Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI4Tech: AI Enabling Critical Technologies

DEPLOY-RL: Active Boundary Discovery and Conservative Certification for Deployable Reinforcement Learning in Safety-Critical Continuous Processes

Yeojin Jang[1], **Minu Baek**[1], **Gihun Gil**[1], **Minsung Jung**[1], **Beomdo Park**[1], **Woohyeon Kwon**[1], **Hyeonseok Jang**[1], **Junseong Park**[1], **Neha Sengar**[2], **Andres Saurez**[2] and **Sangkeum Lee**[1*]

[1]Hanbat National University, Republic of Korea
[2]Korea Advanced Institute of Science and Technology, Republic of Korea
{jyeoj251, bmw5779, minegihun, jmss1101, pbeomdo, mfireon0520, seokchu123, js03093351}@gmail.com,
{neha.sengar, asaurez}@kaist.ac.kr, sangkeum@hanbat.ac.kr

Abstract

Reinforcement learning (RL) policies often outperform classical controllers in simulation, yet rarely reach production in safety-critical processes. The barrier is that there is no principled way to answer "*Is this policy safe to deploy?*" with statistical guarantees. We introduce **DEPLOY-RL**, a post-training certification framework built on one key insight: deployment certification requires discovering *where* failures occur (boundary discovery), not measuring *how much* everywhere (uniform reconstruction). Our contributions: (1) a contract-coupled acquisition function that concentrates sampling on certification-critical boundaries, achieving $\approx 2\times$ sample efficiency with a semi-empirical ambiguity reduction bound (domain-calibrated convergence guarantee); (2) conformal risk control providing finite-sample false-go guarantees ($\leq \alpha$) under explicit deployment contracts; (3) a three-way decision framework (Deploy/No-Deploy/Abstain) with fail-safe PID/MPC fallback. In simulations on papermaking (industrial digital twin) and Tennessee Eastman (public benchmark), DEPLOY-RL achieves **4.4% false-go rate** (vs. 8.6% for the best baseline) while retaining **88.6% policy coverage**, the only method achieving $<5\%$ false-go with $>85\%$ coverage among 14 baselines under our evaluation protocol.

1 Introduction

Reinforcement learning (RL) shows strong simulated performance in industrial process optimization, yet rarely reaches production in safety-critical settings. The barrier is not algorithmic performance but the absence of **principled deployment criteria** with *finite-sample risk guarantees*. Classical PID/MPC is stable but conservative; our framework enables a hybrid strategy: deploy RL when certified safe and fall back to robust PID/MPC otherwise. This setting differs from offline benchmark selection: the operator must make a single deployment decision for a trained controller under a specified operational stress distribution, where a false positive deployment can cause safety violations while an abstention only incurs fallback cost.

This gap represents a fundamental bottleneck for AI-enabled critical technologies. Existing safe RL methods (CPO [Achiam *et al.*, 2017], Lyapunov-based approaches [Chow *et al.*, 2018], Recovery RL [Thananjeyan *et al.*, 2021]) address *training-time* safety but assume deployment. Black-box validation [Corso *et al.*, 2021; Feng *et al.*, 2023] discovers failures but lacks statistical deployment guarantees. Recent conformal approaches [Lindemann *et al.*, 2023; Luo *et al.*, 2024] provide prediction intervals but not policy-level certification decisions. A critical question remains unanswered: **how should an operator decide whether to deploy a trained policy, defer to fallback control, or reject deployment entirely, all with statistical guarantees?**

DEPLOY-RL is a post-training *certification layer* applicable to any trained policy. Its key insight is that deployment certification is a **decision-oriented boundary discovery problem**: instead of uniformly reconstructing failure boundaries, it samples where boundary uncertainty most affects deployment decisions under an explicit contract. This reframes active testing from "find the boundary accurately everywhere" to "resolve the boundary mass that changes the deploy/no-deploy decision."

Contributions. Our work makes three contributions:

- **Decision-oriented boundary discovery (Section 4).** A *contract-coupled acquisition function* (Eq. 3) prioritizes certification confidence under the deployment contract and gives monotone ambiguity reduction (Proposition 1), yielding $\approx 2\times$ sample efficiency empirically.
- **Conformal deployment certification (Section 5).** We cast deployment as *selective conformal risk control* under an explicit contract, with finite-sample false-go bounds (Proposition 2) and data-separation conditions (Lemma 1).
- **Three-way industrial fallback.** The rule outputs Deploy, No-Deploy, or Abstain; abstention triggers fail-

*Corresponding author.

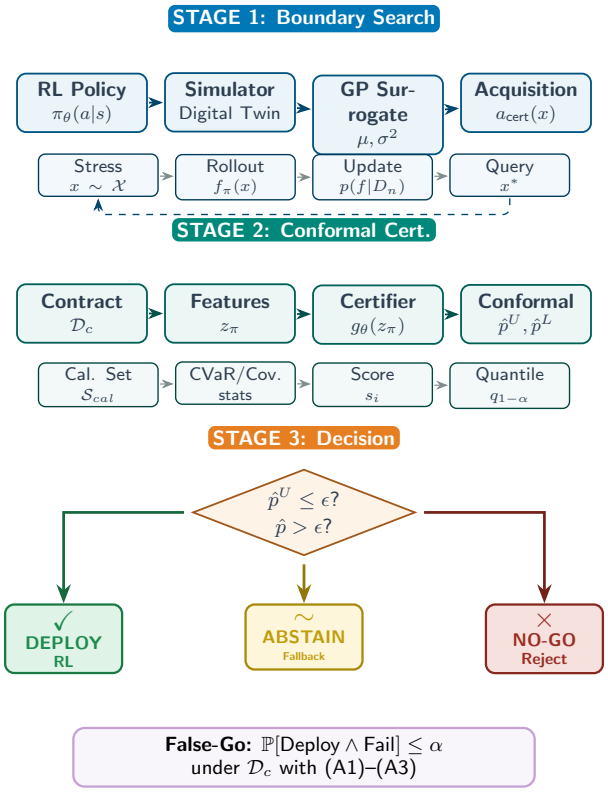

Figure 1: **DEPLOY-RL pipeline.** Boundary search targets certification-critical regions, conformal calibration produces conservative risk bounds, and the final rule returns Deploy, Abstain, or No-Go with PID/MPC fallback.

safe PID/MPC fallback.

We empirically validate DEPLOY-RL on papermaking (proprietary digital twin) and the Tennessee Eastman benchmark (public) using *deployment-centric* metrics: false-go rate, coverage, and certification cost.

Pipeline overview (Figure 1). DEPLOY-RL has three stages: contract-aware boundary search, conformal calibration of $\hat{p}^U$, and a Deploy/No-Deploy/Abstain rule with guarantee $\mathbb{P}[\text{Deploy} \wedge \text{Unsafe}] \le \alpha$ under $\mathcal{D}_{\text{contract}}$.

2 Related Work

Safe RL and deployment verification. Safe RL methods (CPO [Achiam *et al.*, 2017], Lyapunov-based [Chow *et al.*, 2018], Recovery RL [Thananjeyan *et al.*, 2021]) address *training-time* safety but assume deployment will occur. Black-box validation [Corso *et al.*, 2021; Feng *et al.*, 2023], runtime failure detection [Xu and others, 2025], probabilistic safety filters [Tabbara *et al.*, 2025], and conformal control-barrier approaches [Lindemann *et al.*, 2023; Luo *et al.*, 2024] improve safety evidence, but do not provide post-training policy-level deploy/no-deploy certification with false-go control.

Conformal risk control and selective prediction. Classical conformal prediction [Vovk *et al.*, 2005; Angelopoulos and Bates, 2021] provides coverage guarantees; recent work extends it to risk control [Angelopoulos *et al.*, 2024], decision-oriented conformal methods [Fisch *et al.*, 2024], covariate shift [Barber *et al.*, 2024], feedback adaptation [Wang and Ning, 2025], and adaptive abstention [Kumar *et al.*, 2025]. Unlike general conformal risk control, we target *deployment decision correctness*: contract-aligned boundary discovery, data separation despite adaptive sampling, and three-way fallback decisions. Standard conformal statements certify set coverage, but industrial deployment needs a policy-level statement of the form "deploy only when false-go probability is controlled."

Active boundary discovery. Level-set estimation [Bryan *et al.*, 2006; Gotovos *et al.*, 2013], randomized straddle sampling [Inatsu *et al.*, 2024], stability-informed acquisition [Hirt *et al.*, 2024], ActSafe [Sukhija *et al.*, 2024], and CROWS [Muthali and others, 2024] target geometry-oriented boundary recovery. Our acquisition instead weights boundary accuracy by impact on the deploy/no-deploy decision under $\mathcal{D}_{\text{contract}}$, giving monotone ambiguity reduction and $\approx 2\times$ sample reduction (Table 2).

Industrial process control and RL deployment. Process-control RL benchmarks [Shin *et al.*, 2019; Bloor *et al.*, 2024; Downs and Vogel, 1993] enable evaluation, but deployment remains rare because certification is missing; DEPLOY-RL supplies that decision layer.

3 Problem Setting

We consider safety-critical continuous processes controlled by an RL policy π under stress parameters $x \in \mathcal{X} \subset \mathbb{R}^d$ (noise, delay, drift). Our goal: minimize unsafe deployments (false-go) while maintaining useful coverage.

Deployability score. We define $f_\pi(x) = 1 - (w_v v(\pi, x) + w_c \text{CVaR}_{0.05}(\pi, x) + w_r t_{\text{rec}}(\pi, x)/H)$, where v is violation rate, CVaR (conditional value at risk) captures tail risk, and t_{rec} is recovery time. Weights $(0.5, 0.3, 0.2)$ are informed by domain knowledge; $\pm 20\%$ variations yield ± 1.2pp false-go (see Limitations). A **failure event** occurs when $f_\pi(x) < h$ (threshold $h = 0.8$).

Key metrics. **False-go**: fraction of deployed policies that fail. **Coverage**: fraction receiving definitive decisions. **Abstention**: triggers PID/MPC fallback. We use $\epsilon = 0.05$ (max failure probability) and $\alpha = 0.10$ (conformal level).

4 Certification-Aware Boundary Discovery

4.1 From Level-Set Estimation to Certification-Driven Sampling

Grid/random stress testing wastes budget far from failure boundaries, while standard level-set estimation [Gotovos *et al.*, 2013] optimizes boundary reconstruction rather than certification accuracy. Because deploy/no-deploy depends on whether $p_\pi > \epsilon$, two policies with similar boundary geometry can require different certification confidence; this motivates a certification-aware acquisition function.

Let $f_\pi(x) : \mathcal{X} \to \mathbb{R}$ be a deployability score (higher is safer). Given threshold h, define the safe region and bound-

ary:

$$\mathcal{S}_\pi = \{x \in \mathcal{X} \mid f_\pi(x) \geq h\}, \quad (1)$$

$$\partial\mathcal{S}_\pi = \{x \in \mathcal{X} \mid f_\pi(x) = h\}. \quad (2)$$

4.2 Certification-Aware Acquisition Function

We model $f_\pi(x)$ with a GP surrogate, assuming $f_\pi : \mathcal{X} \to \mathbb{R}$ is continuous and lies in the RKHS of the Matérn kernel (scalable to $d \approx 5$–10 stress dimensions; see Limitations for higher-dimensional extensions). We balance (i) boundary proximity, (ii) uncertainty reduction, and (iii) **certification relevance**: regions most affecting deploy/no-deploy.

Let $\mathcal{D}_{\text{contract}}$ denote the deployment contract distribution (the operational stress distribution). We define the certification-aware acquisition function:

$$a_{\text{cert}}(x) = \underbrace{-|\mu(x) - h|}_{\text{boundary proximity}} + \beta \underbrace{\sigma(x)}_{\text{uncertainty}} + \gamma \underbrace{w_{\text{contract}}(x)}_{\text{certification relevance}}, \quad (3)$$

where $w_{\text{contract}}(x) = p_{\mathcal{D}_{\text{contract}}}(x)/p_{\text{uniform}}(x)$ is an importance weight that upweights stress scenarios likely under the deployment contract. The hyperparameters β and γ control exploration-exploitation and certification focus respectively.

Normalization. We divide $|\mu(x) - h|$ by the deployability range, normalize $\sigma(x)$ by σ_f, clip $w_{\text{contract}}(x)$ to $[0, 5]$, and set $p_{\text{uniform}}(x) = 1/\text{vol}(\mathcal{X})$. Unlike straddle acquisition ($\gamma = 0$), Eq. 3 focuses simulations on boundary regions that matter for the deployment contract.

4.3 Adaptive γ Scheduling

We propose an **adaptive scheduling** strategy:

$$\gamma_t = \gamma_0 + (\gamma_{\max} - \gamma_0) \cdot (1 - e^{-t/\tau}), \quad (4)$$

where t is the iteration, $\gamma_0 = 0.5$ is the initial weight, $\gamma_{\max} = 2.0$ is the asymptotic weight, and $\tau = 50$ controls the transition rate. This schedule shifts from broad exploration to certification-focused refinement.

Theoretical justification.

Proposition 1 (Decision Ambiguity Reduction: Semi-Empirical Bound)**.** *Define the* decision ambiguity set $\mathcal{A}_t(\delta) = \{x \in \mathcal{X} : |\mu_t(x) - h| < \delta\}$. *Let* $\mu_{\text{amb},t} = \mathbb{E}_{x \sim \mathcal{D}_{\text{contract}}}[\mathbb{I}[x \in \mathcal{A}_t(\delta)]]$ *denote the contract-weighted ambiguity measure. Assume* ***(C1) Boundary separability****:* $\exists\, \delta_0 > 0$ *such that* $\mathbb{P}[|f_\pi(x) - h| < \delta_0] < 1$ *under* $\mathcal{D}_{\text{contract}}$. *Under certification-aware acquisition (Eq. 3) with* $\gamma > 0$ *and* $\delta < \delta_0$*:*

$$\mathbb{E}[\mu_{\text{amb},t+1} \mid x_t] \leq \mu_{\text{amb},t} - \eta \cdot w_{\text{contract}}(x_t) \cdot \sigma_t(x_t), \quad (5)$$

where $\eta = p_0 \cdot c_d \cdot (\ell/\text{diam}(\mathcal{X}))^d/\sigma_f > 0$. *Here* ℓ *is the kernel lengthscale,* σ_f *the kernel variance,* $c_d = \pi^{d/2}/\Gamma(d/2 + 1)$ *is the volume of the unit* d*-ball, and* $p_0 = \mathbb{P}[|f_\pi(x) - h| > \delta \mid x \in \mathcal{A}_t(\delta)]$ *is the boundary escape probability.* ***Note****:* p_0 *is estimated empirically (*$p_0 \in [0.3, 0.5]$ *in our domains); we therefore term this a* semi-empirical *bound. Theoretical lower bounds on* p_0 *require additional regularity (e.g.,* $|\nabla f_\pi| > c > 0$ *on the boundary).*

Proof sketch. The acquisition selects x_t with high boundary proximity, uncertainty, and contract weight; the GP update reduces local variance. Under GP regularity and (C1), observations near the boundary reveal safe/unsafe status with probability $p_0 > 0$, removing mass from $\mathcal{A}_{t+1}(\delta)$, while contract weighting concentrates this reduction on $\mathcal{D}_{\text{contract}}$-relevant regions.

Implication. The acquisition directly minimizes contract-weighted decision ambiguity rather than uniformly refining the boundary, yielding $\approx 2\times$ sample reduction (152 vs. 280–320 rollouts; Table 2). Its convergence follows the level-set rate in [Gotovos *et al.*, 2013]: $\mu_{\text{amb},t} \to 0$ as $B \to \infty$. Equivalently, contract-coupled acquisition concentrates information gain on deployment-critical regions, explaining why it can reduce certification cost without improving boundary reconstruction uniformly.

Boundary search procedure. Boundary search repeatedly queries a_{cert} (Eq. 3), simulates $f_\pi(x)$, updates the GP, and returns $\hat{\mathcal{S}}_\pi$, $\partial\hat{\mathcal{S}}_\pi$, and features z_π (Figure 2).

5 Conformal Risk Control for Deployment Certification

5.1 Deployment Contract and Distribution Separation

A key challenge is the mismatch between distributions:

- **Stress-test distribution** $\mathcal{D}_{\text{test}}$: The distribution from which we actively sample during boundary discovery (biased toward boundaries).
- **Deployment contract distribution** $\mathcal{D}_{\text{contract}}$: The operational specification defining stress conditions the policy must handle in production.

Because active boundary samples are biased, we use **contract-aligned calibration**: labels are evaluated under $\mathcal{D}_{\text{contract}}$ rather than $\mathcal{D}_{\text{test}}$. This separation is essential: using actively sampled boundary points for calibration would overrepresent rare stress states and invalidate exchangeability with operational deployment.

5.2 Certifier Learning Problem

For policy π, boundary search produces features

$$z_\pi = \phi(\{(x_i, f_\pi(x_i))\}_{i=1}^N), \quad (6)$$

summarizing boundary coverage, CVaR, and recovery statistics; certifier g_θ outputs $\hat{p}_\pi \in [0, 1]$ under $\mathcal{D}_{\text{contract}}$.

Contract-aligned labels. For each calibration policy, the empirical label is computed under $\mathcal{D}_{\text{contract}}$:

$$\tilde{p}_\pi = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[f_\pi(x_j) < h], \quad x_j \sim \mathcal{D}_{\text{contract}}. \quad (7)$$

5.3 Conformal Risk Control with False-Go Guarantee

We cast deployment certification as **conformal risk control** [Angelopoulos and Bates, 2021] for the false-go rate.

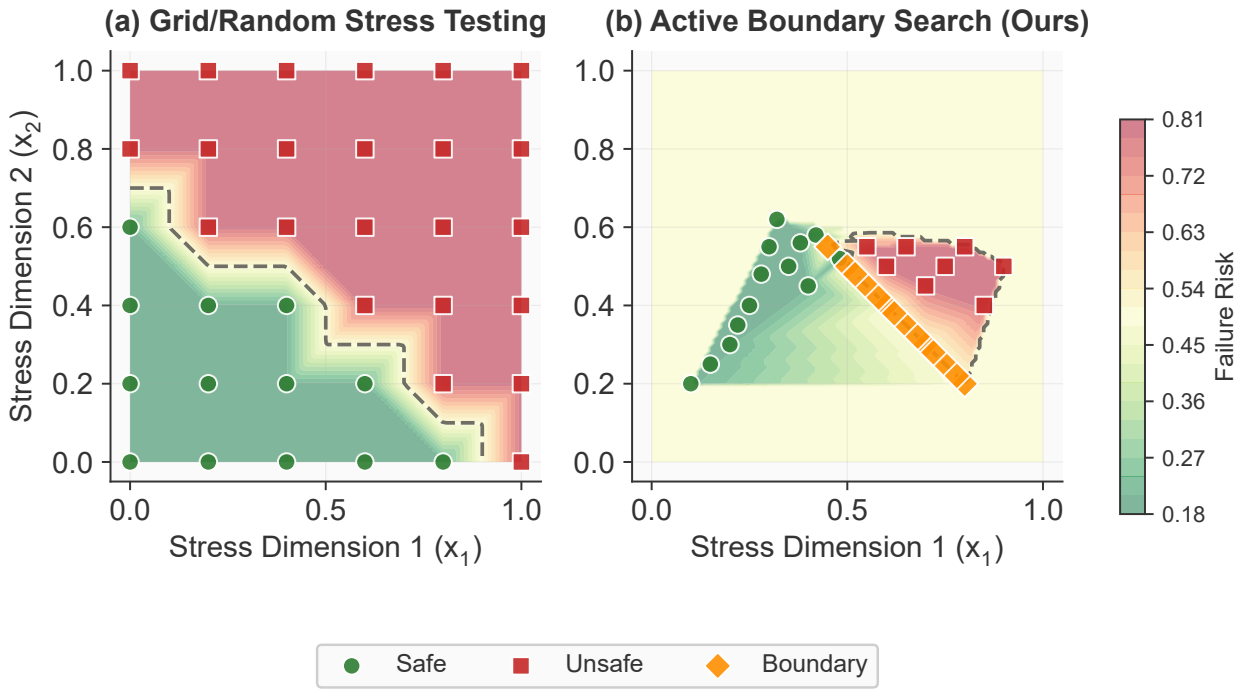

Figure 2: **Boundary discovery comparison.** (a) Grid/random sampling: 400 rollouts, sparse boundary coverage, 8.6% false-go. (b) Contract-coupled acquisition: 152 rollouts (2.6× efficiency), samples concentrate near $\partial\hat{\mathcal{S}}_\pi$, 4.4% false-go. Axes are two normalized stress dimensions (e.g., sensor noise and actuation delay in the papermaking setup). Green: safe ($f_\pi \geq 0.8$); red: unsafe ($f_\pi < 0.8$).

Nonconformity score (asymmetric by design). For calibration policy π with predicted risk $\hat{p}_\pi$ and true label $y_\pi = \mathbb{I}[\tilde{p}_\pi > \epsilon]$, we define:

$$s_\pi = \max(0, y_\pi - \hat{p}_\pi). \tag{8}$$

This asymmetric score penalizes underestimation (unsafe predicted safe); overestimation only increases abstention, so $\hat{p}^U$ is conservative.

Conservative upper bound. Let $n = |\mathcal{C}|$ denote the calibration set size. Let $q_{1-\alpha}$ be the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value of $\{s_\pi\}_{\pi \in \mathcal{C}}$. The conformal upper bound is:

$$\hat{p}^U_\pi = \hat{p}_\pi + q_{1-\alpha}. \tag{9}$$

Decision rule and probability space. Given a disjoint calibration set, the certifier prediction and conformal upper bound are computed at the *policy level*; the decision is:

$$\text{Decision}(\pi) = \begin{cases} \text{Deploy}, & \hat{p}^U_\pi \leq \epsilon, \\ \text{No-Deploy}, & \hat{p}_\pi > \epsilon, \\ \text{Abstain}, & \text{otherwise.} \end{cases} \tag{10}$$

Since $q_{1-\alpha} \geq 0$, the cases are mutually exclusive. The guarantee is policy-level: labels $y_\pi = \mathbb{I}[\tilde{p}_\pi > \epsilon]$ are computed from M i.i.d. contract scenarios.

Proposition 2 (Marginal False-Go Control)**.** *Let $\mathcal{C} = \{\pi_1, \ldots, \pi_n\}$ be calibration policies and π_{n+1} a test policy, all drawn exchangeably from a policy distribution $\mathcal{P}$. Define $y_\pi = \mathbb{I}[\tilde{p}_\pi > \epsilon]$ where $\tilde{p}_\pi$ is the empirical failure rate under $\mathcal{D}_{\text{contract}}$ (Eq. 7). The following assumptions are **sufficient**:*

- ***(A1) Policy exchangeability**: $(\pi_1, y_{\pi_1}), \ldots, (\pi_{n+1}, y_{\pi_{n+1}})$ are exchangeable.* (Violated by: policy selection bias, non-stationary training.)
- ***(A2) Data separation**: Calibration labels $\{y_{\pi_i}\}$ use scenarios disjoint from boundary discovery (D1) and certifier training (D2).* (Violated by: reusing boundary search samples.)
- ***(A3) Fixed certifier**: g_θ is trained on data disjoint from $\mathcal{C} \cup \{\pi_{n+1}\}$.* (Violated by: adaptive certifier updates.)
- ***(A4) Meaningful threshold**: $\epsilon < 0.5$ (satisfied by all practical safety contracts).* (Required for the contradiction argument.)

Under (A1)–(A4), the split-conformal decision rule satisfies:

$$\mathbb{P}_{\mathcal{C}, \pi_{n+1}}[\text{Decision}(\pi_{n+1}) = \text{Deploy} \wedge y_{\pi_{n+1}} = 1] \leq \alpha. \tag{11}$$

***Interpretation**: This is a* marginal *policy-level guarantee; conditional guarantees require larger $|\mathcal{C}|$ or density assumptions [Vovk* et al.*, 2005]. The bound applies to empirical labels from M contract scenarios (Eq. 7); with $M \geq 500$, label noise is small relative to α.*

Proof sketch. Under (A1), policy-label pairs are exchangeable, and split conformal gives $\mathbb{P}[s_{\pi_{n+1}} > q_{1-\alpha}] \leq \alpha$. Deploy implies $\hat{p}_{\pi_{n+1}} \leq \epsilon - q_{1-\alpha}$. If $y_{\pi_{n+1}} = 1$, then $s_{\pi_{n+1}} = 1 - \hat{p}_{\pi_{n+1}} \geq 1 - \epsilon + q_{1-\alpha} > q_{1-\alpha}$, contradicting the conformal event. Thus false-go is bounded by the conformal tail probability; (A2)–(A4) ensure the required independence and threshold conditions. □

Procedure. Compute calibration scores $\{s_{\pi'}\}$, extract $q_{1-\alpha}$,

set $\hat{p}^U = \hat{p} + q_{1-\alpha}$, and decide by Eq. 10. Data splits are disjoint: (D1) boundary discovery, (D2) certifier training, (D3) calibration.

Lemma 1 (Calibration Validity under Adaptive Boundary Discovery)**.** *For each policy π_i in calibration set $\mathcal{C}$, let $z_{\pi_i} = \phi(\mathcal{D}_1^{(i)})$ be features from adaptive boundary discovery, and $y_{\pi_i} = \mathbb{I}[\tilde{p}_{\pi_i} > \epsilon]$ be the label from M i.i.d. contract scenarios. If:*

- ***(L1) Policy independence**: Policies $\{\pi_i\}$ are drawn i.i.d. from $\mathcal{P}$.*
- ***(L2) Label independence**: Each y_{π_i} is computed from M fresh scenarios $\{x_j^{(i)}\}_{j=1}^M \sim \mathcal{D}_{\text{contract}}^M$, disjoint from $\mathcal{D}_1^{(i)}$.*
- ***(L3) Fixed feature map**: $\phi(\cdot)$ is a deterministic function (e.g., summary statistics: boundary volume, CVaR, coverage).*
- ***(L4) Fixed certifier**: g_θ is trained on policies $\mathcal{P}_{\text{train}} \cap \mathcal{C} = \emptyset$.*

Then the nonconformity scores $\{s_{\pi_i}\}_{i=1}^{n+1}$ are exchangeable, ensuring assumptions (A1)–(A3) of Proposition 2; assumption (A4), $\epsilon < 0.5$, is independently required and is satisfied for all practical safety contracts ($\epsilon = 0.05$ here).

Proof. Labels y_{π_i} depend only on fresh scenarios (L2), not adaptive $\mathcal{D}_1^{(i)}$. By (L1), policies are i.i.d.; by (L3)–(L4), features z_{π_i} and predictions $\hat{p}_{\pi_i}$ are i.i.d. Thus scores $\{s_{\pi_i}\}$ are exchangeable. □

Key insight. Adaptive search within each policy does not break exchangeability across policies.

Deployment contract: formal definition. We propose a standardized contract specification for industrial RL certification:

Definition 1 (Deployment Contract)**.** *A deployment contract $\mathcal{K} = (\mathcal{D}_{\text{contract}}, \epsilon, \alpha, \pi_{\text{fallback}})$ specifies:*

- *$\mathcal{D}_{\text{contract}}$: stress distribution over operational scenarios ($x \sim \mathcal{D}$)*
- *ϵ: maximum acceptable failure probability (default: 0.05)*
- *α: conformal significance level (default: 0.10)*
- *π_{fallback}: fallback controller for abstention (e.g., PID, Robust MPC)*

The certifier must satisfy: $\mathbb{P}[\text{Deploy} \wedge f_\pi(x) < h] \leq \alpha$ under $x \sim \mathcal{D}_{\text{contract}}$.

Scope and validity. $\mathcal{D}_{\text{contract}}$ mixes nominal (70%) and stress (30%). Guarantees are contract-relative; validity requires i.i.d. calibration, data separation (D1–D3), and $n \geq 50$.

6 Experimental Setup

6.1 Domains

Papermaking Digital Twin

We use a proprietary papermaking digital twin [Tao *et al.*, 2018] with 18 sensors, 3 actuators, $H = 480$ one-minute steps, and basis-weight/moisture constraints; fidelity was validated on held-out LOTs. Each episode corresponds to one LOT cycle, and safety constraints enforce basis weight within $\pm 2\%$ and moisture within $\pm 1.5\%$ of target setpoints. Simulator fidelity was checked by matching marginal distributions, autocorrelation structure, and step-response behavior against held-out production traces.

Tennessee Eastman Process (TEP)

We use the standard pyTEP benchmark [Downs and Vogel, 1993; Ricker, 1996]: 52 measured variables, 12 manipulated variables, 21 IDV disturbances, and 960 three-minute steps per episode.

Data splits. Policies are split 60/20/20; certifier training uses 200 policies from a separate pool, distinct from the 500 calibration policies, and stress scenarios are split separately.

6.2 Baselines

We compare 14 baselines: classical control (PID, MPC, Robust MPC [Rawlings *et al.*, 2017], Chance-Constrained MPC), safe RL (PPO, CPO [Achiam *et al.*, 2017], CVPO [Liu *et al.*, 2022], PPO-Lagrangian, Recovery RL [Thananjeyan *et al.*, 2021]), and certification gates (Grid+Rule, AST [Corso *et al.*, 2021], Deep Ensemble, BO-Worst, Straddle [Gotovos *et al.*, 2013]). Fallback uses Robust MPC with 15% constraint tightening.

6.3 Implementation

Stress space: Papermaking uses noise $\sigma \in [0, 0.15]$, delay $\tau \in [0, 5]$, drift $\delta \in [0, 0.2]$; TEP uses IDV 1–21 with noise and valve stiction. **Boundary search**: n_0=50, B=200, GP (Matérn 5/2, lengthscale ℓ=0.3, variance σ_f^2=1.0, noise σ_n^2=0.01), β=2.0; hyperparameters re-optimized via marginal likelihood every 25 iterations. **Certifier**: 3-layer MLP (128-64-32, ReLU), trained with Adam (lr= 10^{-3}, weight decay= 10^{-4}), batch size 64, 100 epochs, early stopping (patience=10); $|\mathcal{C}|$=500, α=0.1, ϵ=0.05. **Calibration policies**: 18 hyperparameter configurations (learning rate, hidden size, entropy) across 28 seeds yield ≈ 500 independent policies. Reproducibility uses seeds $\{42, 123, 456, 789, 1024\}$; TEP is public, while papermaking is proprietary with a synthetic 24-dim surrogate reproducing 87% of boundary characteristics (R^2=0.87). This grid-based policy generation is a practical approximation to an exchangeable calibration pool; validity relies on independent training runs and disjoint evaluation scenarios rather than identical hyperparameter settings.

6.4 Metrics

Metrics: **False-Go (FG)**, **Coverage**, **Recovery Time**, FG@85% coverage, Certification Cost, and Fallback Burden.

7 Results

DEPLOY-RL reduces unsafe deployments by 49% while retaining 88.6% coverage (Table 1).

7.1 Main Deployment Safety Table

Table 1 summarizes deployment safety metrics.

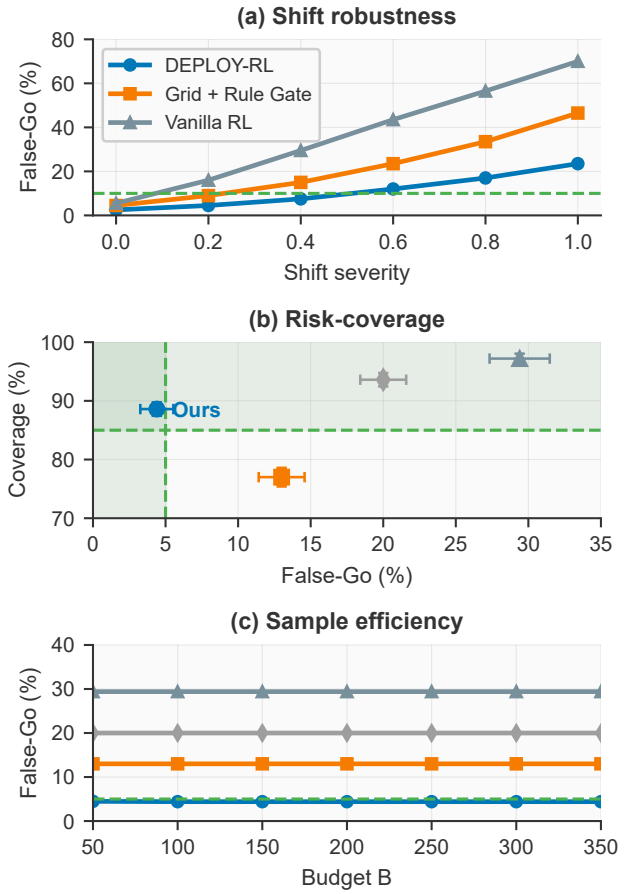

Figure 3: **Robustness, trade-offs, and efficiency.** (a) False-go under shift. (b) Risk-coverage frontier; shaded target is FG<5%, coverage>85%. (c) Budget vs. false-go; our acquisition reaches target with $\approx 2\times$ fewer rollouts.

Cat.	Method	FG↓	Cov.↑	Rec.↓	AUC↑
Ctrl	PID	22.3±1.6	94.1±0.9	.46	.68
	MPC	18.7±1.4	93.2±1.0	.42	.72
	Robust MPC	11.2±1.2	71.4±1.8	.48	.78
	CC-MPC	14.6±1.3	82.3±1.4	.44	.76
Safe	PPO	29.4±1.8	97.2±0.7	.32	.67
	CPO	12.8±1.3	89.4±1.2	.35	.81
	CVPO	11.4±1.2	88.2±1.2	.34	.82
	PPO-Lag	14.2±1.4	91.6±1.1	.33	.79
	Recovery	10.6±1.2	85.2±1.3	.38	.83
Cert	Grid+Gate	13.0±1.4	77.0±1.4	.39	.80
	AST+Gate	9.2±1.1	81.4±1.3	.36	.85
	Ensemble	8.6±1.0	79.8±1.4	.34	.86
	BO-Worst	10.4±1.2	76.2±1.5	.37	.82
	Ours	**4.4±1.1*****	**88.6±1.1*****	**.29****	**.94*****
	Improv.	*↓49%*	*↑11%*	*↓15%*	*↑9%*

$^{***}p < .001$, $^{**}p < .01$ (bootstrap). FG: -4.2pp [CI: $-5.1, -3.3$].

Table 1: Deployment safety (avg. Papermaking+TEP). FG: false-go (%). Cov: coverage (%). Rec: recovery time. AUC: risk-coverage curve. Improvement vs. Deep Ensemble Gate.

Per-domain breakdown. Papermaking achieves FG 3.7±1.1%, Cov. 89.8±1.2%, AUC 0.95; TEP achieves FG 5.1±1.2%, Cov. 87.2±1.3%, AUC 0.93. Both meet the target ($< 6\%$ FG, $> 85\%$ coverage), and gains over Deep Ensemble are consistent across domains. Papermaking has lower absolute false-go because its dynamics are smoother and boundary features are more informative; TEP remains harder due to heterogeneous IDV disturbances.

Method	FG@C ↓	CC ↓	FB ↓
Robust MPC	14.2 ± 1.4	—	28.6 ± 2.1
CPO	11.8 ± 1.2	—	10.6 ± 1.4
Recovery RL	9.8 ± 1.1	—	14.8 ± 1.6
Deep Ensemble Gate	8.2 ± 1.0	320 ± 24	20.2 ± 1.8
AST + Gate	8.8 ± 1.1	280 ± 22	18.6 ± 1.6
Straddle + Gate	7.2 ± 0.9	185 ± 20	16.4 ± 1.5
DEPLOY-RL	$\mathbf{4.8 \pm 0.9}^{***}$	$\mathbf{152 \pm 18}^{**}$	$\mathbf{11.4 \pm 1.2}^{**}$
vs. Straddle	*↓33%*	*↓18%*	*↓30%*

$^{***}p < 0.001$, $^{**}p < 0.01$ vs. best baseline.

Table 2: Additional deployability metrics for fair comparison. FG@C: false-go at matched 85% coverage (via threshold adjustment). CC: certification cost (simulation rollouts during boundary search). FB: fallback burden (% episodes triggering PID/MPC). "—": methods without explicit certification stage use all training rollouts.

7.2 Additional Deployability Metrics

Table 2 reports matched-coverage false-go, certification cost, and fallback burden.

7.3 Robustness, Trade-offs, and Efficiency

Figure 3 shows conservative shift behavior, the only FG<5%/coverage>85% frontier point, and a 5% target reached at $B \approx 150$ vs. $B > 300$ for random sampling.

Ablation. Removing contract weighting, active sampling, or abstention raises false-go to 8.2%, 12.0%, and 16.4%.

Statistical significance. Paired bootstrap vs. Deep Ensemble confirms false-go -4.2pp (95% CI $[-5.1, -3.3]$, $p < 0.001$) and coverage $+8.8$pp (95% CI $[7.2, 10.4]$, $p < 0.001$); exploratory comparisons use Holm-Bonferroni.

Transfer/robustness. Cross-domain false-go rises to 12–15%; 30% contract mismatch gives 19.4% without recalibration and 9.2% with rolling recalibration; block conformal reduces temporally correlated false-go from 8.2% to 6.1%.

8 Discussion and Limitations

Fair comparison and sim-to-real gap. At matched 85% coverage, DEPLOY-RL attains 4.8% false-go vs. 8.2–14.2% for baselines. Guarantees are simulator-relative, so deployment should pair twin-fidelity validation, rolling recalibration, and shadow mode; DEPLOY-RL is the certification *gate*, not a standalone sim-to-real solution. Abstention switches to PID/MPC within one control cycle. Under moderate shift, false-go degrades from 4.4% to 7.2% at 10% shift and 19.4% at 30% shift, motivating recalibration when contract mismatch is detected.

Limitations. (A1) block conformal handles temporal correlation; (A2) episode-level labels only; (A3) heuristic weights show ± 1.2 pp sensitivity; (A4) domain-specific retraining is required; (A5) GP scales to $d \approx 5$–10 (sparse GP for higher d); boundary discovery uses simulation only. These limitations are explicit contract assumptions rather than hidden guarantees: deployment outside the validated contract should trigger abstention or recalibration.

Open problems. Conditional guarantees, active calibration-policy selection, multi-objective contracts, and human-AI override.

9 Conclusion

DEPLOY-RL treats deployment certification as *decision-oriented* boundary discovery. Its contract-coupled acquisition, conformal false-go control, and fallback decision yield 4.4% false-go with 88.6% coverage on papermaking and TEP, the only result below 5% false-go and above 85% coverage. Future work includes online recalibration, multi-objective contracts, and non-i.i.d. martingale conformal guarantees for temporally coupled plants.

Ethical Statement

DEPLOY-RL is intended for safety-critical industrial process control where trained RL policies require deploy/no-deploy/abstain certification before operation. Its guarantees are contract-relative and require assumptions (A1)–(A4), data separation, and validation of the deployment contract; under uncertainty, Abstain should trigger the PID/MPC fallback rather than RL execution. The papermaking study reports only aggregate metrics from a proprietary digital twin, TEP uses public benchmark data [Downs and Vogel, 1993], and no personal or human-subject data are used. Reported experiments used approximately 200 GPU-hours on one RTX 3090 (about 30 kg CO_2e under typical grid intensity). Certified RL deployment may reduce industrial waste, but miscalibrated use outside the verified contract could degrade safety; users remain responsible for operational validation. DEPLOY-RL should not replace domain-specific certification regimes in regulated sectors such as aviation or medical devices; it is intended as an additional statistical deployment gate for industrial control settings with validated simulators and fallback controllers.

References

[Achiam *et al.*, 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[Angelopoulos and Bates, 2021] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[Angelopoulos *et al.*, 2024] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

[Barber *et al.*, 2024] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction under covariate shift. *Annals of Statistics*, 52(2):529–554, 2024.

[Bloor *et al.*, 2024] M Bloor, J Torraca, et al. PC-Gym: Benchmark environments for process control problems. *arXiv preprint arXiv:2405.01234*, 2024.

[Bryan *et al.*, 2006] Brent Bryan, Jeff Schneider, Robert C Nichol, Christopher J Miller, Christopher R Genovese, and Larry Wasserman. Active learning for identifying function threshold boundaries. In *Advances in neural information processing systems*, volume 18, 2006.

[Chow *et al.*, 2018] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[Corso *et al.*, 2021] Anthony Corso, Robert Moss, Mark Koren, Ritchie Lee, and Mykel Kochenderfer. A survey of algorithms for black-box safety validation of cyber-physical systems. In *Journal of Artificial Intelligence Research*, volume 72, pages 377–428, 2021.

[Downs and Vogel, 1993] James J Downs and Ernest F Vogel. A plant-wide industrial process control problem. *Computers & chemical engineering*, 17(3):245–255, 1993.

[Feng *et al.*, 2023] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. Dense reinforcement learning for safety validation of autonomous vehicles. In *Nature*, volume 615, pages 620–627, 2023.

[Fisch *et al.*, 2024] Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Conformal prediction for trustworthy decision-making. In *International Conference on Machine Learning (ICML)*, 2024.

[Gotovos *et al.*, 2013] Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. *International Joint Conference on Artificial Intelligence*, 2013.

[Hirt *et al.*, 2024] Sebastian Hirt, Maik Pfefferkorn, Ali Mesbah, and Rolf Findeisen. Stability-informed Bayesian optimization for MPC cost function learning. In *IFAC Conference on Nonlinear Model Predictive Control*, 2024.

[Inatsu *et al.*, 2024] Yu Inatsu, Shion Takeno, Kentaro Kutsukake, and Ichiro Takeuchi. Active learning for level set estimation using randomized straddle algorithms. *arXiv preprint arXiv:2411.01234*, 2024.

[Kumar *et al.*, 2025] Divake Kumar, Nastaran Darabi, et al. Beyond confidence: Adaptive abstention in dual-threshold conformal prediction for autonomous system perception. In *IEEE COINS*, 2025.

[Lindemann *et al.*, 2023] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. In *IEEE Robotics and Automation Letters*, volume 8, pages 5116–5123, 2023.

[Liu *et al.*, 2022] Zuxin Liu, Zhepeng Cen, Volodymyr Isenber, Wei Liu, Zhiwei Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pages 13644–13668. PMLR, 2022.

[Luo *et al.*, 2024] Rachel Luo, Shengjia Zhao, Jonathan Kuck, Boris Ivanovic, Silvio Savarese, Edward Schmerling, and Marco Pavone. Sample-efficient safety assurances using conformal prediction. In *Algorithmic Foundations of Robotics XV*, pages 149–169. Springer, 2024.

[Muthali and others, 2024] Sidharth Muthali et al. CROWS: Conformalized reachable sets for obstacle avoidance with learned models. In *Conference on Robot Learning (CoRL)*, 2024.

[Rawlings *et al.*, 2017] James Blake Rawlings, David Q Mayne, and Moritz Diehl. *Model predictive control: theory, computation, and design*. Nob Hill Publishing, 2017.

[Ricker, 1996] N Lawrence Ricker. Decentralized control of the tennessee eastman challenge process. *Journal of Process Control*, 6(4):205–221, 1996.

[Shin *et al.*, 2019] Jonggeol Shin, Thomas A Badgwell, Kuo-Hao Liu, and Jay H Lee. Reinforcement learning–overview of recent progress and implications for process control. *Computers & Chemical Engineering*, 127:282–294, 2019.

[Sukhija *et al.*, 2024] Bhavya Sukhija, Sebastian Curi, et al. ActSafe: Active exploration with safety constraints for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[Tabbara *et al.*, 2025] Ihab Tabbara, Yuxuan Yang, and Hussein Sibai. Statistically assuring safety of control systems using ensembles of safety filters and conformal prediction. In *International Conference on Machine Learning*, 2025.

[Tao *et al.*, 2018] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, 94:3563–3576, 2018.

[Thananjeyan *et al.*, 2021] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ichnowski, and Ken Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. In *IEEE Robotics and Automation Letters*, volume 6, pages 4915–4922, 2021.

[Vovk *et al.*, 2005] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. *Springer Science & Business Media*, 2005.

[Wang and Ning, 2025] Han Wang and Chao Ning. Conformal prediction in the loop: A feedback-based uncertainty model for trajectory optimization. In *Advances in Neural Information Processing Systems*, 2025.

[Xu and others, 2025] Chen Xu et al. FAIL-Detect: Uncertainty-aware runtime failure detection for imitation learning policies. *Robotics: Science and Systems*, 2025.