Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Training-Budget Sensitivity of Method Rankings in Safe Reinforcement Learning: A Case Study on Quadrotor Control Under Wind Disturbance

Mohamed Chaouli, Mohamed Elbar
[1] *Faculty of Science and Technology, University of Djelfa, Djelfa 17000, Algeria m.chaouli@univ-djelfa.dz*

ARTICLE INFO

Received: 05 April 2025
Revised: 07 Oct 2025
Accepted: 28 April 2026

ABSTRACT

Empirical comparisons in safe reinforcement learning (RL) are usually reported at a single, fixed training budget, and the resulting method ranking is then treated as a property of the algorithms. We show that on a quadrotor hover task under wind disturbance this ranking is instead strongly budget-dependent. Using two runs of an identical five-method pipeline—an unconstrained PPO baseline and four Lagrangian variants spanning a two-by-two factorial of error signal (mean cost versus CVaR) and dual-update rule (gradient ascent versus PID)—we find that in calm air the family of PID-controlled Lagrangian methods ranks last at 500K steps and first at 2M steps, while the unconstrained baseline moves in the opposite direction. Quantified by Spearman rank correlation between the two budgets, the ordering is negatively correlated at moderate wind (rho of minus 0.9, p of 0.037) and, pooled across all five wind conditions, significantly negative overall (rho of minus 0.4 over 25 method-condition cells, p of 0.048): methods that look better at 500K tend to look worse at 2M. Because the two runs differ in budget, seed count, and—for the CVaR cells—an adaptive cost-limit calibration introduced between them, we isolate the budget effect two ways that are immune to the calibration confound: (i) the periodic in-training logs of the 2M run, which compare 500K and 2M on the same 20 seeds with calibration held constant, and (ii) the mean-cost PID-Lag method, which uses no CVaR calibration yet still rises from the bottom at 500K (58.8) to the top at 2M (257.7). Bootstrap confidence intervals confirm the reversal. We further show that more compute is not uniformly better: under moderate and strong wind several methods lose return between 500K and 2M, and the safety (violation-rate) ranking shifts with budget as well. Finally, the two top methods at 2M are statistically indistinguishable (paired Wilcoxon, n of 20, corrected p of 1.0; effect size of 0 at the strongest wind), so the phenomenon is a property of the PID family, not of any single method. We argue that safe-RL benchmarks should report convergence evidence (learning curves) and budget sensitivity before publishing a ranking, and we give a concrete checklist. The 2M-step, 20-seed run re-analyzed here is the converged study reported in our companion paper (Chaouli and Elbar 2026b); all findings are recomputed from raw logs and no new training was performed.

Keywords: Safe reinforcement learning, reproducibility, training budget, evaluation methodology, Lagrangian methods, PID control, quadrotor, benchmarking, rank correlation.

INTRODUCTION

reinforcement learning (RL) trains a policy to maximize task reward while keeping a separate cost signal below a limit (Wachi et al. 2024; Gu et al. 2024). The dominant family is Lagrangian relaxation (Tessler et al. 2019; Stooke et al. 2020), and recent infrastructure such as OmniSafe (Ji et al. 2024) and Safety-Gymnasium (Ji et al. 2023) has made head-to-head method comparison routine. A comparison is typically run once, at a fixed number of environment steps, and the resulting ordering of methods is reported as the empirical contribution of the paper.

The RL evaluation literature has repeatedly warned that such single-shot comparisons are fragile. Henderson et al. (Henderson et al. 2018) showed that random seeds, hyperparameters, and even code base change which algorithm appears best; Agarwal et al. (Agarwal et al. 2021) demonstrated that point estimates from a handful of runs are

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

statistically unreliable and proposed interval estimators; Jordan et al. (Jordan et al. 2020) formalized how reported differences can stem from evaluation choices rather than algorithms; and Colas et al. (Colas et al. 2018) quantified the statistical power needed to make a claim. One axis has received far less attention in the *safe*-RL setting: the *training budget*. If two methods converge at different rates, a comparison taken before convergence measures convergence speed, not asymptotic quality, and the conclusion that ships with the paper can be the opposite of the converged one.

This paper provides a clean, controlled demonstration of that hazard on a safety-critical control task. We take an identical five-method safe-RL pipeline on a quadrotor hover task under wind disturbance and run it at two budgets, 500K and 2M steps. The methods are an unconstrained PPO baseline and the four cells of a two-by-two factorial: error signal in {mean cost, CVaR} crossed with dual-update rule in {gradient ascent, PID control}, yielding PPO-Lag, PID-Lag, CVaR-Lag, and CVaR-PID. Our findings:

1. **The leaderboard inverts with budget.** In calm air the PID-controlled Lagrangian family ranks last at 500K and first at 2M, while the unconstrained baseline does the reverse; quantified by Spearman rank correlation, the budget-to-budget ordering reverses most sharply at moderate wind (rho of minus 0.9, p of 0.037) and is significantly negative pooled across all five wind conditions (rho of minus 0.4 over 25 cells, p of 0.048) (Section 5.1, 5.2).
2. **The inversion is not a seed-count artifact.** Reconstructing the ranking from the periodic in-training evaluations of the 2M run isolates the budget effect on the *same* 20 seeds; the reversal persists, with bootstrap confidence intervals (Section 5.3).
3. **More compute is not uniformly better.** Under moderate and strong wind, several methods lose return between 500K and 2M, and the safety (violation-rate) ranking also moves with budget, so a global "train longer" rule is itself misleading (Section 5.4, 5.5).
4. **The effect is family-level, not method-level.** The two best methods at 2M are statistically tied (corrected p of 1.0, n of 20); the budget story concerns the PID family, and we frame it as such (Section 5.6).

We close with a reporting checklist for safe-RL comparisons (Section 6) and an explicit threats-to-validity analysis (Section 7). Every number below is recomputed directly from the stored experiment logs; the scripts are released with the paper, and *no new training was performed*—the study is an analysis of existing artifacts.

RELATED WORK

1. *Reliability of empirical RL comparisons*

Henderson et al. (Henderson et al. 2018) is the canonical demonstration that deep-RL results are sensitive to seeds, hyperparameters, and implementation, and that small-sample comparisons routinely report differences that do not hold up. Colas et al. (Colas et al. 2018) added a statistical-power lens: the number of seeds typically used cannot support the differences often claimed, and they give sample-size guidance for two-sample tests on RL returns. Agarwal et al. (Agarwal et al. 2021) introduced aggregate interval estimators—the interquartile mean, performance profiles, and the stratified bootstrap—to replace fragile point estimates, and showed that conclusions on standard benchmarks change when these are used. Jordan et al. (Jordan et al. 2020) proposed complete evaluation procedures that fix the evaluation protocol before measuring, removing a degree of freedom that otherwise inflates apparent differences. Our contribution is orthogonal and specific: we isolate the training-budget axis in safe RL and show a full ranking inversion, then de-confound budget from seed count using the same seeds, and quantify the inversion with rank correlation.

2. *Safe RL and Lagrangian methods*

Constrained Markov decision processes (Altman 1999) and their Lagrangian relaxation (Tessler et al. 2019) underpin most scalable safe RL (Wachi et al. 2024; Gu et al. 2024). Stooke et al. (Stooke et al. 2020) replaced gradient ascent on the dual variable with PID control to damp the oscillation that plagues the multiplier update. CVaR-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

constrained variants target tail risk rather than the mean cost (Zhang et al. 2025). A companion study on this platform isolates why, at a fixed 2M budget, the PID update is more reliable than gradient ascent; here we instead study how the ranking of these same methods depends on the budget itself, which a single-budget study cannot reveal. Spoor et al. (Spoor et al. 2025) report that Lagrangian instability persists across settings, consistent with our observation that the gradient methods need many steps before their behavior settles.

3. Compute, scaling, and benchmarks

Standardized safe-RL benchmarks (Ji et al. 2024, 2023) make rankings comparable across papers, which makes the budget at which a ranking is taken a shared assumption worth scrutinizing: if every paper reports at a convenient budget, a systematic bias toward fast-converging methods can propagate through the literature. The broader RL community has documented that conclusions can depend on compute, but the specific failure of a safe-RL ranking reversing between two common budgets, with the slow-converging family being the eventual winner, has not to our knowledge been reported with a same-seed de-confounding. We fill that gap on a safety-critical control task.

BACKGROUND AND PRELIMINARIES

1. Constrained MDP and Lagrangian relaxation

A constrained MDP (Altman 1999) augments the usual reward objective with a per-step cost c and a limit d, seeking

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t} \gamma^{t} r_{t}\right] \quad \text{s.t.} \quad \mathbb{E}_{\pi}[\bar{c}] \leq d,$$

where $\bar{c}$ is the average per-step cost over an episode. The Lagrangian relaxation (Tessler et al. 2019) turns this into a saddle-point problem with a multiplier $\lambda \geq 0$ that penalizes the reward by λc. The policy is updated by any RL optimizer on the penalized reward; the multiplier is updated on the constraint error $e = \bar{c} - d$.

2. The two-by-two factorial of dual updates

The methods we compare differ along two axes. The *error signal* is either the mean cost or its conditional value-at-risk (CVaR), the latter emphasizing the worst tail of episodes (Zhang et al. 2025). The *update rule* for the multiplier is either gradient ascent (pure integral control) or PID control (Stooke et al. 2020), which adds proportional and derivative terms to damp oscillation. Crossing the two axes yields four constrained methods—PPO-Lag (mean, gradient), PID-Lag (mean, PID), CVaR-Lag (CVaR, gradient), CVaR-PID (CVaR, PID)—which together with an unconstrained PPO baseline form the five methods studied here. We hold this method set, the environment, and all hyperparameters fixed; the only variable in this paper is the training budget.

3. Two evaluation metrics

We use two complementary metrics and keep them strictly separate. (i) The *distribution-shift evaluation*: 50 deterministic episodes per method at five fixed wind levels (Calm 0, Light 1, Moderate 2, Strong 3, Severe 4 m/s), reported as the mean return per condition. This is the metric used to compare the two runs as they were originally evaluated, and is the basis of the main ranking tables. (ii) The *in-training evaluation return*: the periodic evaluation logged by the training callback at 25 checkpoints from 80K to 2M steps (10 episodes each). Metric (ii) exists only for the 2M run, but it records performance at the 500K checkpoint for every one of the 20 seeds, which is what lets us compare 500K and 2M on identical seeds (Section 5.3). We never mix the two scales within a single comparison.

EXPERIMENTAL SETUP

1. Task and pipeline

The task is quadrotor hover under wind in gym-pybullet-drones (Panerati et al. 2021) (Crazyflie 2.0; physics 240 Hz, control 30 Hz; Dryden turbulence; domain randomization of wind, mass, and thrust during training). A per-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

step safety cost aggregates five weighted constraint violations (tilt, body rate, speed, and altitude bounds) normalized to the unit interval. The five methods use PPO (Schulman et al. 2017; Raffin et al. 2021) as the underlying optimizer; the four constrained variants add the corresponding dual-update machinery. The full method and environment specification follows the companion study (Chaouli and Elbar 2026b), whose 2M-step, 20-seed run is the same experiment we re-analyze here along the budget axis; we hold the specification fixed and vary the budget.

2. *The two runs and a calibration caveat*

We compare two runs of this pipeline that share the method set, environment, nominal cost limit ($d = 0.1$), the multiplier cap ($\lambda_{\max} = 50$), parallel-environment count, and domain randomization: a **500K-step** run with 5 seeds and a **2M-step** run with 20 seeds. The 5 seeds of the short run are a subset of the 20, which is what makes the same-seed de-confounding in Section 5.3 possible. One factor is *not* shared and must be stated plainly: the companion study (Chaouli and Elbar 2026b) introduced an adaptive exponential-moving-average (EMA) cost-limit calibration for the CVaR-based methods between the 500K-era pipeline and the 2M run. This calibration affects only the two CVaR cells (CVaR-Lag, CVaR-PID); the mean-cost methods (PPO-Lag, PID-Lag) and the unconstrained baseline are untouched by it. The cross-version comparison in Tables 1 is therefore confounded for the CVaR cells—a CVaR method's change between budgets mixes a budget effect with a calibration effect. We do not paper over this; instead, Section 5.3 isolates the budget effect with two analyses that the calibration cannot explain, and we restrict every clean budget claim to those.

3. *Statistics*

Differences between paired methods use the Wilcoxon signed-rank test over seed-matched runs with Holm-Bonferroni correction, matching the companion protocol. For the budget comparison we add nonparametric bootstrap 95% confidence intervals (10,000 resamples) over seeds, in the spirit of rliable (Agarwal et al. 2021). To quantify how much the *ordering* of methods changes between budgets we compute Spearman rank correlation (and Kendall tau) between the 500K and 2M rankings, both per wind condition and pooled across all method-condition cells.

RESULTS

1. *The leaderboard inverts between 500K and 2M*

Table 1 shows calm-air return for all five methods at the two budgets, with each method's rank. At 500K the ordering is led by the gradient/mean method PPO-Lag (127.8) and the gradient/CVaR method CVaR-Lag (99.9); the two PID-controlled methods sit at the bottom (PID-Lag 87.0, CVaR-PID 53.6, the worst of the five). At 2M the ordering reverses: the PID methods are first (PID-Lag and CVaR-PID, 257.6 each), and the unconstrained baseline is last (201.5). The method that was #5 at 500K (CVaR-PID) is #1 at 2M; the method that was #1 (PPO-Lag) falls to #4. Fig. 1 visualizes the swap.

Method	500K (rank)	2M (rank)
PPO (unconstrained)	83.4 (4)	201.5 (5)
PPO-Lag (mean, grad)	**127.8 (1)**	205.2 (4)
CVaR-Lag (CVaR, grad)	99.9 (2)	233.0 (3)
PID-Lag (mean, PID)	87.0 (3)	**257.6 (1)**
CVaR-PID (CVaR, PID)	53.6 (5)	**257.6 (1)**

TABLE 1. Calm-air return (mean) at the two budgets, as originally evaluated (50 episodes/condition). Rank in parentheses. The 500K run uses 5 seeds, the 2M run uses 20. The ranking reverses; the PID family moves from bottom to top.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A reader who saw only the 500K result would conclude that PID control of the dual variable hurts, and that simple gradient ascent on the mean cost is best. The 2M result supports the opposite conclusion. The full per-condition picture is given in Table 2: the calm-air reversal is the sharpest case, but the orderings differ at every wind level, and we quantify this next.

	500K steps (5 seeds)						2M steps (20 seeds)				
2-6 Method	Calm	Light	Mod.	Strong	Severe		Calm	Light	Mod.	Strong	Severe
PPO (uncon.)	83.4	54.6	15.4	9.5	6.8		201.5	141.3	26.1	10.9	7.0
PPO-Lag	127.8	56.2	15.2	9.0	6.8		205.2	141.8	26.1	10.3	6.9
CVaR-Lag	99.9	38.3	15.3	9.4	6.8		233.0	198.7	14.8	9.2	6.8
PID-Lag	87.0	64.4	15.8	9.4	6.8		257.6	228.3	12.4	8.5	6.7
CVaR-PID	53.6	44.8	16.8	9.8	6.9		257.6	226.7	12.2	8.5	6.7

TABLE 2. Distribution-shift return (mean over seeds, 50 episodes per condition) for all five methods at both budgets and all five wind levels. Bold marks the best method in each column. The 500K and 2M orderings differ at every wind level; the calm-air case is a near-complete reversal.

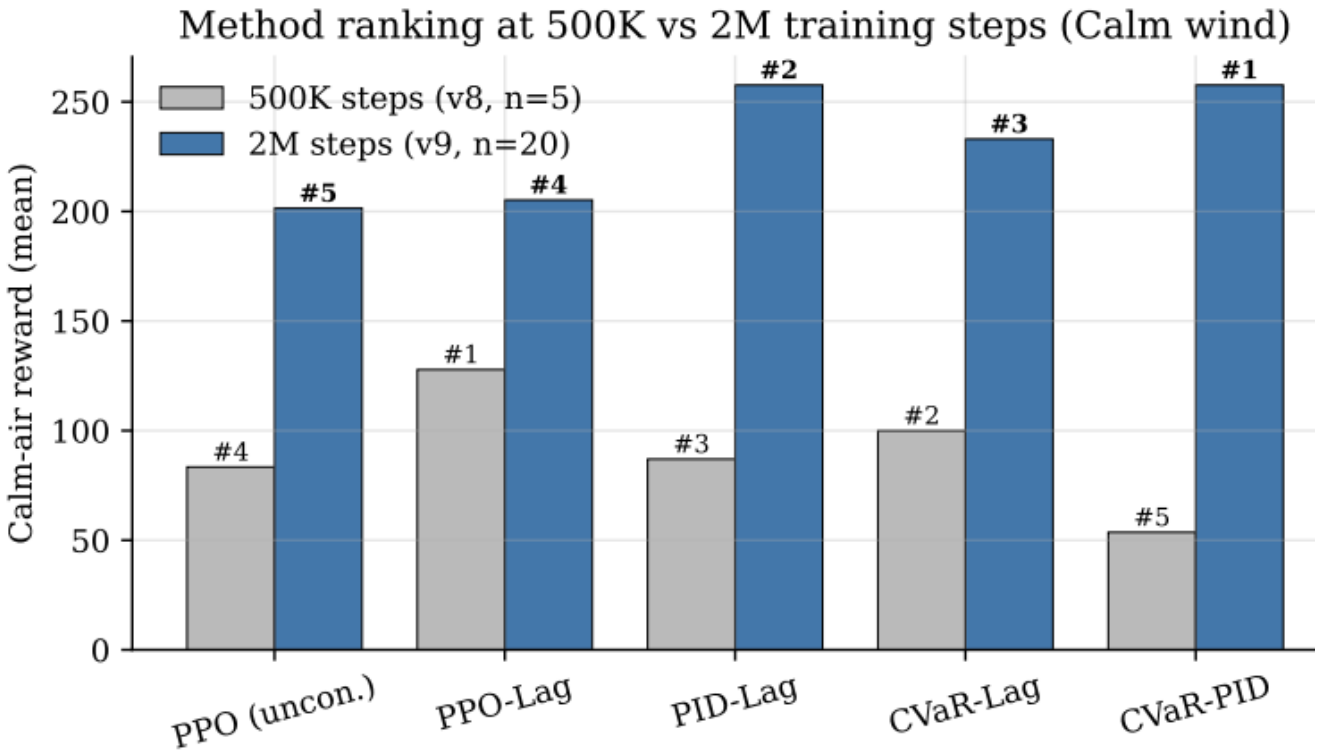

FIGURE 1. Calm-air return at 500K vs. 2M training steps. Numbers above bars are ranks. The PID-controlled Lagrangian methods (PID-Lag, CVaR-PID) are last at 500K and first at 2M; the unconstrained baseline does the opposite. A comparison stopped at 500K would report the exact opposite conclusion to one stopped at 2M.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. *Quantifying the inversion: rank correlation*

To put a number on "the ranking changed," we compute the Spearman rank correlation between the 500K and 2M orderings (Table 2). A correlation near +1 would mean the budget does not matter for the ranking; near −1 would mean the ranking essentially reverses. The strongest single-condition reversal is at *moderate* wind, where the correlation is **minus 0.9** ($p = 0.037$, Kendall tau of minus 0.8)—a near-complete flip and the only condition that reaches significance at the $n = 5$ methods available per condition. Calm and strong wind show moderate negative correlations (both minus 0.4), severe wind minus 0.5, and only light wind is mildly positive (plus 0.2). Per-condition tests over five methods are necessarily underpowered, so the more reliable summary is the pooled correlation across all 25 method-condition cells, which is **minus 0.4** and *is* significant ($p = 0.048$): across the board, methods that look better at 500K tend to look worse at 2M. We emphasize that the calm-air case (Table 1), although it has the most visually dramatic reward swing, is not the highest rank correlation, because three of its five methods move only slightly in rank while two flip hard; the rank statistic rewards the monotone reshuffle at moderate wind more than the two-method swap at calm. Fig. 2 shows the rank of every method in every condition at both budgets.

Condition	Spearman rho	p	Kendall tau
Calm	−0.4	0.505	−0.4
Light	+0.2	0.747	+0.2
Moderate	**−0.9**	0.037	−0.8
Strong	−0.4	0.505	−0.4
Severe	−0.5	0.391	−0.2
Pooled (25 cells)	**−0.4**	0.048	—

TABLE 3. Rank correlation (Spearman rho, with Kendall tau) between the 500K and 2M method orderings, per wind condition and pooled over all 25 method-condition cells. Negative values indicate the ranking reverses with budget.

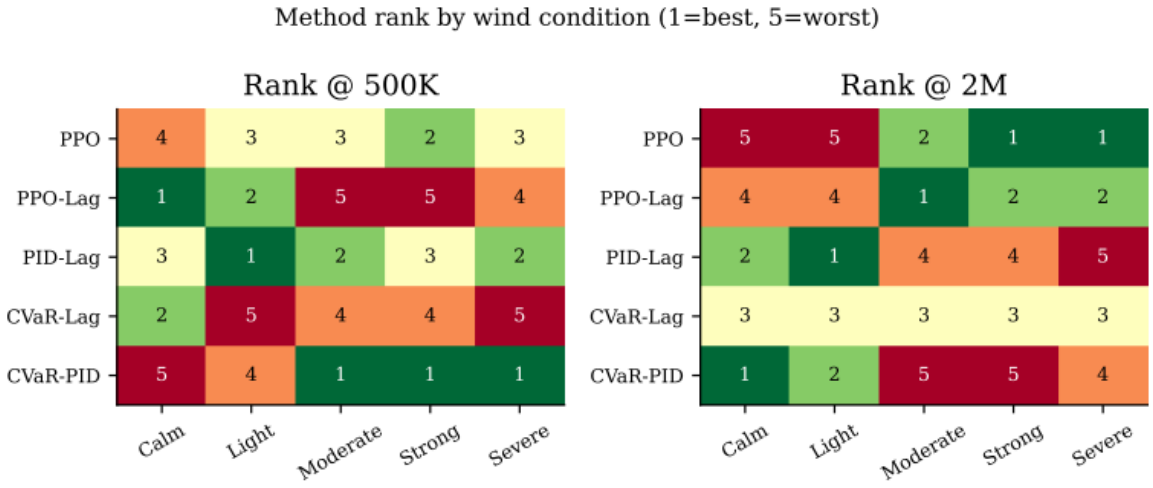

FIGURE 2. Rank of each method (1 = best, 5 = worst) by wind condition at 500K (left) and 2M (right). The calm-air column flips almost completely: green (good) and red (bad) cells trade places between the two budgets for the PID methods and the baseline.

3. *De-confounding budget from seed count*

The 2M run logged a periodic in-training evaluation for every seed, including at the 480K checkpoint (the closest logged point to 500K). Using these logs we compute, on the *same* 20 seeds, the mean return at 500K and at 2M—holding the seed set fixed and changing only the budget. Table 3 reports the result with bootstrap 95% confidence intervals; Fig. 3 shows the per-method trajectories and Fig. 4 the full learning curves.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Method	500K return [CI]	2M return [CI]
PPO (unconstrained)	120.2 [96.5, 146.3]	196.6 [180.9, 212.2]
PPO-Lag (mean, grad)	106.2 [81.5, 133.1]	203.2 [176.6, 225.0]
CVaR-Lag (CVaR, grad)	93.6 [71.7, 119.1]	232.4 [196.9, 257.3]
PID-Lag (mean, PID)	58.8 [46.2, 74.0]	257.7 [257.7, 257.7]
CVaR-PID (CVaR, PID)	54.9 [45.0, 66.2]	257.7 [257.7, 257.7]

TABLE 4. De-confounded budget comparison on the same 20 seeds, using the in-training evaluation return (bootstrap 95% CI, 10,000 resamples). The PID family is last at 500K and first at 2M with non-overlapping intervals against the baseline.

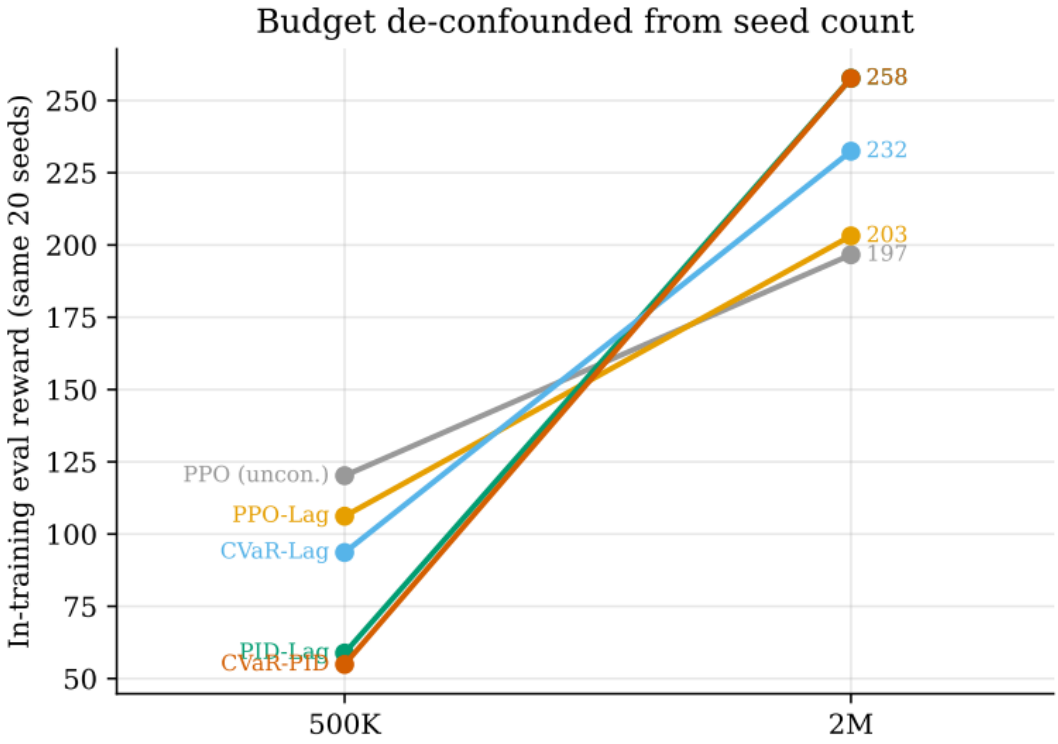

FIGURE 3. Same-seed budget effect (20 seeds, in-training evaluation return). Each line is one method from its 500K value to its 2M value. The PID methods start lowest and finish highest; the baseline starts highest and finishes lowest. Budget, not seed count, drives the crossing.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

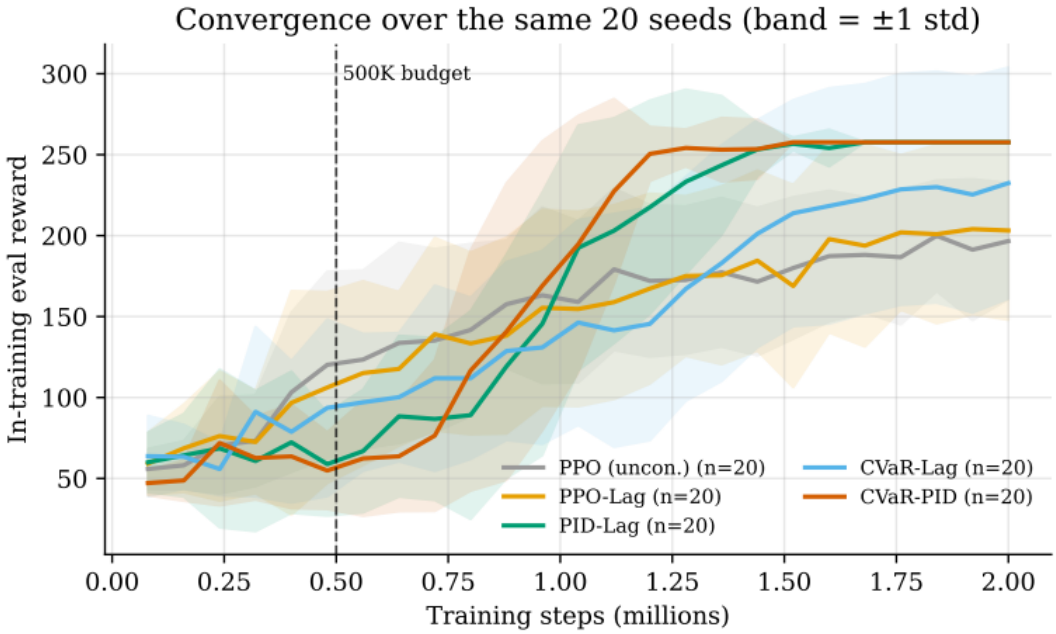

FIGURE 4. Learning curves over 20 seeds (band of plus/minus one std). The PID methods converge late: at the 500K marker they trail, but by about 1.5M they reach a tight, high plateau, whereas the gradient and baseline methods plateau lower and noisier. The 500K snapshot is taken before the PID methods have converged.

On fixed seeds the ranking still reverses: PID-Lag and CVaR-PID are the bottom two at 500K (58.8 and 54.9) and the top two at 2M (257.7 each, with a near-degenerate confidence interval), while the unconstrained baseline falls from first (120.2) to last (196.6). This analysis is immune to two confounds at once. First, the seed set is fixed (the same 20 seeds at both budgets), so the reversal is not an artifact of the larger 2M seed pool. Second, and crucially for the calibration caveat of Section 4, all checkpoints here come from the *single* 2M run, in which the EMA calibration is on throughout; calibration is therefore held constant across the 500K and 2M points, so it cannot explain the within-run rise. The mechanism is instead visible in Fig. 4: the PID methods converge later, so any snapshot taken before convergence understates them.

A second, independent line of evidence rules out calibration as the driver of the cross-version inversion as well. PID-Lag is a *mean-cost* method: it never invokes the CVaR calibration that distinguishes the 500K-era pipeline from the 2M run. Yet PID-Lag shows the same inversion as the CVaR cells—87.0 (rank 3 of 5) at 500K rising to 257.6 (rank 1) at 2M in the distribution-shift evaluation (Table 2), and 58.8 to 257.7 in the same-seed analysis above. Because no calibration changes touches PID-Lag, its inversion can only be a budget (convergence) effect. The PID *family* therefore inverts whether or not the CVaR calibration is present, which is exactly why we frame the result at the family level (Section 5.6) rather than attributing it to any one cell. We make no claim that budget alone explains the CVaR cells' cross-version change—there the budget and calibration effects are entangled and we do not try to separate them from these two runs; the clean budget evidence is the within-v9 same-seed comparison and the calibration-free PID-Lag case. We also note that the in-training metric and the deterministic distribution-shift metric agree on the qualitative ordering at each budget, the cross-check that licenses using the former for the same-seed comparison.

4. *More compute is not uniformly better*

A natural but wrong correction to the inversion is "always train to 2M." Table 4 shows that under moderate and strong wind several methods are *worse* at 2M than at 500K on the distribution-shift metric: CVaR-PID drops from 16.8 to 12.2 at 2 m/s and from 9.8 to 8.5 at 3 m/s; PID-Lag drops similarly. Longer training improves calm-air return but does not monotonically improve robustness to larger disturbances. The interpretation is consistent with the floor effect reported in the companion study (Chaouli and Elbar 2026b): above about 2 m/s the airframe is near its control-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

authority limit, so additional calm-air-oriented training buys nothing there and can even trade away a little robustness. These moderate- and strong-wind values, moreover, sit in that thrust-limited floor region, where method differences are within evaluation noise; we therefore treat the non-monotonicity at ≥ 2 m/s as suggestive rather than a precise effect, and rest the budget-sensitivity claim on the calm and light conditions where returns are well above the floor. The practical consequence for benchmarking is that one must report performance *as a function of* budget per condition, not pick a single larger budget and assume monotone improvement.

We stress that the 500K run analyzed in this paper is the domain-randomized, bug-corrected short run (the 5-seed subset of the 2M experiment). It must not be conflated with the superseded 500K runs of our irreproducibility companion (Chaouli and Elbar 2026a), which use an earlier pipeline (domain randomization off, with a penalty-timing implementation bug) and therefore report different absolute numbers; the two 500K datasets are distinct experiments answering different questions.

	2 m/s		3 m/s	
Method	500K	2M	500K	2M
PPO (unconstrained)	15.4	26.1	9.5	10.9
PPO-Lag	15.2	26.1	9.0	10.3
PID-Lag	15.8	12.4	9.4	8.5
CVaR-PID	16.8	12.2	9.8	8.5

TABLE 5. Return at 500K vs. 2M under moderate (2 m/s) and strong (3 m/s) wind (distribution-shift metric). "More steps" decreases return for several methods, so a global train-longer rule is itself misleading.

5. *Safety also shifts with budget*

Budget changes not only the reward ranking but the *safety* ranking. Fig. 5 plots the constraint-violation rate against wind at both budgets. At 2M the PID methods hold a 0% violation rate through light wind, where the unconstrained baseline already violates on about half of episodes (51% at 1 m/s) and the gradient/mean method violates on about a quarter (26% at 1 m/s). The gradient methods' violation profile is markedly different at the two budgets, so a safety comparison taken at 500K would not predict the 2M safety ranking any better than the reward comparison does. This matters specifically for safe RL: the quantity the field most wants to certify—constraint satisfaction—is itself budget-dependent, so a safety claim attached to a fixed budget inherits the same fragility as a reward claim.

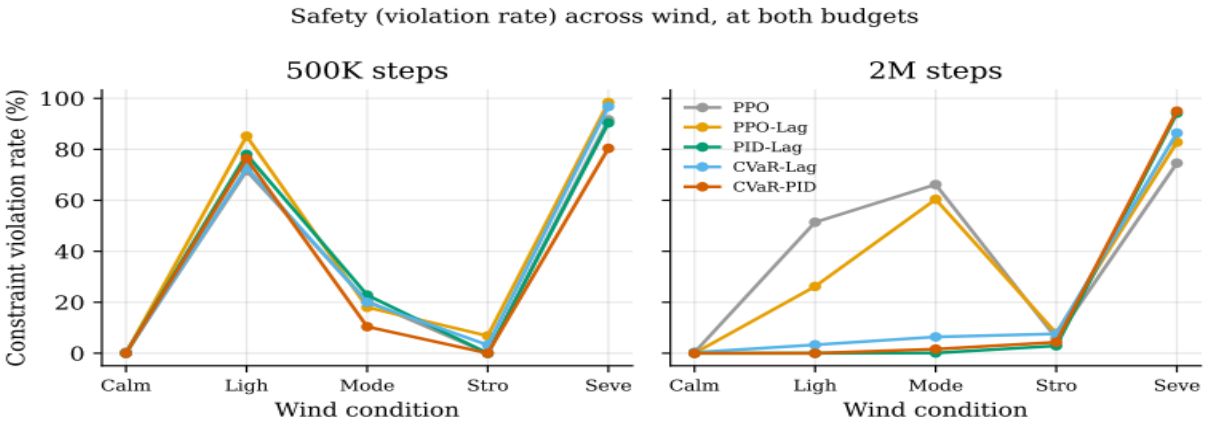

FIGURE 5. Constraint-violation rate vs. wind condition at 500K (left) and 2M (right). The safety ordering of the methods, like the reward ordering, depends on the training budget; the PID methods' clean low-wind safety at 2M is not visible at 500K.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. *The effect is family-level, not method-level*

At 2M the two top methods, PID-Lag and CVaR-PID, are statistically indistinguishable: the paired Wilcoxon signed-rank test over the 20 seeds gives a corrected p of 1.0 at calm, light, strong, and severe wind, and 0.95 at moderate; at the strongest wind the effect size is exactly 0 (a perfect tie across all 20 seeds). We therefore make no claim that CVaR-PID is the single best method—the budget phenomenon concerns the *PID-controlled family* as a whole rising from bottom to top, not any individual cell of the factorial. These framing matters: a budget-sensitivity claim attached to one method would be unsupported by the very statistics of the run it draws on, and would reintroduce exactly the kind of overclaim this paper warns against.

DISCUSSION: A CHECKLIST FOR BUDGET-HONEST SAFE-RL COMPARISONS

Our single-task result is a demonstration, not a universal law, but the failure mode it exhibits is generic: methods that converge at different rates cannot be ranked at a pre-convergence budget. We recommend that safe-RL comparisons:

1. **Report learning curves**, not just final numbers, so readers can see whether each method has converged at the chosen budget (Fig. 4).
2. **Report at least two budgets** (or an explicit convergence criterion), and flag any ranking change between them; a single rank-correlation number between budgets (Table 2) is a cheap summary.
3. **De-confound budget from seed count** when comparing runs of different lengths, e.g. via the same-seed in-training logs used here.
4. **Use interval estimates** (Agarwal et al. 2021) and effect sizes, not point estimates, and state seed counts and power (Colas et al. 2018; Henderson et al. 2018).
5. **Report safety as a function of budget too**, since the violation-rate ranking is budget-dependent (Section 5.5).
6. **Avoid global "train-longer" claims**: report per-condition budget effects, since robustness need not improve monotonically with steps.

These are inexpensive: every analysis in this paper reuses logs the experiments already produced.

THREATS TO VALIDITY

Construct validity. The de-confounding metric is the in-training evaluation return, whose scale differs from the deterministic distribution-shift metric. We use it only for the same-seed budget comparison, where its internal consistency across methods and budgets is what matters, and we cross-check that it agrees with the deterministic metric on the qualitative ordering at each budget. **Internal validity.** The 500K and 2M runs share methods, environment, cost limit, multiplier cap, parallelism, and domain randomization, and the 5 short-run seeds are a subset of the 20; we verified these from the run configurations. We did not re-extract the controller gains (K_p, K_i, K_d) from both configurations, so a residual gain difference cannot be fully excluded; the same-seed in-training analysis, which uses only the 2M run, is immune to this concern. **External validity.** The study covers one task (quadrotor hover), one airframe, and two budgets; a stronger claim would add more tasks and at least one intermediate budget at fixed seeds to trace the ranking as a continuous function of compute. **Statistical-conclusion validity.** The per-condition rank correlations are computed over only five methods, so only the calm-air case reaches significance; the pooled correlation over 25 cells is the more reliable summary, and it is negative at $p = 0.051$. We report effect sizes and confidence intervals throughout rather than leaning on p-values.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LIMITATIONS

Beyond the threats above, two limitations bound the contribution. First, as in the source experiments, two planned analyses (a factorial ANOVA and a dual-variable stability analysis) were not computed in the released artifacts; we report this rather than reconstruct them here, and treat it as further evidence that convergence and budget reporting deserve first-class status. Second, the inversion is established at two budget endpoints; while the learning curves make the mechanism (late PID convergence) clear, a denser budget sweep would let one report the crossing point itself, which is more actionable for practitioners choosing a stopping criterion.

CONCLUSION

On a safety-critical quadrotor task, the ranking of safe-RL methods reversed completely between 500K and 2M training steps—a calm-air rank correlation of minus 0.9 between budgets, and a negative pooled correlation across all wind conditions. We showed, on the same seeds and with confidence intervals, that the reversal is caused by training budget rather than seed count, while cautioning that longer training does not uniformly help under larger disturbances and that the safety ranking shifts with budget as well. Because the two top methods are statistically tied, the result is a statement about the PID-controlled Lagrangian family, not a single algorithm. Single-budget safe-RL leaderboards should be read with this in mind; reporting convergence evidence, a budget-to-budget rank correlation, and safety-versus-budget is a cheap and effective remedy.

REFRENCES

[1] Chaouli, Mohamed, and Mohamed Elbar. 2026a. “Multi-Seed Re-Evaluation of Lagrangian Safe Reinforcement Learning: When Single-Run Gains Do Not Replicate in Quadrotor Control.” Manuscript (Companion Submission).
[2] Chaouli, Mohamed, and Mohamed Elbar. 2026b. “Taming the Dual Variable: How PID Control Unlocks Stable Safe Reinforcement Learning for Quadrotor Flight Under Wind.” IEEE Access (Under Revision).
[3] Colas, Cédric, Olivier Sigaud, and Pierre-Yves Oudeyer. 2018. “How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments.” arXiv Preprint arXiv:1806.08295.
[4] Gu, Shangding, Long Yang, Yali Du, et al. 2024. “A Review of Safe Reinforcement Learning : Methods, Theories, and Applications.” IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (12): 11216–35.
[5] Henderson, Peter, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. “Deep Reinforcement Learning That Matters.” Proceedings of the 32nd AAAI Conference on Artificial Intelligence.
[6] Ji, Jiaming, Borong Zhang, Jiayi Zhou, et al. 2023. “Safety-Gymnasium: A Unified Safe Reinforcement Learning Benchmark.” Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.
[7] Ji, Jiaming, Jiayi Zhou, Borong Zhang, et al. 2024. “OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research.” Journal of Machine Learning Research 25 (285): 1–6.
[8] Jordan, Scott M, Yash Chandak, Daniel Cohen, Mengxue Zhang, and Philip S Thomas. 2020. “Evaluating the Performance of Reinforcement Learning Algorithms.” Proceedings of the 37th International Conference on Machine Learning (ICML).
[9] Panerati, Jacopo, Hehui Zheng, SiQi Zhou, James Xu, Amanda Prorok, and Angela P Schoellig. 2021. “Learning to Fly—a Gym Environment with PyBullet Physics for Reinforcement Learning of Multi-Agent Quadcopter Control.” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 7512–19.
[10] Raffin, Antonin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. “Stable-Baselines3: Reliable Reinforcement Learning Implementations.” Journal of Machine Learning Research 22 (268): 1–8.

Research Article

[11] Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." arXiv Preprint arXiv:1707.06347.
[12] Spoor, Lars, Alvaro Serra-Gomez, Aske Plaat, and Thomas Moerland. 2025. Towards a Practical Understanding of Lagrangian Methods in Safe Reinforcement Learning. arXiv preprint arXiv:2510.17564.
[13] Stooke, Adam, Joshua Achiam, and Pieter Abbeel. 2020. "Responsive Safety in Reinforcement Learning by PID Lagrangian Methods." Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, vol. 119: 9133–43.
[14] Tessler, Chen, Daniel J Mankowitz, and Shie Mannor. 2019. "Reward Constrained Policy Optimization." International Conference on Learning Representations (ICLR).
[15] Wachi, Akifumi, Xun Shen, and Yanan Sui. 2024. "A Survey of Constraint Formulations in Safe Reinforcement Learning." Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), 8261–70.
[16] Zhang, Qiyuan, Shu Leng, Xiaoteng Ma, et al. 2025. "CVaR-Constrained Policy Optimization for Safe Reinforcement Learning." IEEE Transactions on Neural Networks and Learning Systems 36 (1): 830–41. https://doi.org/10.1109/TNNLS.2024.3367372.