize the policy families and the six case studies, describe the certificate parameterization, and finally collect all fixed experimental settings in a single table. Numerical robustness results are reported separately so that benchmark definitions are not mixed with the outcomes of a particular certification run. The complete codebase, experiment configurations, and per-specification results, including the learned certificate parameters, are available at this link.

D.1 Benchmark formulation

Each case study instantiates the finite-horizon stochastic closed-loop process of Sec. 2. For every tested environment–policy pair, nominal and attacked trajectories start from the same initial state and use the shared episode-randomness semantics of Definition 2. The operational attack law is instantiated separately for each benchmark using the attack method, attacked action channel, and norm budget reported in Table 2. Each realized attack is fixed before the episode, perturbs exactly K decision times, and satisfies the corresponding per-step norm bound. Thus, K controls temporal sparsity, whereas b controls the perturbation magnitude.

Because the horizons differ across tasks, we report temporal attack density as K/T in addition to the absolute cardinality K. The final grids are chosen so that comparable attack densities are represented

across environments. The formal event always uses the raw paired-degradation threshold $D_{\mathrm{traj}} \le \alpha$. For cross-task presentation we additionally report

$$\alpha_{\%} := 100\,\frac{\alpha}{S_{\mathrm{env}}}, \qquad \alpha = \frac{\alpha_{\%}}{100} S_{\mathrm{env}}, \tag{97}$$

where $S_{\mathrm{env}} > 0$ is a fixed environment-level return scale. The same S_{env} is used for every policy compared within one environment. Hence $\alpha_{\%}$ is neither a success probability nor a policy-specific percentage drop from that policy's nominal return; it is only a common reporting scale for the raw terminal-degradation threshold. The values used in the current experiments are reported in Table 2.

The operating law also fixes the source of episode randomness. In the custom control tasks, the random input enters the declared stochastic state update. For the current Hopper, HalfCheetah, and Ant campaigns, the environmental randomness follows the corresponding native reset semantics, with no artificial per-step actuation or process noise. Formal verification approximates the conditional successor expectations using fresh Monte Carlo successor batches, with the corresponding sample counts reported in Table 2.

D.2 Policy families

The formulation of Sec. 2 is agnostic to how the fixed deployed policy is obtained: π may be a learned policy or an analytically designed decision rule, provided its deployment-time execution satisfies Assumption 2. The main empirical study focuses on learned RL policies.

Specifically, we evaluate frozen policies trained with TD3 (Fujimoto et al., 2018), SAC (Haarnoja et al., 2018), and PPO (Schulman et al., 2017). The exact policy coverage, actor architectures, and deployment conventions are collected in Table 2, while the resulting complete robustness profiles are shown in Fig. 6. Each checkpoint is trained before certification and then held fixed throughout attack generation, paired rollout generation, barrier synthesis, calibration, and formal verification. Thus the training algorithm does not alter the certification problem: the certified object is the deployed policy itself, including its chosen deployment-time randomization mode.

For the three custom tasks, we additionally retain analytic reference policies: swing-up with local LQR stabilization for the inverted pendulum, headway regulation for ACC, and waypoint/pure-pursuit feedback for the unicycle. These are supplementary instances of the same fixed-policy formulation, not competing certification methods. All main-text empirical results report RL policies only.

D.3 Case studies

D.3.1 Inverted pendulum

The inverted-pendulum task captures nonlinear swing-up and stabilization under bounded torque and stochastic angular-velocity disturbances. With physical state $x = (\theta, \omega)$, the implementation uses the benchmark semi-implicit discretization with $\Delta t = 0.05$ and an additive Gaussian term on the angular-velocity update. The RL policy observes $(\cos\theta, \sin\theta, \omega)$ and acts through a scalar torque in $[-2, 2]$. The initial state is sampled near the downward configuration, $\theta_0 \in \pi \pm 0.1$ and $\omega_0 \in [-0.1, 0.1]$. The per-step reward penalizes wrapped-angle error, angular velocity, and control effort,

$$r_t = -\big(\bar{\theta}_t^2 + 0.1\omega_t^2 + 10^{-3}u_t^2\big), \qquad \bar{\theta} := \mathrm{wrap}(\theta). \tag{98}$$

This task is deliberately nonlinear over the full angular domain: a bounded action perturbation applied during swing-up can affect the subsequent closed-loop transient long after the attacked time step.

D.3.2 ADAPTIVE CRUISE CONTROL

ACC models longitudinal car following with state $x = (d, v_E, v_L)$, consisting of inter-vehicle distance, ego speed, and lead speed. With $\Delta t = 0.1$,

$$v_{t+1}^L = v_t^L + \Delta t\, a_t^L + \sigma \xi_t, \tag{99}$$

$$v_{t+1}^E = v_t^E + \Delta t\, u_t, \tag{100}$$

$$d_{t+1} = d_t + \Delta t\, (v_t^L - v_t^E), \tag{101}$$

where the lead-vehicle stochastic input is shared between the nominal and attacked branches. The ego action is bounded in $[-3, 2]$ and the state is clipped to the declared speed/headway box. Initial speeds satisfy $v_E(0) = v_L(0) \sim \mathcal{U}[20, 25]$ with the initial headway placed at the desired constant-time-headway reference. Writing $e = d - (5 + 1.4 v_E)$ and $\dot{e} = v_L - v_E$, the reward is

$$r_t = -(e/10)^2 - 0.1(\dot{e}/5)^2 - 0.02(u/3)^2. \tag{102}$$

The case study combines a physically interpretable stochastic disturbance with a direct action-space attack on the throttle/brake command.

D.3.3 UNICYCLE WAYPOINT TRACKING

The unicycle task represents planar goal-directed motion with physical state $x = (p_x, p_y, \theta)$ and two control channels, forward speed and turn rate. The discrete-time model is

$$p_{x,t+1} = p_{x,t} + \Delta t\, v_t \cos\theta_t, \tag{103}$$

$$p_{y,t+1} = p_{y,t} + \Delta t\, v_t \sin\theta_t, \tag{104}$$

$$\theta_{t+1} = \mathrm{wrap}(\theta_t + \Delta t\, \omega_t + \sigma \xi_t), \tag{105}$$

with $\Delta t = 0.1$ and stochastic heading evolution. The action satisfies $v \in [0, 1]$ and $\omega \in [-1.5, 1.5]$. Initial positions are sampled from an annulus of radius $[2, 4]$ around the goal with a random initial heading. The reward combines squared distance-to-goal with normalized forward-speed and turn-rate penalties. Because the attacked channels have heterogeneous physical units, the fixed channel scaling used to map the scalar budget b to (v, ω) perturbations is part of the experimental specification.

D.3.4 HOPPER-V5

Hopper-v5 is a MuJoCo planar locomotion task with intermittent ground contact (Towers et al., 2024; Todorov et al., 2012). Its generalized dynamics can be written schematically as

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + g(q) = Bu + J(q)^\top \lambda, \tag{106}$$

where the simulator resolves the contact forces λ. The deployed policy observes the standard 11-dimensional observation and controls three bounded joint-torque channels. The retained runs use the native Hopper-v5 reset randomness with reset-noise scale 0.005, and no artificial per-step transition or actuation noise is added. The default unhealthy-state termination is retained. This task tests certification under hybrid contact transitions, where a sparse action perturbation can change both the accumulated return and the subsequent contact sequence.

D.3.5 HALFCHEETAH-V5

HalfCheetah-v5 is a higher-dimensional planar MuJoCo locomotion task with six coupled torque inputs (Towers et al., 2024; Todorov et al., 2012). As in Hopper, its multibody dynamics include state-dependent inertial coupling and ground-contact forces. The policy uses the standard 17-dimensional observation and actions are clipped to $[-1, 1]^6$. The retained campaign uses the native HalfCheetah-v5 reset randomness (reset-noise scale 0.1) and no artificial per-step actuation or process noise. Small action perturbations can shift the phase and coordination of the learned gait over the finite horizon, which makes the paired return degradation sensitive to both attack magnitude and attack timing.

D.3.6 ANT-V5

Ant-v5 is a three-dimensional quadrupedal MuJoCo locomotion task with eight actuated joints (Towers et al., 2024; Todorov et al., 2012). The current policy interface uses the standard 105-dimensional Ant-v5 observation and an 8-dimensional action in $[-1, 1]^8$. The default simulator control interval is 0.05 s, the retained horizon is $T = 1000$, and the environment uses the native Ant reset/health semantics of the final run. The reward combines forward motion and healthy-state reward with control and contact penalties. The task provides the highest-dimensional locomotion setting in the current suite and therefore probes the scalability of the learned certificate beyond the planar MuJoCo systems.

D.4 CERTIFICATE PARAMETERIZATION

All experiments use the bounded neural barrier parameterization introduced in Sec. 4.1. The barrier is a scalar structured MLP on features of the augmented nominal–attacked state, degradation coordinate, and time, with a sigmoid output bounded by $B_{\max}$. Its learned constants b_1^*, b_2^*, b_3^* and κ^* retain the meanings in Definition 6: they summarize the initial-set bound, terminal separation, stochastic one-step drift, and multiplicative propagation factor. The accumulated finite-horizon contribution uses $\mathcal{G}_T(\kappa) = \sum_{j=0}^{T-1} \kappa^j$ as in Definition 6, rather than replacing the geometric sum by the looser factor T.

The hidden layers use SiLU activations in all retained certificates. This is an optimization choice, not a theoretical restriction. During synthesis, the conditional drift term is estimated using empirical averages over M_s fresh successor samples. Formal verification likewise uses empirical averages over M fresh successor samples together with the finite-sample correction of Lemma 1. The values of N_s, M_s, N, and M used in each case study are collected in Table 2.

For the synthesis sets in equation 6, if $U = \varnothing$, its loss contribution is defined as zero. This provides no evidence of robustness, since the resulting candidate remains subject to independent verification. The safe pool S is retained for supplementary analysis and does not enter the loss in equation 7. If auxiliary unsafe examples are used during synthesis, their full augmented trajectories are retained so that intermediate state–attack–time tuples may also contribute to the sampled drift condition; such data affect candidate construction only and do not themselves establish a formal guarantee.

D.5 FIXED EXPERIMENTAL SETTINGS

Table 2 provides the complete experimental specification for the six retained case studies, including environment and policy settings, attack methods and norm budgets, attack and degradation-tolerance grids, normalization constants, certificate parameterization, and synthesis/verification sample budgets. Unless stated otherwise, all formal FoSeRL experiments use target reliability $\rho = 0.90$ and confidence parameters $\beta = \beta_0 = 0.001$.

Computational setup. All FoSeRL experiments reported in this paper were run on a workstation with an Intel Xeon w5-3433 CPU, 128 GB DDR5-4400 RAM, and an NVIDIA RTX 4500 Ada Generation GPU with 24 GB GDDR6 ECC memory. Reported runtimes are wall-clock certification times and exclude policy training.

D.6 CERTIFICATE CONSTRUCTION AND REPORTING WORKFLOW

This subsection summarizes how each tested attack specification is converted into a reported FoSeRL certificate. The formal definitions and guarantees are given in Sec. 4.2; here we describe the experimental workflow used to produce the robustness profiles in Fig. 6.

From a tested specification to a certified point. Fix an environment–policy pair and a specification (K, b, α, ρ). Candidate construction first learns the bounded time-dependent barrier B_{ϕ^*} and the associated constants b_1^*, b_2^*, b_3^* and κ^* using the synthesis procedure of Sec. 4.1. The candidate and all verification-design choices are then frozen. Independent calibration episodes are drawn under the operational law $\mathbb{Q}_{K,b}$, and each episode is assigned the score S_i of Definition 9. For every finite conformal rank k, the score determines the residual allowance $r_k = \max\{0, S_{(k)}\}$ and the

Table 2: Complete experimental specification for the six FoSeRL case studies. For each environment, we report the environment and policy configuration, attack specification, degradation-tolerance grid, certificate architecture, and synthesis/verification sample budgets. N_s and M_s denote the synthesis episodes and successor samples, while N and M denote the corresponding formal-verification sample counts. All formal runs target $\rho = 0.90$ with $\beta = \beta_0 = 0.001$.

	Parameter	Inv. Pend.	ACC	Unicycle	Hopper	HalfCheetah	Ant
Environment	Model	Custom	Custom	Custom	Gym v5	Gym v5	Gym v5
	n_{obs}	3	3	3	11	17	105
	n_x	2	3	3	12	18	29
	n_a	1	1	2	3	6	8
	Δt	0.05	0.10	0.10	0.008	0.05	0.05
	T	200	300	500	1000	500	1000
	γ	0.99	0.99	0.99	0.99	0.99	0.99
	Task const.	θ wrapped, $\lvert\omega\rvert \le 8$ $g = 10,\ m = \ell = 1$	$d \in [0, 60]$, $v_E, v_L \in [0, 40]$ $d_0 = 5,\ h = 1.4$	$(p_x, p_y) \in [-5, 5]^2$ goal $(0, 0)$	Native v5	Native v5	Native v5
	Randomness	ω noise	Lead-process noise	Heading noise	Reset	Reset	Reset
	Noise scale	$\sigma = 0.01$	$\sigma = 0.30$	$\sigma = 0.05$	0.005	0.1	0.1
	Initial law	$\theta_0 \in \pi \pm 0.1$ $\omega_0 \in [-0.1, 0.1]$	$v_E(0) = v_L(0) \sim \mathcal{U}[20, 25]$	annulus $[2, 4]$ random θ_0	Native	Native	Native
	$\mathcal{A}$	$[-2, 2]$	$[-3, 2]$	$v \in [0, 1]$ $\omega \in [-1.5, 1.5]$	$[-1, 1]^3$	$[-1, 1]^6$	$[-1, 1]^8$
Policy	Policy	TD3, SAC, PPO	TD3, SAC, PPO	TD3, SAC, PPO	TD3, SAC, PPO	TD3, SAC, PPO	TD3, SAC, PPO
	Analytic ref.	Swing-up + LQR	Headway LQR	Pure pursuit	–	–	–
	Actor	$3 \to 64 \to 64 \to 1$	$3 \to 64 \to 64 \to 1$	$3 \to 64 \to 64 \to 2$	$11 \to 64 \to 64 \to 3$	$17 \to 64 \to 64 \to 6$	$105 \to 256 \to 256 \to 8$
	Activation	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
	Output	Bounded	Bounded	Bounded	Bounded	tanh / bounded	tanh / bounded
Threat Model	Channel	Torque	Ego accel.	(v, ω)	3 actuators	6 actuators	8 actuators
	Attack method	PGD / Uniform	PGD	PGD	LAS / Uniform	PGD / Uniform	DE / Uniform
	Norm budget	ℓ_∞	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞	ℓ_∞
	$\alpha_{\%}$	$\{0.5, 1, 2.5, 5\}$	$\{0.5, 1, 2.5, 5\}$	$\{0.5, 1, 2.5, 5\}$	$\{0.5, 1, 2.5, 5\}$	$\{5, 10, 15, 20\}$	$\{5, 10, 15, 20\}$
	S_{env}	453.340	6.229	22.921	174.525	682.400	336.554
Synthesis	Certificate	MLP	MLP	MLP	MLP	MLP	MLP
	Input dim.	12	18	21	39	57	92
	Widths	$48, 48$	$48, 48$	$48, 48$	$64, 64$	$96, 96, 64$	$128, 128, 96$
	Output dim.	1	1	1	1	1	1
	Activation	SiLU	SiLU	SiLU	SiLU	SiLU	SiLU
	(N_s, M_s)	$(1024, 8)$	$(1024, 8)$	$(1024, 8)$	$(1000, 10)$	$(1000, 10)$	$(2000, 40)$
Verification	ρ	0.90	0.90	0.90	0.90	0.90	0.90
	β	0.001	0.001	0.001	0.001	0.001	0.001
	β_0	0.001	0.001	0.001	0.001	0.001	0.001
	N	2000	1395–2000	1566–2000	2000	2000	3000
	M	49,152	49,152	49,152	30,000	50,000	20,000

corresponding out-of-sample violation bound $\bar{v}_k$ from equation 16. The selected rank k^* is the one maximizing the resulting end-to-end reliability lower bound, as in equation 17.

Reliability decomposition. For reporting and diagnosis, we separate the reliability bound into three terms:

$$\begin{aligned} \mathrm{Base} &:= \frac{b_1^*(\kappa^*)^T + b_3^*\,\mathcal{G}_T(\kappa^*)}{b_2^*}, \\ \mathrm{Margin}_k &:= r_k \frac{(\kappa^*)^T + \mathcal{G}_T(\kappa^*)}{b_2^*}, \\ \mathrm{Bridge}_k &:= \frac{\bar{v}_k}{1-\beta_0}. \end{aligned} \tag{107}$$

The corresponding certified reliability at rank k is

$$\rho_{\mathrm{cert}}(k) = 1 - \Big(\mathrm{Base} + \mathrm{Margin}_k + \mathrm{Bridge}_k\Big). \tag{108}$$

Hence k^* equivalently maximizes $\rho_{\mathrm{cert}}(k)$ over the finite ranks, and the tested magnitude b is accepted exactly when $\rho_{\mathrm{cert}}(k^*) \geq \rho$, which is equivalent to the acceptance condition $\bar{q}_4 \leq 0$ in equation 18. Base is the contribution of the frozen barrier itself, Margin is the propagated calibration slack, and Bridge accounts for violations of the exact adjusted barrier conditions that remain after finite-sample calibration. These are components of a certified lower bound, not empirical terminal-failure frequencies.

Finite-successor correction. When the conditional expectation in the drift condition is estimated from M fresh successors, the correction ϵ_M is chosen to satisfy Lemma 1; the tight Hoeffding choice at fixed M is $\epsilon_M = B_{\max}\sqrt{\log(1/\beta_0)/(2M)}$. This correction enters the empirical drift residual and therefore affects the calibration scores and the selected Margin. It is not subtracted again from ρ_{cert}.

From certified points to robustness profiles. For each fixed (K, α), attack magnitudes are explored over the prescribed experimental grid. The quantity b_{emp} is used only as a pre-verification design reference for candidate construction and search; it is not a formal certificate. A tested magnitude b contributes to the certified frontier only after the frozen candidate passes the independent verification procedure above. Thus, the reported $b_{\mathrm{cert}}^{\mathrm{op}}$ is an accepted tested magnitude under the operational criterion of Sec. 4, not an interpolated robustness radius.

Repeating this procedure across the tested values of α and K produces the paired empirical and certified profiles in Fig. 6. Each displayed certified point has its own per-specification confidence statement from Theorem 2; connecting lines are visual guides only and do not certify intermediate budgets or imply simultaneous validity across the displayed collection. Adaptive or collection-level reporting follows the confidence-control procedures in Appendix B.7.

Practical certificate design. The decomposition in equation 108 provides a direct way to diagnose a weak certificate. A large Base points to the synthesized witness itself: the propagated initial and drift contributions are large relative to terminal separation. A large Margin indicates that the calibration scores require substantial residual slack. A large Bridge indicates limited episode-level statistical resolution. These effects should be treated separately rather than increasing every synthesis and verification budget indiscriminately.

Long horizons require particular care because both the learned drift term and the calibrated residual slack propagate through $\mathcal{G}_T(\kappa)$. When $\kappa = 1$, $\mathcal{G}_T(\kappa) = T$, so even a small residual allowance can become dominant for large T. For long-horizon problems, selecting or fixing $\kappa < 1$ during candidate design can substantially reduce this propagation: $\mathcal{G}_T(\kappa)$ is then bounded by $(1-\kappa)^{-1}$ and κ^T decays with the horizon. This is nevertheless a design trade-off rather than a rule that smaller κ is always better, since decreasing κ also tightens the one-step drift condition and may increase b_3^* or the resulting residuals. Any such choice must be made before independent calibration and then kept fixed during verification.

The verification sample sizes play different roles. Increasing the number of calibration episodes N improves the resolution of the conformal violation bound and can reduce Bridge, whereas increasing

the successor count M reduces ϵ_M and can improve the drift residuals that determine Margin. The barrier output bound $B_{\max}$ also enters this trade-off through the finite-successor requirement: an unnecessarily large $B_{\max}$ increases the required successor budget, while a value that is too small can restrict the witness class. In practice, certificate refinement is therefore most efficient when the limiting contribution in equation 108 is identified first and the corresponding synthesis or verification component is improved.

D.7 Complete FoSeRL certification results

Figure 6 reports the complete robustness profiles for all retained RL-policy settings. Each environment–policy specification is shown as a paired set of panels: the upper panel reports the formally certified FoSeRL frontier, while the lower panel shows the corresponding empirical pre-verification reference. Within each panel, curves correspond to the tested temporal attack densities K/T, the horizontal axis gives the attack magnitude b, and the vertical axis gives the normalized degradation tolerance $\alpha_{\%}$. All formal FoSeRL results use the target reliability $\rho = 0.90$ and per-specification confidence parameter $\beta = 0.001$. Across environments and policy families, relaxing the degradation tolerance generally permits larger attack magnitudes, whereas increasing the temporal attack density typically contracts the admissible budget. The empirical and certified profiles exhibit the same broad structure, but the gap between them varies substantially across settings. The aggregate view in Fig. 5 shows that this retention remains broadly stable across the low-to-high density regimes, with a larger reduction appearing in the very-high-density bin. In some cases the formal frontier closely tracks the empirical reference, while in others verification produces a visibly more conservative budget. This variation reflects the cost of converting a design-stage empirical frontier into an independently verified certificate and also shows that the conservatism of the certification stage is environment– and policy-dependent rather than uniform across tasks. Each marker corresponds to an individually tested specification; connecting lines are included only as visual guides and do not imply certification of intermediate or untested budgets.

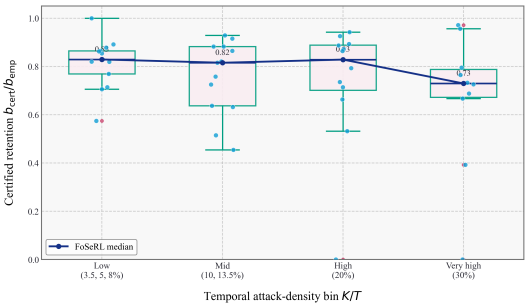


Figure 5: FoSeRL certified-budget retention $b_{\mathrm{cert}}/b_{\mathrm{emp}}$ across temporal attack-density bins. Points denote individual reported specifications and the blue line connects bin medians.

D.8 Comparison with policy smoothing

Table 3 reports the detailed external-baseline comparison, while Fig. 7 summarizes how often each method returns a nonzero certificate across the reported specifications. Since no existing formal method matches our complete setting, we align policy smoothing (PS) with FoSeRL on the action channel, paired degradation event, and reliability/confidence targets, while retaining the native ℓ_2 perturbation geometry of the smoothing certificate. The PS guarantee is worst-case over the corresponding ℓ_2 perturbation set and therefore does not depend on a particular attack optimizer; uniform precommitted perturbations are used only for the associated empirical rollout calculations. Full benchmark-specific attack and norm specifications are given in Table 2, and the broader positioning of PS relative to FoSeRL is discussed in Appendix F. Comparable PS certificates are obtained for Inverted Pendulum, Hopper, and HalfCheetah. ACC is excluded because the available smoothing-noise channel does not coincide with the attacked ego-control channel, while Unicycle does not meet the required nominal-reliability condition at the tested tolerance. For Ant, PS yields no nonzero certified bud-

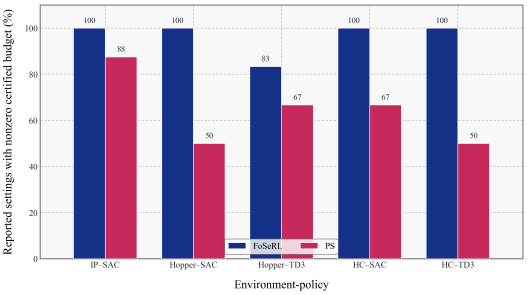


Figure 7: Fraction of reported settings with a nonzero certified budget for FoSeRL and policy smoothing (PS), grouped by environment–policy pair.

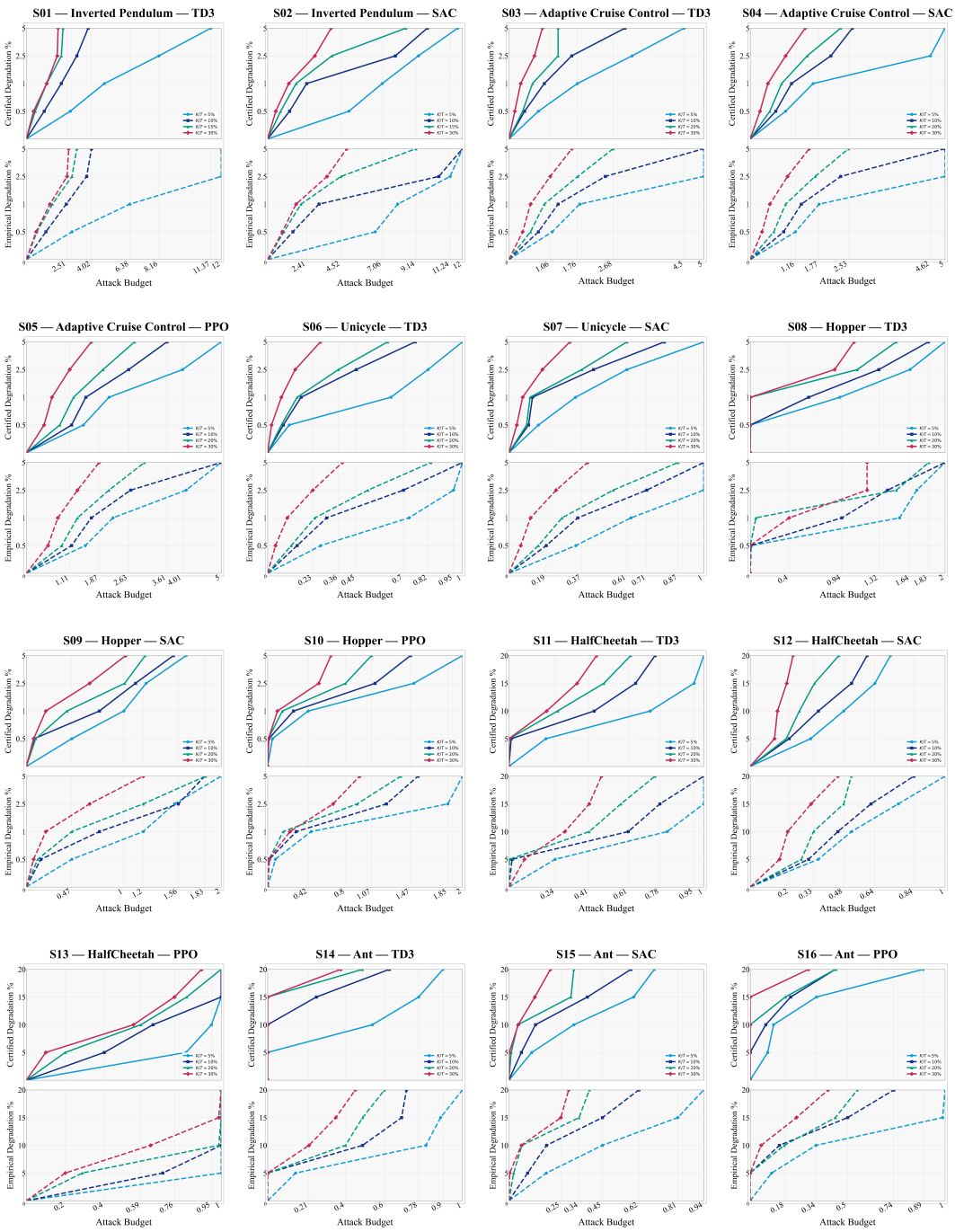


Figure 6: Complete empirical and certified robustness profiles for all retained RL-policy experiments. Each environment–policy setting is shown as a paired panel: the upper plot gives the FoSeRL-certified frontier and the lower plot the corresponding empirical pre-verification reference. Curves represent the tested temporal attack densities K/T; the horizontal axis is the attack magnitude b, and the vertical axis is the normalized degradation tolerance $\alpha_{\%}$. Formal FoSeRL certificates target reliability $\rho = 0.90$ with per-specification confidence parameter $\beta = 0.001$. Across settings, larger degradation tolerances generally support larger attack budgets, while denser temporal attacks tend to reduce the admissible magnitude. The gap between the empirical and certified frontiers varies across environments and policies, revealing setting-dependent verification conservatism. Markers denote individually tested specifications; connecting lines are visual guides only and do not certify interpolation or untested budgets.

get at the tested settings. Across the comparable cases, FoSeRL consistently certifies substantially larger budgets, including several specifications for which PS certifies only the zero budget.

Table 3: Detailed FoSeRL–policy-smoothing (PS) comparison on aligned environment–policy settings. For each setting, we report the search limit $b_{\max}$, attack cardinality K and density K/T, normalized degradation tolerance $\alpha_{\%}$, empirical design reference b_{emp}, and the certified budgets returned by FoSeRL and PS. The normalized gap is $\left|b_{\text{cert}}^{\text{FoSeRL}} - b_{\text{cert}}^{\text{PS}}\right| / b_{\max}$. A zero PS entry indicates that no nonzero budget was certified at that specification.

Environment	Policy	$b_{\max}$	K	K/T	$\alpha_{\%}$	b_{emp}	b_{cert}		Norm. gap
							FoSeRL	PS	
Inverted Pendulum ($T = 200$)	SAC	4	7	3.5%	5%	4.00	4.00	0.0198	0.9951
			16	8%	2.5%	2.60	2.00	0.0129	0.4968
			16	8%	5%	3.47	3.00	0.0132	0.7467
			27	13.5%	1%	1.02	**0.65**	**0.0000**	0.1625
			27	13.5%	2.5%	1.84	1.50	0.0101	0.3725
			27	13.5%	5%	3.47	3.00	0.0101	0.7475
			40	20%	2.5%	1.72	1.50	0.0081	0.3730
			40	20%	5%	2.80	2.50	0.0083	0.6229
Hopper ($T = 1000$)	SAC	2	50	5%	1%	1.40	**1.00**	**0.0000**	0.5000
			50	5%	5%	1.84	1.64	0.0032	0.8184
			100	10%	1%	0.99	**0.75**	**0.0000**	0.3750
			100	10%	5%	1.84	1.51	0.0023	0.7539
			200	20%	1%	0.79	**0.42**	**0.0000**	0.2100
			200	20%	5%	1.84	1.22	0.0016	0.6092
			300	30%	1%	0.51	**0.20**	**0.0000**	0.1000
			300	30%	5%	1.05	1.02	0.0013	0.5094
	TD3	2.5	50	5%	1%	1.11	**0.92**	**0.0000**	0.3680
			50	5%	2.5%	1.92	1.64	0.0020	0.6552
			50	5%	5%	2.44	2.00	0.0032	0.7987
			100	10%	1%	0.95	**0.60**	**0.0000**	0.2400
			100	10%	2.5%	1.82	1.32	0.0014	0.5274
			100	10%	5%	1.97	1.83	0.0023	0.7311
			200	20%	1%	0.95	0.00	**0.0000**	0.0000
			200	20%	2.5%	1.54	1.10	0.0010	0.4396
			200	20%	5%	1.62	1.50	0.0016	0.5994
			300	30%	1%	0.91	0.00	**0.0000**	0.0000
			300	30%	2.5%	0.91	0.87	0.0008	0.3477
			300	30%	5%	1.46	1.07	0.0013	0.4275
HalfCheetah ($T = 500$)	SAC	1	25	5%	5%	0.54	**0.31**	**0.0000**	0.3100
			25	5%	10%	0.68	0.48	0.0024	0.4776
			25	5%	20%	0.82	0.72	0.0032	0.7168
			50	10%	5%	0.44	**0.20**	**0.0000**	0.2000
			50	10%	10%	0.68	0.35	0.0017	0.3483
			50	10%	20%	0.68	0.60	0.0023	0.5977
			100	20%	5%	0.22	**0.19**	**0.0000**	0.1900
			100	20%	10%	0.34	0.25	0.0012	0.2488
			100	20%	20%	0.58	0.46	0.0016	0.4584
			150	30%	5%	0.17	**0.13**	**0.0000**	0.1300
			150	30%	10%	0.21	0.14	0.0010	0.1390
			150	30%	20%	0.32	0.22	0.0013	0.2187
	TD3	1.5	25	5%	15%	1.16	**0.95**	**0.0000**	0.6333
			25	5%	20%	1.16	1.00	0.0007	0.6662
			50	10%	15%	0.71	**0.65**	**0.0000**	0.4333
			50	10%	20%	0.85	0.75	0.0005	0.4997
			100	20%	15%	0.52	**0.49**	**0.0000**	0.3267
			100	20%	20%	0.71	0.63	0.0004	0.4197
			150	30%	15%	0.44	**0.35**	**0.0000**	0.2333
			150	30%	20%	0.62	0.45	0.0003	0.2998

Qualitative comparison. The quantitative separation in Table 3 is also visible in representative closed-loop rollouts. Figures 8 and 9 show nominal executions, executions at FoSeRL-certified budgets, and attacks beyond those certified budgets for Inverted Pendulum, Hopper, and HalfCheetah. At the reported FoSeRL budgets, the illustrated degradation traces remain below the prescribed terminal tolerance, whereas the larger uncertified perturbations produce substantially greater degradation and, in the Hopper example, early collapse. The corresponding PS certificates remain confined to much smaller perturbation magnitudes, consistent with the quantitative comparison above.

Scope of the comparison. These visualizations should be interpreted together with the differing certificate scopes. Policy smoothing provides a worst-case guarantee over its corresponding ℓ_2 perturbation set, whereas the data-driven FoSeRL certificate is distributional under the declared attack law. The beyond-certificate panels are deliberately selected from episodes that violate the degrada-

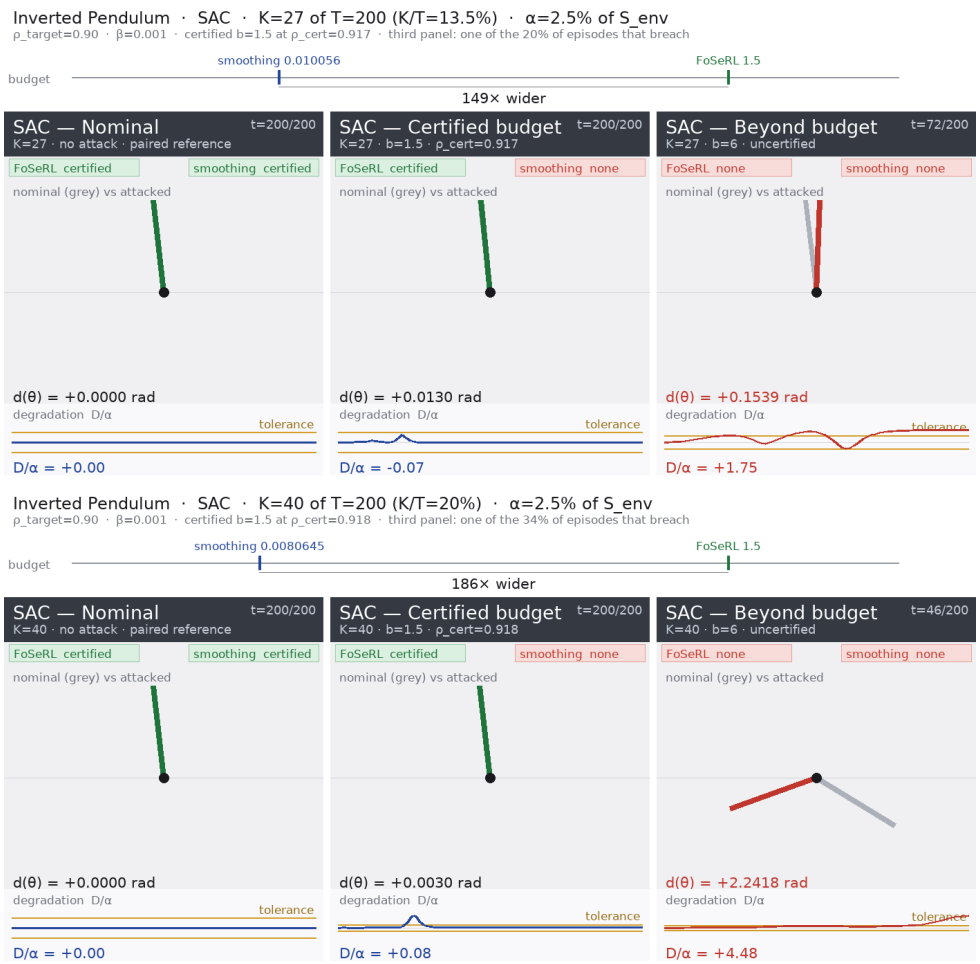
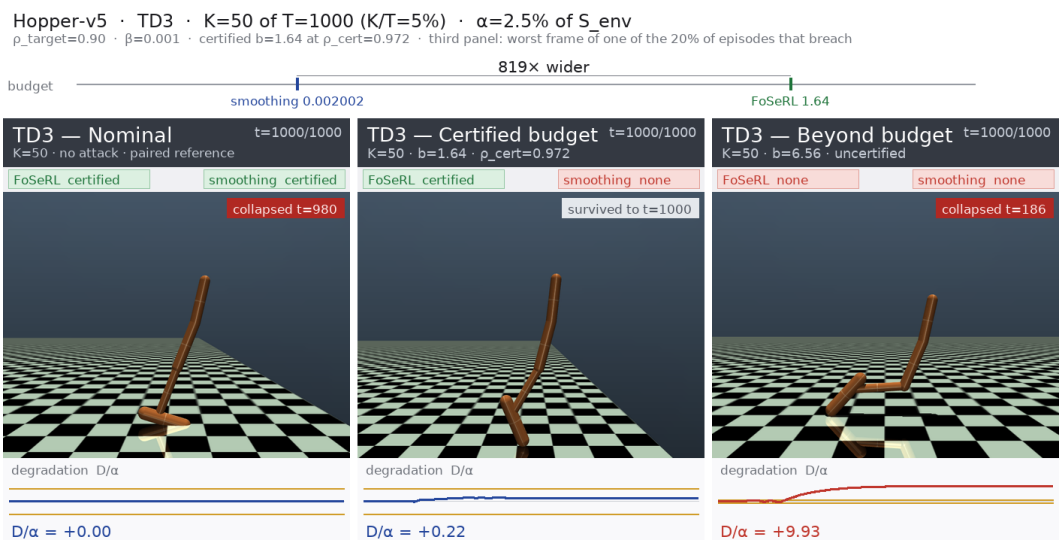
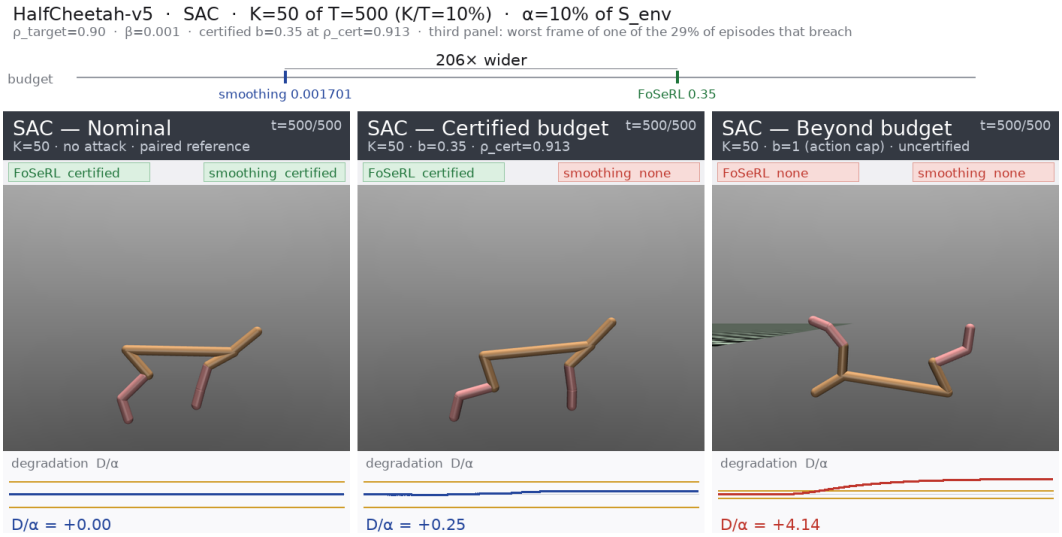


Figure 8: Inverted pendulum (SAC), at $K = 27$ (top block) and $K = 40$ (bottom), $\alpha = 2.5\%$ of S_{env}, $\rho_{\text{target}} = 0.90$, $\beta = 10^{-3}$. Each block shows the nominal rollout, the rollout at FoSeRL's certified budget, and a rollout beyond it, with the attacked pole in colour against the nominal in grey. The budget bar contrasts the two certificates; note the scopes differ, smoothing being worst-case over $\|\delta\|_2 \leq b\sqrt{n_a K}$ and FoSeRL distributional under the declared attack law. D/α is signed, so a marginally negative value (-0.07, top middle) indicates an attack realisation that slightly helped that episode. The third panel of each block is one of the 20% ($K = 27$) and 34% ($K = 40$) of episodes that breach beyond the certified budget; breaching outside the certified region is the certificate behaving as specified.

tion tolerance outside the certified region; they illustrate the behavior beyond the guarantee rather than a failure of the certificate.



(a) Hopper-v5 (TD3), $K = 50$ of $T = 1000$, $\alpha = 2.5\%$; certified $b = 1.64$ at $\rho_{\text{cert}} = 0.972$, against a smoothing budget of 2.0×10^{-3}.



(b) HalfCheetah-v5 (SAC), $K = 50$ of $T = 500$, $\alpha = 10\%$; certified $b = 0.35$ at $\rho_{\text{cert}} = 0.913$, against a smoothing budget of 1.7×10^{-3}.

Figure 9: Certified and uncertified regimes on MuJoCo locomotion, the setting in which neither Kumar et al. (2022) nor Mu et al. (2024) report results. Panels show the nominal rollout, the rollout at FoSeRL's certified budget, and a rollout beyond it, with the running degradation D/α beneath each. Because the two branches share the noise realisation but diverge after the first perturbation, their termination times differ: in a the nominal episode terminates at $t = 980$ while the attacked one at the certified budget survives to the horizon, and beyond the budget it collapses at $t = 186$. Third panels are drawn from the 20% (Hopper) and 29% (HalfCheetah) of episodes that breach outside the certified region.

E Additional Discussion

This section expands on Sec. 6, explaining the long-horizon verification design, the information supplied by the barrier witness, and the interpretation and refinement of the resulting certificates. Technical boundary cases and sampling requirements are collected in Appendix B; literature positioning and external-baseline selection are discussed separately in Appendix F.

E.1 From training-time robustness to long-horizon certification

The framework of Taheri & Zamani (2026) models stochastic training as a dynamical system and certifies the accuracy of its terminal parameters under data perturbations, already incorporating a training horizon and a time-dependent barrier. FoSeRL changes both the intervention and the certified object: the policy is already fixed, perturbations affect its executed actions, and performance is measured by the accumulated loss relative to a matched nominal episode. The paired-state construction and return-gap coordinate make this sequential objective a terminal safety property, while the separate cardinality and magnitude budgets capture the temporal allocation of attack resources. Long horizons also require controlling the accumulation of local drift allowances. With the earlier linear term $b_3 T$, a verification correction η_3 contributes $\eta_3 T/b_2$ to the failure bound; for example, $\eta_3/b_2 = 10^{-3}$ already contributes one at $T = 1000$, exceeding the failure allowance $1 - \rho = 0.1$. FoSeRL retains the geometric accumulation $\mathcal{G}_T(\kappa)$ and calibrates the residual allowance using the final reliability bound. For $\kappa < 1$, $\mathcal{G}_T(\kappa) \le \min\{T, (1-\kappa)^{-1}\}$, avoiding unnecessary linear accumulation without removing the horizon dependence of the problem (Appendix B.6). Nor can T simply be replaced by K: an altered state can affect later decisions and rewards after the last attacked action.

E.2 Episode-level calibration and statistical validity

FoSeRL calibrates complete episodes rather than treating their dependent transitions as independent observations. The score in Definition 9 aggregates initial and drift residuals while retaining terminal separation as a hard condition. All calibration episodes remain in the sample, including those assigned infinite scores because of unsafe-terminal separation violations. No parametric score distribution is required; validity instead relies on independent episodes and fresh conditional successor samples. Calibration balances the propagated residual margin against the probability of episode-level condition violations. Simultaneous validity over ranks permits selecting this balance from the calibration data without an additional rank-selection penalty. The finite-successor bound then connects the sampled drift residuals to their exact counterparts: conditional on a violating episode, its first exact drift violation is detected with probability at least $1 - \beta_0$. This yields the transfer penalty $\bar{v}_{k^*}/(1-\beta_0)$ without a union bound over time. Combined with the good-prefix barrier argument, these steps establish the end-to-end operational guarantee of Theorem 2, rather than only marginal coverage of a score.

E.3 Reliability, confidence, and violation levels

The target ρ concerns a fresh attack–episode draw: it is the required probability of satisfying $D_{\mathrm{traj}} \le \alpha$ under $\mathbb{Q}_{K,b}$. The reported ρ_{cert} is the lower bound in equation 108, and acceptance requires $\rho_{\mathrm{cert}} \ge \rho$. In contrast, β controls uncertainty over calibration data: conditional on the frozen specification, Theorem 2 bounds false acceptance by β. The auxiliary β_0 controls the one-sided successor approximation at a fixed state–attack–time tuple through Lemma 1. Its effect is incorporated into the transfer penalty, not deducted again from the outer confidence, which remains $1 - \beta$. Finally, $\bar{v}_{k^*}$ bounds out-of-sample violations of the selected empirical episode conditions; it is neither an empirical failure frequency nor the terminal-degradation risk itself. The latter also includes the dynamical contribution, as made explicit by the Base–Margin–Bridge decomposition in equation 107–equation 108.

E.4 Barrier witnesses versus direct terminal certification

FoSeRL provides a dynamical witness in addition to a scalar robustness guarantee. Direct terminal certification applies a one-sided binomial bound to independent observations of the same paired terminal event under the same operational specification. It is a useful formal reference when the required output is only the terminal probability, and can yield larger certified budgets with less computation because it does not require barrier synthesis or conditional successor queries. FoSeRL's additional contribution is the structured certificate supporting that probability and the analysis it enables.

State–time structure. The witness $B_{\phi^*}(\hat{x}, t)$ links initial conditions, expected one-step evolution, and terminal separation through the conditions of Definition 6. These relations make the dynamics

part of the certification argument, rather than reducing each episode solely to a terminal success indicator. The same structure underlies the ideal uniform guarantee and its finite-sample operational counterpart, with their respective scopes stated in Theorems 1 and 2.

Condition-specific diagnosis and refinement. The residuals in Definition 8 distinguish initial-value, terminal-separation, and drift violations, including the sampled phases at which drift conditions are difficult to satisfy. The reliability decomposition further separates the frozen witness contribution from calibration slack and episode-level violations. This provides concrete directions for refinement: concentrated late-horizon drift violations can motivate targeted synthesis data or time-dependent features, whereas poor terminal separation calls for improving the terminal witness. The scalar binomial certificate alone does not provide this condition-based decomposition. These diagnostics concern the witness, not certified conditional failure probabilities at arbitrary states.

In Table 1, FoSeRL retains approximately 71–99% of Direct's certified magnitude across the representative settings while additionally returning this state–time witness. Thus, the comparison is not between formal and empirical validity: both methods provide formal operational guarantees. It is a trade-off between scalar budget tightness and computation on one side, and a dynamically structured certificate with targeted diagnostic value on the other. When the purpose includes understanding and refining the certifying conditions, the barrier provides information beyond the terminal bound itself.

E.5 Generality, simulator access, and attack coverage

FoSeRL separates policy evaluation from policy construction. Under the stated execution and sampling assumptions, its verification argument does not require a particular learning algorithm, policy architecture, differentiable controller, or analytical dynamics family. The deployed policy remains unchanged, including its chosen randomization mode; the learned barrier is an auxiliary witness, not a replacement controller. This generality uses a concrete interface: synchronously coupled rollouts and repeatable conditional successor queries, with fresh complete one-step randomness regenerated as specified in Sec. 4. The resulting guarantee concerns the execution process being sampled. The distinction from offline RL and evaluation from single-branch logs is discussed in Appendix F. The operational attack law need not be uniform. It may emphasize difficult precommitted plans or design-selected mixtures, provided it is fixed before calibration and attacks remain independent of the initial state and episode randomness as required by Definition 7. Theorem 2 certifies reliability under this law; uniform coverage of all admissible precommitted attacks instead follows from the ideal barrier conditions. State-adaptive adversaries require an extension of the present attack model.

E.6 Reading robustness frontiers and selecting specifications

The robustness profiles expose how attack frequency and magnitude interact with the allowed return loss. Their interpretation retains the declared policy execution, attack channel and norm, horizon, coupling, and statistical target. The PS comparison in Appendix D.8 also retains the distinction between operational and worst-case attack coverage. Normalization by S_{env} supplies a common reporting scale within an environment, not a percentage of each policy's nominal return or a universal ranking across tasks. The empirical design budget supports candidate construction, whereas an accepted operational budget is supported by independent verification. Neither is asserted to be the optimal robustness radius. The ideal frontier has the nesting properties of Appendix B.3; operational certificates do not automatically inherit them when changing the budget changes the attack law. Connecting lines therefore do not certify interpolation or untested specifications. New specifications can nevertheless be added incrementally without predeclaring the entire future experiment family. Each frozen specification receives its own audit; selecting a budget after exploration requires fresh final calibration or appropriate selection-aware confidence control. Appendix B.7 gives the finite-collection and sequential protocols, including accounting for repeated tests and stopping.

E.7 Performance robustness versus pathwise safety

Paired degradation complements nominal performance by measuring the loss relative to the same episode without an attack. Shared randomness does not force equal actions or states after a perturbation: each branch follows its own closed-loop evolution, preserving the attack's subsequent consequences. The running gap can increase or recover because individual reward differences can

have either sign. Accordingly, crossing the tolerance before T is not itself terminal failure, and the maximum in the verification score aggregates residuals rather than return loss. This formulation certifies cumulative performance without imposing an additional pathwise constraint. Collision avoidance or all-time state safety can be studied alongside it, but requires an explicit safety event and corresponding verification conditions. A safety-violation indicator in the augmented state offers one route for such an extension.

E.8 Certificate refinement and policy improvement

Separating construction from verification allows the witness to improve without changing the meaning of acceptance. Richer features, alternative bounded architectures, and targeted synthesis data may improve candidate quality; the synthesis law may emphasize difficult cases rather than match the operational law. Each candidate and verification design must then be frozen before fresh independent calibration. The decomposition in Appendix D.6 also supports targeted resource allocation. The episode count N controls calibration resolution, whereas the successor count M controls the approximation correction. The choice of κ balances geometric propagation against the strictness of the local drift inequality. These are distinct design levers, not interchangeable ways to improve a certificate (Appendix B.6). Beyond witness refinement, residual diagnostics offer a direction for certificate-informed policy improvement, with each updated policy treated as a new target and independently re-certified.

E.9 Limitations and outlook

FoSeRL's simulation-based interface supports broad applicability but requires repeated coupled rollouts and conditional successor queries. Long horizons and expensive policy or simulator evaluations can therefore increase computational cost, as quantified in Appendix B.6. The reported case studies use the original execution process without surrogate-assisted synthesis. For larger-scale applications, lightweight surrogates could reduce candidate-construction cost by generating synthesis rollouts. The candidate would still be frozen and formally verified using fresh paired episodes and conditional successors from the original execution process. Provided the verification assumptions and confidence protocol are retained, surrogate accuracy would affect candidate quality and certifiability, not the stated validity of an accepted certificate. Such acceleration would concern synthesis; formal verification would retain its original-process sampling requirements. The sufficient barrier conditions also trade some budget tightness for a state–time witness; their conservatism is discussed in Appendix B.4. These computational and tightness considerations accompany the additional certificate structure, while policy and environment generality hold within the execution, sampling, and attack-coverage scope specified in Appendix E.5. The separation between synthesis and verification provides a route to improving efficiency and witness quality while retaining independent formal validation.

F Positioning, Baseline Selection, and Extended Comparisons

Here we specify what FoSeRL certifies and how that object relates to existing guarantees. We first place the relevant literature on the axes that distinguish certified quantities from one another, and separate the methods that share our problem formulation from those that solve a different one (Sec. F). We then state the criteria a method must meet to serve as a quantitative baseline here, apply them, and explain why exactly one external method qualifies (Sec. F.2). Finally, we record the regime in which that baseline has previously been evaluated, which differs from the higher-dimensional, longer-horizon settings considered in our experiments (Sec. F.3).

F.1 Positioning

Here we illustrate, using Table 4 the exact characterization of our problem and what it certifies.

Relation to offline reinforcement learning. FoSeRL shares surface features with offline RL (Levine et al., 2020): the policy is fixed and no improvement is performed. The statistical problem differs, however. Offline RL evaluates or improves a target policy from trajectories logged by a potentially different behaviour policy, which makes distribution shift, importance weighting, and

Method	Formal	Channel	Budget	Fixed π	Level	Form	Paired	Resolved
Attacks on deployed policies								
FGSM on policies (Huang et al., 2017)	✗	obs	ℓ_p	✓	ep	—	✗	✗
Strategically-timed (Lin et al., 2017)	✗	obs	(K, b)	✓	ep	—	✗	✗
Critical-point (Sun et al., 2020)	✗	obs	(K, b)	✓	ep	—	✗	✗
MAS action attack (Lee et al., 2020)	✗	act	(K, b)	✓	ep	—	✗	✗
Robust training: hardens a policy, certifies none								
RARL (Pinto et al., 2017)	✗	dyn	set	✗	ep	—	✗	✗
NR-/PR-MDP (Tessler et al., 2019)	✗	act	ℓ_p	✗	ep	—	✗	✗
SA-MDP (Zhang et al., 2020)	○	obs	ℓ_p	✗	ep	mean	✗	✗
ATLA (Zhang et al., 2021)	✗	obs	ℓ_p	✗	ep	—	✗	✗
WocaR (Liang et al., 2022)	○	obs	ℓ_p	✗	ep	mean	✗	✗
RADIAL (Oikarinen et al., 2021)	○	obs	ℓ_p	✗	st	wc	✗	✗
Certified robustness for RL								
CARRL (Everett et al., 2021)	✓	obs	ℓ_p	✗	st	wc	✗	✓
Policy smoothing (Kumar et al., 2022)	✓	obs	ℓ_2	✗	ep	mean	✗	✗
CROP (Wu et al., 2022a)	✓	obs	ℓ_2	✗	ep	mean	✗	✗
COPA (Wu et al., 2022b)	✓	data	ℓ_0	✗	ep	wc	✗	✗
ReCePS (Mu et al., 2024)	✓	act	ℓ_p	✗	ep	mean	✗	✗
CAROL (Yang et al., 2024)	✓	obs	ℓ_p	✗	ep	wc	✗	✗
Guarantees without an adversary model								
Robust MDPs (Iyengar, 2005; Nilim & El Ghaoui, 2005)	✓	kern	set	✗	ep	wc	✗	✓
Offline RL / OPE (Levine et al., 2020)	✗	—	—	✓	ep	mean	✗	✗
HCOPE (Thomas et al., 2015)	✓	—	—	✓	ep	mean	✗	✗
Statistical model checking (Legay et al., 2010)	✓	—	—	✓	ep	tail	✗	✗
Barrier certificates								
Stochastic barriers (Prajna et al., 2007; Steinhardt & Tedrake, 2012)	✓	—	—	✓	ep	tail	✗	✓
Neural synthesis (Abate et al., 2021; Anand & Zamani, 2023)	✓	—	—	✓	ep	tail	✗	✓
CRABS (Luo & Ma, 2021)	✓	dyn	set	✗	ep	wc	✗	✓
Static robustness certification (parent framework)								
Poisoning certificates (Taheri & Zamani, 2026)	✓	data	ℓ_p	✓	—	tail	✓	✗
FoSeRL (this paper)	✓	**act**	$(\mathbf{K}, \mathbf{b})$	✓	ep	**tail**	✓	✓

Table 4: Related work, and where FoSeRL differs. **Formal**: a mathematical guarantee rather than an empirical rate (○: formal bounds used internally, empirical robustness reported). **Channel**: what the adversary perturbs — *act*=executed actions, *obs*=observations, *data*=training data, *dyn*=dynamics/disturbance, *kern*=transition kernel. **Budget**: geometry of the attack set — (K, b)=cardinality *and* magnitude, ℓ_p=norm ball at every step, ℓ_0=count only, *set*=uncertainty set. **Fixed** π: the guaranteed object is the deployed policy itself, not a smoothed, retrained, robust-argmax or aggregated surrogate. **Level**: *ep*=whole episode, *st*=per state. **Form**: *tail*=$\Pr[\cdot \le \alpha] \ge \rho$, *mean*=expected return, *wc*=worst case. **Paired**: degradation measured against the same episode/-model unattacked. **Resolved**: the output is a function of state (and time), not a scalar. To the best of our knowledge, the cardinality–magnitude budget appears in the *attack* literature but in no prior *certificate*. For FoSeRL, this denotes a state–time witness and diagnostics, not certified conditional reliability at each state.

coverage its central concerns (Precup et al., 2000; Chen & Jiang, 2019). We instead assume generative access and collect fresh coupled executions of the certified policy under the declared attack law, so no off-policy correction or concentrability assumption enters the analysis. The gap is sharper for the paired objective. A single-branch nominal log does not contain the attacked counterfactual under the same realization of the episode randomness, and for deterministic policies trajectory-wise importance sampling degenerates once the attacked execution leaves the support of the nominal behaviour. Recovering the paired quantity from logs alone would therefore require the realized randomness together with a simulator, or a learned model able to generate the counterfactual branch. In this respect FoSeRL is closer to high-confidence policy evaluation (Thomas et al., 2015; 2019), though it returns a state- and time-dependent certificate for a degradation tail event rather than a scalar bound on policy value. One connection does carry over. Because the barrier is synthesized and calibrated from sampled coupled trajectories, pointwise readings of it outside regions the operational data represent are extrapolation, echoing the support concerns of offline RL (Fujimoto et al., 2019). This does not affect the distributional certificate, which concerns fresh draws from the declared law, but it bounds how the barrier should be read away from that coverage. Pessimistic regularization (Kumar et al., 2020; Yu et al., 2020), learned-model routes to certification from deployment logs (Yu et al.,

Table 5: Related formal robustness and certification frameworks, compared against FoSeRL. All three provide formal (high-probability) guarantees; they differ in what is certified and under which perturbation model. Methods with a different problem formulation are listed for completeness but are not used as quantitative baselines.

Method	Problem formulation and certified property	Attack model and main assumptions	Comparable to FoSeRL?
Per-state policy smoothing	Per-decision robustness of the policy map: certifies action agreement at individual observations, with no statement about accumulated trajectory degradation	ℓ_2 ball applied per observation; each visited state certified independently, with fresh sampling at every point and no coupling across timesteps	No
Gaussian policy smoothing	Sequential robustness under actuator-channel perturbation: certifies the same event as FoSeRL, $P(D_{\mathrm{traj}} \le \alpha) \ge \rho$	Worst case over all $\|\delta\|_2 \le b\sqrt{n_a K}$ via a Neyman–Pearson shift bound, paired by a union bound. Requires noise $\mathcal{N}(0, \sigma^2 I)$ shared by both branches and entering the *same* channel as the attack; τ grid fixed on design data	Yes
ReCePS	Robustness of expected return: certifies a lower bound on $\mathbb{E}[J]$ rather than a per-trajectory event	Observation-space ℓ_2 ball of radius B; guarantee holds on the mean, not per trajectory. Smoothed policy evaluated at inference	No

2020; Kidambi et al., 2020), and certificate-aware policy improvement in the spirit of safe policy improvement (Laroche et al., 2019) are natural directions, and lie outside the present scope.

F.2 Why policy smoothing is the only external baseline

A method qualifies as a quantitative baseline for FoSeRL only if it certifies the *same object under the same threat*: (i) a formal guarantee rather than an empirical robustness rate; (ii) on the per-trajectory event $\{D_{\mathrm{traj}} \le \alpha\}$ rather than on an expectation or a per-state property; (iii) under perturbations of the executed action with a cardinality budget; and (iv) for the policy as deployed. Applying these four criteria to the literature of Table 4 leaves exactly one method.

Gaussian policy smoothing (comparable). When the smoothing noise occupies the same channel the attack perturbs, the attacked branch is the nominal closed loop driven by shifted noise, and a Neyman–Pearson shift bound certifies the same terminal event we do. It is therefore our external baseline. Its guarantee is in one respect *stronger* than ours: it holds worst-case over every δ with $\|\delta\|_2 \le b\sqrt{n_a K}$, a set strictly containing our attack class, whereas our data-driven certificate holds under the declared attack law. The comparison is accordingly a budget–scope trade, and the certified budgets differ by $140\times$–$1575\times$ in the comparable cells.

Direct terminal-event certification (internal reference). Direct applies a one-sided Clopper–Pearson bound to the paired terminal event under the same operational specification as FoSeRL. It provides a formal scalar reference without barrier synthesis or conditional successor queries. Appendix E.4 explains the additional state–time witness and condition-specific diagnostics supplied by FoSeRL.

Methods excluded, and why. *Per-state smoothing* certifies action agreement at individual observations and makes no statement about accumulated degradation, so no budget of ours maps onto it. *ReCePS* bounds $\mathbb{E}[J]$ rather than a per-trajectory event and states its guarantee in observation space; a comparison would require transplanting $B = b\sqrt{n_a K}$ onto the action channel, and our reimplementation yields smoothed-return evaluations but no certified bound. *COPA* certifies against training-data poisoning, a different attack phase entirely. *Robust-training methods* (RARL, NR-/PR-MDP, ATLA, WocaR, RADIAL) harden a policy but emit no certificate for a given one, and change the policy in the process, violating (i) and (iv). *Robust MDPs* synthesise a policy under kernel ambiguity rather than auditing a fixed one. *Stochastic and neural barrier certificates* supply our machinery but bound state-safety events, not adversarial performance loss. Reporting numbers from any of these against ours would compare different certified quantities, not different certification methods.

Table 6: Formal comparison with the one method addressing the same problem formulation as FoSeRL. All quantities are formally certified rather than empirically estimated, at $\rho_{\text{target}} = 0.90$ and confidence $1 - \beta = 0.999$ ($\beta = \beta_0 = 10^{-3}$). FoSeRL gives a distributional guarantee; smoothing gives a worst-case guarantee over the ℓ_2 ball $\|\delta\|_2 \leq b\sqrt{n_a K}$. One representative cell per environment; the last column is the certified budget ratio over the reported nonzero-PS rows in Table 3

				FoSeRL		Smoothing		
Environment	K	$\alpha_{\%}$	T	b_{cert}	ρ_{cert}	b_{cert}	ρ_{cert}	Ratio
Pendulum (SAC)	27	2.5%	200	1.5	0.9174	1.01×10^{-2}	0.9000	148.5–301.2×
Hopper (TD3)	50	2.5%	1000	1.64	0.9724	2.00×10^{-3}	0.9000	512.5–1100×
HalfCheetah (SAC)	50	10%	500	0.35	0.9132	1.70×10^{-3}	0.9000	140–1575×

Fairness of the smoothing comparison. Two conditions bound where the baseline applies, and both are properties of the environment rather than of our implementation. The noise must occupy the attack's channel, which fails on ACC, where the process noise acts on the lead vehicle's speed while the attack perturbs the ego control. And the nominal return distribution must place sufficient mass in an α-wide window, since at $b = 0$ the paired bound collapses to that mass; this is an infinite-data ceiling, not a sample-size artefact, and it is why Unicycle ($S_{\text{env}}/\sigma = 6.45$, against 217 on the pendulum) cannot be certified by this baseline at any n. We also verified that the baseline is not sample-starved: with β split over the τ grid, the Clopper–Pearson cap on $\rho(b{=}0)$ alone manufactures failures below $n \approx 2048$, so every smoothing number reported here uses $n = 4096$, where that cap is 0.994 against a 0.90 target.

F.3 Evaluation regime: where smoothing has been tested, and where it has not

The smoothing line of work has been evaluated primarily on low-dimensional classic-control and Atari benchmarks. Kumar et al. (2022) report Cartpole, Mountain Car, Pong, and Freeway, using DQN for the discrete tasks and DDPG for Mountain Car; Mu et al. (2024) follow a similar evaluation regime on Cartpole, Mountain Car, and Freeway. The only continuous-action environment in these evaluations is Mountain Car, with a single actuated dimension. Neither work reports results on MuJoCo locomotion. This distinction becomes important as the action dimension and attack cardinality increase. Under the aligned smoothing construction used here, the trajectory-level perturbation radius scales as $b\sqrt{n_a K}$, so larger n_a and K make the resulting certified per-step budget increasingly restrictive. Our Hopper and HalfCheetah experiments therefore evaluate the baseline in a substantially higher-dimensional and longer-horizon regime than the settings considered in the original smoothing evaluations.

Representative quantitative and qualitative comparisons in these regimes are reported in Appendix D.8.