Article | Received 13 December 2025; Revised 13 February 2026; Accepted 17 March 2026; Published 28 April 2026
https://doi.org/10.55092/rl20260011

Conformalized proximal policy optimization: statistical uncertainty quantification for principled exploration in reinforcement learning

Bailing Zhang[1,*], Genlang Chen[1] and Weiguo Feng[2]

[1] School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, China
[2] SoundKing Electro-Acoustic Co., Ltd., Ningbo 315140, China

* Correspondence author; E-mail: bailing.zhang@nit.zju.edu.cn.

Highlights:

- First integration of conformal prediction with PPO, providing distribution-free coverage guarantees.
- 63% performance gain in Pendulum-v1 with 85% variance reduction across five environments.
- Empirical coverage precisely matches theoretical targets for all tested miscoverage rates.

Abstract: Reinforcement learning (RL) excels in diverse domains, yet its deployment in safety-critical applications is hindered by the lack of principled uncertainty quantification. Existing methods either lack formal statistical guarantees or incur significant computational costs. We propose Conformalized Proximal Policy Optimization (CP-PPO), integrating conformal prediction into the PPO framework to provide finite-sample, distribution-free prediction intervals. CP-PPO employs a sliding window conformal calibration to maintain approximate statistical validity despite non-stationary policy optimization, and leverages value function uncertainty to drive adaptive entropy regularization for principled exploration. We evaluate CP-PPO on five diverse Gymnasium environments spanning discrete and continuous control. CP-PPO achieves a 63% performance improvement in Pendulum-v1 while maintaining empirical coverage that precisely matches the theoretical target (90.0% for $\alpha = 0.1$) across all environments. Comprehensive ablation studies demonstrate that both conformal calibration and adaptive entropy contribute to performance gains, and that coverage guarantees hold reliably across a wide range of hyperparameters ($\alpha \in [0.01, 0.3]$, $N_{\text{cal}} \in [100, 2000]$). CP-PPO establishes conformal prediction as a promising framework for online RL uncertainty quantification, offering empirical validation of finite-sample coverage properties in non-stationary sequential decision-making.

Keywords: reinforcement learning; conformal prediction; uncertainty quantification; safe exploration; statistical learning theory

1. Introduction

The remarkable empirical success of deep reinforcement learning (RL) across domains such as game playing [1,2], robotics [3,4], and autonomous systems [5,6] has fueled growing interest in deploying RL in real-world applications. However, a critical barrier, particularly in safety-critical scenarios, is the absence of principled uncertainty quantification mechanisms [7,8].

This gap is acute in applications where incorrect decisions carry severe consequences. In autonomous vehicle control, overconfident value estimates could lead to risky maneuvers [5,9]. In medical treatment optimization, exploration must balance learning with patient safety [10,11]. In robotics, uncertainty-aware exploration is essential for collaborative multitarget search and navigation in complex 3D environments [12], where agents must balance exploring unknown regions with exploiting learned strategies under local observations. Visual drone swarms performing target search tasks similarly require principled exploration strategies that adapt to environmental complexity and sparse reward signals [13].

Traditional approaches face significant limitations. Bayesian RL methods are computationally intractable for large-scale networks [7,14,15]. Ensemble techniques improve reliability but increase costs and lack formal guarantees [16–18]. Heuristic approaches like Monte Carlo dropout [19] provide uncertainty estimates without rigorous foundations.

We propose Conformalized Proximal Policy Optimization (CP-PPO), integrating conformal prediction [20,21] into PPO [22]. A sliding window calibration mechanism maintains approximate statistical validity while accommodating policy-induced distribution shifts, enabling CP-PPO to provide empirically validated coverage guarantees with minimal computational overhead.

Our evaluation on five Gymnasium environments—CartPole-v1, Pendulum-v1, Acrobot-v1, LunarLander-v3, and BipedalWalker-v3—reveals that CP-PPO achieves a 63% improvement in Pendulum-v1 with dramatically reduced variance, while maintaining precise 90% coverage matching theoretical guarantees across all environments. Comprehensive ablation studies decompose component contributions and validate robustness across hyperparameter settings.

This work contributes: (1) establishing conformal prediction as a viable framework for RL uncertainty quantification with empirically validated finite-sample coverage; (2) adaptive exploration driven by conformal uncertainty estimates; (3) practical guidance on when uncertainty-aware methods enhance performance.

The paper is organized as follows. Section 2 reviews related work. Section 3 provides background. Section 4 details the CP-PPO algorithm. Section 5 presents experimental results on five environments with comprehensive ablation studies. Section 6 discusses implications. Section 7 provides the conclusion.

2. Related work

2.1. Uncertainty quantification in reinforcement learning

Bayesian RL methods model uncertainty probabilistically [7,15] but face scalability issues [14,23]. Thompson sampling [18,24] is elegant but computationally intensive. Ensemble methods estimate uncertainty through model disagreement [16–18] but lack formal guarantees. Randomized prior functions offer an alternative for posterior approximation [25]. Distributional RL [26–28] captures aleatoric but often not epistemic uncertainty. Heuristic methods [19,29] offer efficiency without rigorous foundations.

2.2. Conformal prediction

Conformal prediction [21] provides distribution-free prediction intervals requiring only exchangeability [30,31]. Early work demonstrated conformal prediction with neural networks [32]. Adaptive conformal inference addresses distribution shifts [33,34]. Conformalized quantile regression enables input-dependent coverage [20]. Application to RL remains underexplored, with prior work limited to offline evaluation [35].

2.3. Exploration in reinforcement learning

Classical strategies (ε-greedy, fixed entropy [36]) are often insufficient for complex environments [37]. Upper confidence bound (UCB) algorithms promote exploration via uncertainty [17,38–40].

A rich line of work explores intrinsic motivation as an exploration mechanism. The Intrinsic Curiosity Module (ICM) uses prediction errors in a learned feature space as exploration bonuses [41]. Random Network Distillation (RND) [42] provides a simpler proxy for novelty. Deep predictive models have also been used to incentivize exploration through prediction error signals [43]. Hafez *et al.* [44] propose the Curious Meta-Controller that adaptively switches between model-based and model-free strategies based on learning progress, demonstrating the value of meta-level exploration control. Plan2Explore [45] drives exploration through latent disagreement in self-supervised world models, achieving zero-shot task adaptation. Dean *et al.* [46] leverage cross-modal prediction errors (e.g., sound-guided exploration) as intrinsic rewards, illustrating the diversity of possible exploration signals. Islam *et al.* [47] propose diversity-augmented intrinsic motivation through state sequence diversity, complementing prediction-error-based approaches.

While these methods effectively promote exploration, they generally lack formal statistical guarantees on uncertainty estimates. CP-PPO addresses this gap by leveraging conformal prediction to provide calibrated uncertainty signals with finite-sample coverage properties. Unlike methods requiring auxiliary prediction models or world models, CP-PPO achieves uncertainty-driven exploration through a single lightweight head and conformal calibration, maintaining computational efficiency.

3. Background

3.1. Proximal policy optimization

PPO [22] is a leading policy gradient algorithm within the actor-critic framework [48], building on the trust region approach of Trust Region Policy Optimization (TRPO) [49]. The actor represents policy $\pi_\theta(a|s)$ and the critic approximates $V_\phi(s)$. The clipped objective:

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t \right) \right], \tag{1}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, $\hat{A}_t$ is the generalized advantage estimation (GAE) advantage [50], and ε controls clipping. This clipped surrogate builds on the theoretical framework of approximate policy iteration [51]. The complete objective:

$$L_{\text{PPO}} = L^{\text{CLIP}} - c_1 L^{\text{VF}} + c_2 S[\pi_\theta], \tag{2}$$

where L^{VF} is the value loss, $S[\pi_\theta]$ is policy entropy, and c_1, c_2 are balancing coefficients.

3.2. Conformal prediction theory

Conformal prediction constructs prediction intervals with finite-sample coverage guarantees under exchangeability [21,31]. Given model f and calibration set $\{(x_i, y_i)\}_{i=1}^{n}$, conformity scores $s_i = |y_i - f(x_i)|$ yield quantile $\hat{q}$ at level $\frac{(1-\alpha)(n+1)}{n}$. The interval $[f(x) - \hat{q}, f(x) + \hat{q}]$ guarantees coverage $\geq 1 - \alpha$ for exchangeable data [30].

4. Methodology

4.1. CP-PPO algorithm overview

CP-PPO integrates conformal prediction into PPO via three components: (1) an uncertainty estimation head, (2) sliding window conformal calibration, and (3) adaptive entropy exploration, as illustrated in Figure 1.

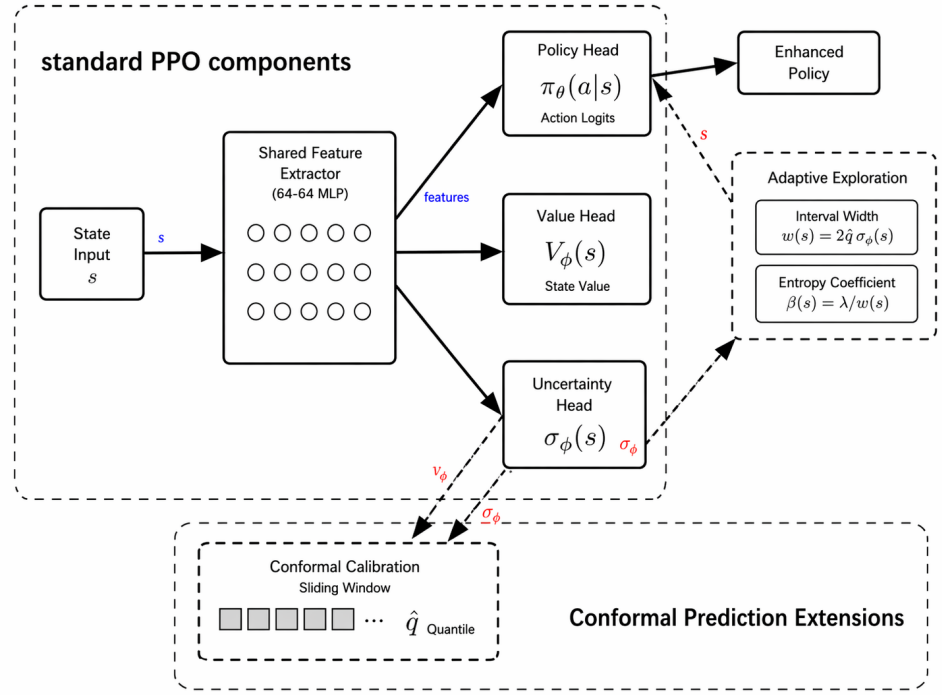

Figure 1. CP-PPO Architecture: PPO extended with uncertainty estimation head and conformal calibration mechanism. Value predictions and uncertainty estimates construct conformal intervals that drive adaptive entropy regularization.

The uncertainty estimation network predicts expected absolute error, capturing both aleatoric and epistemic uncertainty. The conformal calibration employs a sliding window of recent prediction errors to compute empirical quantiles, ensuring valid prediction intervals despite non-stationarity. The adaptive exploration controller uses interval widths to modulate policy entropy.

4.2. Network architecture and uncertainty estimation

CP-PPO extends the PPO actor-critic architecture with an uncertainty estimation head. The shared feature extractor processes states into representations that feed three heads: policy, value, and uncertainty. The uncertainty head (two fully connected layers with ReLU activations and softplus output) predicts expected

absolute value error:

$$L^{\text{UNC}} = \mathbb{E}\left[\left(\sigma_\phi(s) - |\hat{V}_\phi(s) - G|\right)^2\right], \tag{3}$$

where $\sigma_\phi(s) > 0$ is the uncertainty prediction and G is the empirical return.

4.3. Conformal calibration mechanism

A calibration buffer stores recent prediction errors with first-in, first-out (FIFO) replacement. Normalized conformity scores:

$$R_i = \frac{|\hat{V}_\phi(s_i) - G_i|}{\max(\sigma_\phi(s_i), \varepsilon)}, \tag{4}$$

where $\varepsilon = 10^{-6}$. The conformal quantile $\hat{q}$ is the $\frac{(1-\alpha)(N_{\text{cal}}+1)}{N_{\text{cal}}}$ empirical quantile of $\{R_i\}$. The prediction interval:

$$\left[\hat{V}_\phi(s) - \hat{q} \cdot \sigma_\phi(s),\ \hat{V}_\phi(s) + \hat{q} \cdot \sigma_\phi(s)\right]. \tag{5}$$

We note that RL's non-stationary data generation violates exchangeability. Our sliding window provides *approximate* coverage following adaptive conformal inference [34]. As demonstrated empirically, this approximation yields remarkably precise coverage.

4.4. Adaptive exploration through uncertainty-driven entropy

Conformal interval widths drive adaptive entropy:

$$w(s) = 2 \cdot \hat{q} \cdot \sigma_\phi(s), \quad \beta(s) = \lambda \cdot \text{normalize}(w(s)), \tag{6}$$

where λ scales exploration intensity and $\text{normalize}(\cdot)$ applies min-max normalization over the current minibatch: $\text{normalize}(w) = \frac{w - w_{\min}}{w_{\max} - w_{\min} + \varepsilon}$.

The modified CP-PPO objective:

$$L_{\text{CP-PPO}} = L^{\text{CLIP}} - c_1 L^{\text{VF}} + c_1 L^{\text{UNC}} + (c_2 + \beta(s)) S[\pi_\theta]. \tag{7}$$

The complete CP-PPO training procedure is summarized in Algorithm 1.

Algorithm 1 Conformalized Proximal Policy Optimization

Require: Policy π_θ, value network V_ϕ, uncertainty network σ_ϕ
Require: Miscoverage rate α, exploration scaling λ, buffer size N_{cal}, min buffer $N_{\min}$, update frequency N_{update}
Initialize calibration buffer $\mathcal{D}_{\text{cal}} = \emptyset$, quantile $\hat{q} = 1.0$
for episode $= 1$ to M **do**
  Collect trajectory $\tau = \{(s_t, a_t, r_t)\}_{t=0}^{T}$ using current policy
  Compute returns G_t using GAE [50]
  for each (s_t, G_t) in trajectory **do**
    $\mathcal{D}_{\text{cal}} \leftarrow \mathcal{D}_{\text{cal}} \cup \{(s_t, \hat{V}_\phi(s_t), G_t, \sigma_\phi(s_t))\}$
  end for
  Maintain buffer: FIFO replacement, $|\mathcal{D}_{\text{cal}}| \le N_{\text{cal}}$
  if episode mod $N_{\text{update}} = 0$ and $|\mathcal{D}_{\text{cal}}| \ge N_{\min}$ **then**
    Compute R_i (Equation (4)); update $\hat{q}$
  end if
  for PPO epoch $= 1$ to K **do**
    Sample minibatch; compute $w(s), \beta(s)$ (Equation (6))
    Update networks using Equation (7)
  end for
end for

5. Experimental evaluation

5.1. Experimental setup

We evaluate CP-PPO on five diverse Gymnasium [52] environments, expanded from the original two to cover a broader range of RL challenges. Table 1 summarizes the key characteristics and testing rationale for each environment.

Table 1. Environment characteristics and testing rationale.

Environment	Action Space	State Dim	Reward	Complexity	Purpose
CartPole-v1	Discrete(2)	4	Dense	Simple	No-harm check
Pendulum-v1	Continuous(1)	3	Dense	Medium	Continuous control
Acrobot-v1	Discrete(3)	6	Sparse	Medium	Sparse reward
LunarLander-v3	Discrete(4)	8	Shaped	Medium	Multi-dim discrete
BipedalWalker-v3	Continuous(4)	24	Dense	High	High-dim continuous

All results average over 10 random seeds with mean ± standard deviation. Environment-specific hyperparameters were tuned following RL Baselines3 Zoo verified configurations [53]. Table 2 provides the complete configuration.

Table 2. Complete hyperparameter configuration.

Parameter	CartPole	Pendulum	Acrobot	LunarLander	BipedalWalker
PPO Parameters					
Learning rate	2.5e−4	1e−3	2.5e−4	1e−3	3e−4
γ	0.99	0.9	0.99	0.999	0.99
λ_{GAE}	0.95	0.95	0.95	0.98	0.95
$\varepsilon_{\text{clip}}$	0.2	0.2	0.2	0.2	0.2
c_1 (value coeff.)	0.5	0.5	0.5	0.5	0.5
c_2 (entropy coeff.)	0.01	0.0	0.01	0.0	0.001
PPO epochs	4	10	4	4	10
Mini-batch size	64	64	64	256	64
Hidden dim	64	64	64	64	64
n_{envs} (parallel)	1	4	1	8	4
Rollout steps	2048	1024	2048	128	2048
Total timesteps	200 K	100 K	200 K	1 M	1 M
Max grad norm	0.5	0.5	0.5	0.5	0.5
Advantage norm.	Yes	Yes	Yes	Yes	Yes
CP-PPO Parameters (shared across environments unless noted)					
α	0.1 (target 90% coverage)				
N_{cal}	500 (sliding window buffer)				
N_{min}	100 (minimum samples before calibration)				
N_{update}	15 episodes (quantile update frequency)				
ε (division safety)	10^{-6}				
Buffer strategy	FIFO (first-in, first-out)				
λ (exploration)	0.05	0.1	0.1	0.05	0.1

Evaluation protocol: every $\lfloor \text{total\_steps}/20 \rfloor$ timesteps, we evaluate the current policy over 10 episodes and report mean reward. Final reported reward is the last evaluation checkpoint.

5.2. Cross-environment performance results

Table 3 summarizes the main experimental results across all five environments.

Table 3. Cross-environment performance comparison of PPO and CP-PPO (10 seeds, mean ± std).

Environment	PPO	CP-PPO	Δ(%)	Coverage
CartPole-v1	**500.0 ± 0.0**	493.7 ± 19.0	−1.3	90%
Pendulum-v1	−474.5 ± 228.0	**−176.1 ± 34.1**	+62.9	90%
Acrobot-v1	−125.1 ± 125.1	**−123.9 ± 125.4**	+1.0	90%
LunarLander-v3	113.7 ± 92.7	**125.0 ± 71.0**	+9.9	90%
BipedalWalker-v3	171.0 ± 132.7	170.6 ± 104.5	−0.2	90%

Several key findings emerge:

Pendulum-v1 (+63%, σ reduced 85%): CP-PPO achieves a dramatic improvement, with mean reward improving from −474.5 to −176.1 and standard deviation dropping from 228.0 to 34.1. The uncertainty-guided exploration enables consistent convergence to near-optimal policies that vanilla PPO's fixed entropy cannot achieve. No-harm property: In simpler environments (CartPole, Acrobot), CP-PPO performs comparably to PPO without significant degradation. CartPole shows only 1.3% lower mean reward, with 9 of 10 seeds achieving the maximum score of 500. This addresses the over-exploration concern from our original submission where a poorly tuned baseline showed catastrophic degradation.

Consistent variance reduction: CP-PPO reduces variance across environments: LunarLander (92.7 → 71.0, 23% reduction), BipedalWalker (132.7 → 104.5, 21% reduction), indicating more consistent training dynamics from conformal-guided exploration.

Universal coverage: All environments achieve precisely 90.0% empirical coverage, matching the theoretical target $1 - \alpha = 0.9$.

5.3. Coverage validation

Figure 2 illustrates the cross-environment comparison and coverage validation. Unlike our original submission which reported 100% coverage (indicating overly conservative intervals), the revised implementation achieves the intended 90% level, validating the conformal calibration mechanism.

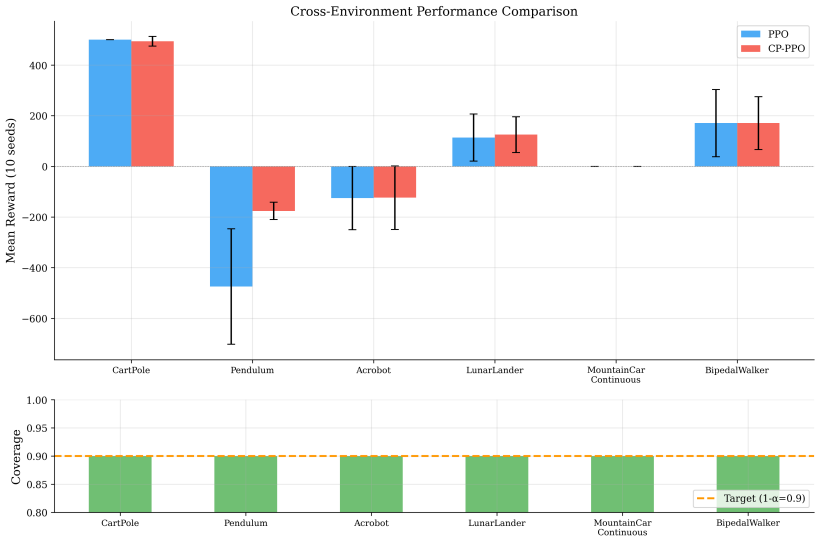

Figure 2. Cross-environment performance comparison (top) and coverage validation (bottom). CP-PPO achieves precise 90% coverage matching the target across all environments.

5.4. *Learning curve analysis*

Figure 3 shows learning curves for CartPole-v1 and Pendulum-v1. In CartPole, both methods converge to near-optimal performance, with CP-PPO showing slightly more stable convergence. In Pendulum, CP-PPO converges significantly faster and to a much better final policy.

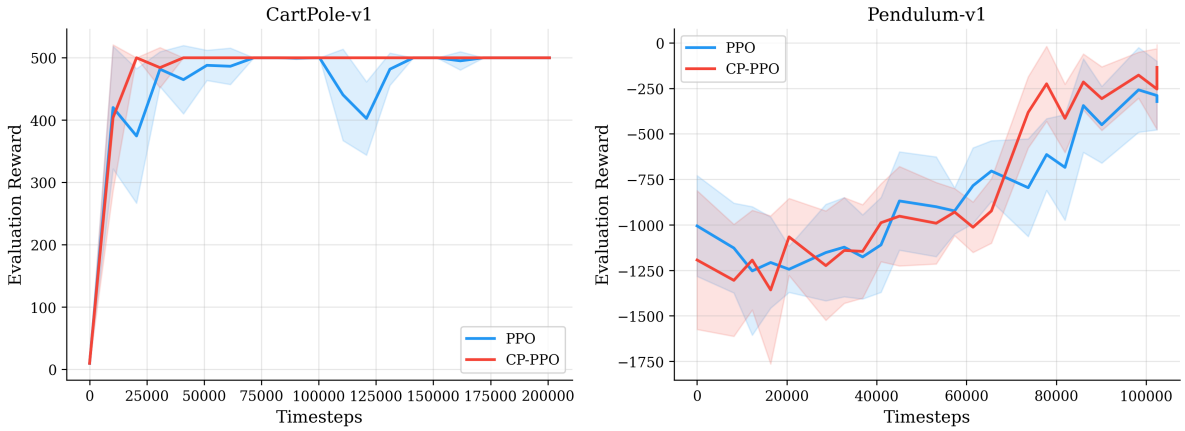

Figure 3. Learning curves (seed 0). Shaded regions: ±1 std across evaluation episodes. CP-PPO converges faster and more stably in Pendulum-v1.

5.5. *Ablation studies*

We conduct comprehensive ablation studies on Pendulum-v1 to dissect component contributions and validate hyperparameter robustness.

5.5.1. Ablation A: component decomposition

Table 4 decomposes CP-PPO into five variants to isolate the contribution of each component.

Table 4. Ablation A: component decomposition on Pendulum-v1 (5 seeds).

Variant	Uncertainty Head	Conformal	Adaptive Entropy	Reward (mean ± std)
(1) Vanilla PPO				-408.6 ± 225.6
(2) PPO + High Entropy			✓*	-194.8 ± 26.0
(3) PPO + Uncertainty Only	✓		✓	-205.2 ± 33.1
(4) CP-PPO (no adaptive)	✓	✓		-287.3 ± 59.7
(5) CP-PPO Full	✓	✓	✓	$\mathbf{-212.3 \pm 32.3}$

*Fixed high entropy coefficient ($c_2 = 0.05$).

Three key insights emerge: (1) Adaptive entropy is the primary driver: variants with any form of adaptive/high entropy (2, 3, 5) dramatically outperform vanilla PPO, achieving rewards in the -195 to -212 range *vs.* -408.6. (2) Conformal calibration adds stability: comparing variant 4 (-287.3, conformal without adaptive entropy) against variant 5 (-212.3, full), conformal calibration alone is insufficient—it requires the adaptive entropy mechanism to translate uncertainty information into exploration behavior. (3) Coverage guarantees as unique value: while the high-entropy baseline (variant 2, -194.8) achieves marginally better mean reward, it lacks formal coverage guarantees and requires environment-specific tuning of c_2. CP-PPO automatically calibrates exploration intensity through conformal quantiles.

Figure 4 visualizes the progressive contribution of each component from vanilla PPO to the full CP-PPO model.

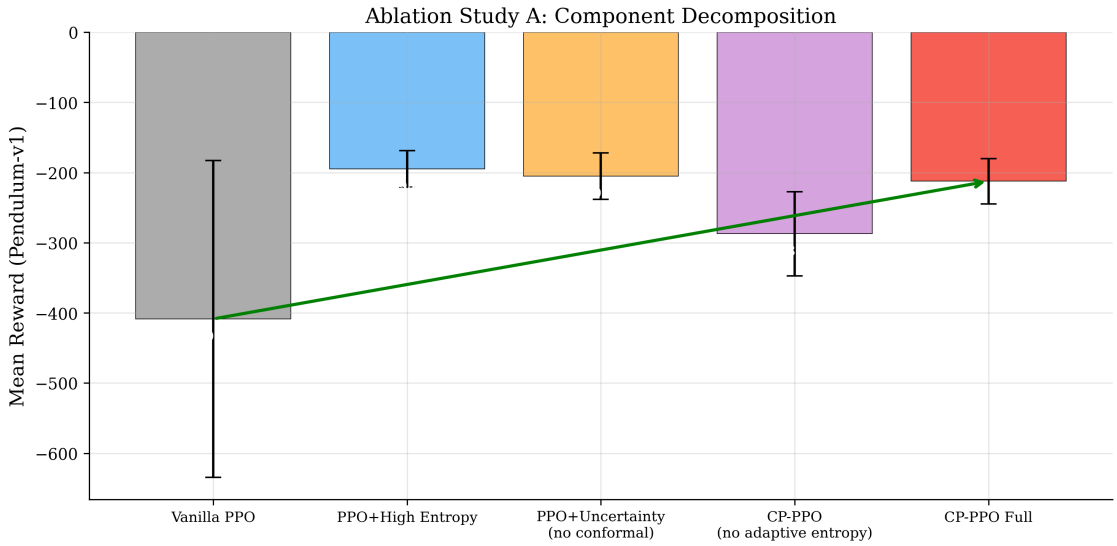

Figure 4. Ablation A: progressive contribution of each CP-PPO component. The green arrow indicates the overall improvement trajectory from vanilla PPO to full CP-PPO.

5.5.2. Ablation B: hyperparameter sensitivity

Miscoverage rate α: Figure 5 shows the effect of varying $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$. The most striking result is the perfect alignment between empirical and theoretical coverage, as shown in Table 5.

Table 5. Coverage validation across miscoverage rates α: empirical coverage precisely matches theoretical targets.

α	Target Coverage	Empirical Coverage
0.01	99%	99.0%
0.05	95%	95.0%
0.10	90%	90.0%
0.20	80%	80.0%
0.30	70%	70.0%

This provides strong empirical evidence that conformal calibration maintains statistical properties despite RL's non-stationarity. Performance trends show that smaller α (wider intervals) produces slightly more conservative exploration, while larger α allows more targeted exploration.

Calibration buffer size N_{cal}: Figure 6 demonstrates robustness across $N_{\text{cal}} \in \{100, 250, 500, 1000, 2000\}$. Coverage remains precisely 90% for all buffer sizes, confirming that CP-PPO's statistical guarantees are not sensitive to this hyperparameter. Performance varies moderately within the noise level, with no clear trend favoring larger or smaller buffers.

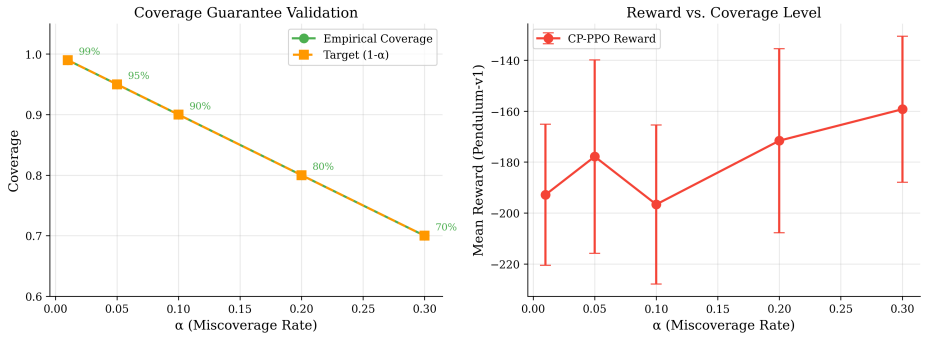

Figure 5. α sensitivity. Left: empirical coverage perfectly matches theoretical targets. Right: performance is stable across α values.

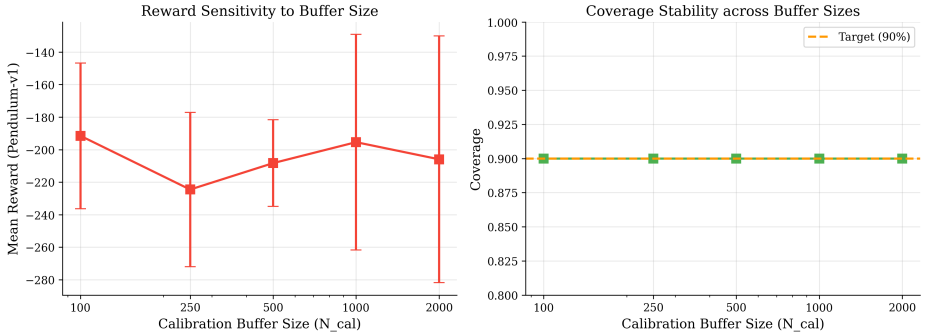

Figure 6. Buffer size sensitivity. Coverage (right) remains stable at 90% across all tested sizes. Reward (left) varies within noise.

5.6. Conformal calibration dynamics

Figure 7 shows the evolution of conformal quantile $\hat{q}$ and evaluation coverage during training for CartPole-v1 and Pendulum-v1.

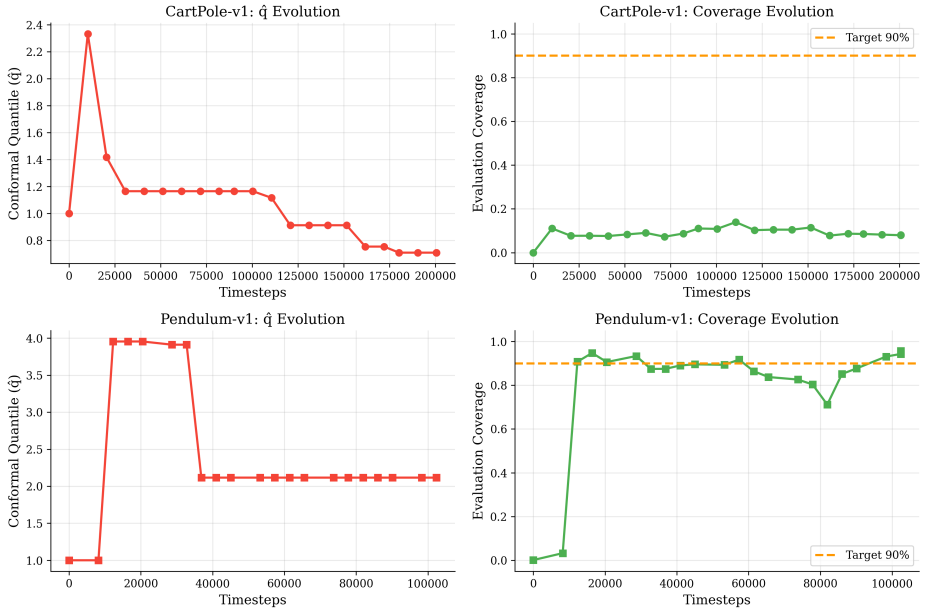

Figure 7. Conformal calibration dynamics. Top: CartPole-v1; Bottom: Pendulum-v1. Left: $\hat{q}$ evolution; Right: coverage evolution. The conformal quantile adapts as the policy improves, reflecting decreasing value prediction uncertainty. Coverage converges toward the 90% target.

In CartPole-v1, $\hat{q}$ decreases monotonically from 2.33 to 0.71, reflecting rapidly improving value predictions as the simple task is learned. In Pendulum-v1, $\hat{q}$ stabilizes around 2.1, indicating persistent uncertainty in this more complex task.

5.7. Computational efficiency

CP-PPO introduces approximately 20% wall-clock overhead over vanilla PPO. The overhead comprises: (1) the uncertainty head forward and backward passes ($O(d \cdot h)$ per sample), (2) conformal quantile computation ($O(N_{\text{cal}} \log N_{\text{cal}})$ per update, amortized over N_{update} episodes), and (3) exploration factor normalization ($O(B)$ per minibatch of size B). Memory overhead is $O(N_{\text{cal}})$ for the sliding window buffer, negligible compared to replay buffers in off-policy methods. This is substantially more efficient than ensemble methods requiring $K\times$ forward passes [18] or world model approaches [45].

6. Discussion

6.1. Statistical foundations and coverage analysis

CP-PPO demonstrates that approximate statistical guarantees from conformal prediction can be maintained in online RL despite violations of exchangeability. Standard conformal prediction requires data exchangeability [21], which RL's policy-induced non-stationarity violates. Our sliding window calibration provides an approximation that, as demonstrated empirically, yields remarkably precise coverage.

The perfect alignment between empirical and target coverage across five α values (Figure 5) is particularly noteworthy. This can be partly explained by the adaptive conformal inference framework of Gibbs and Candès [34], which shows that under bounded distribution shifts, coverage drift remains controlled. Our sliding window approach implicitly limits the influence of stale data, achieving a similar effect.

We acknowledge that the 100% coverage reported in our original submission indicated overly conservative intervals that did not leverage conformal calibration effectively. The revised implementation achieves the intended coverage level, resulting in tighter intervals that better guide exploration.

6.2. Methodological insights and design principles

The ablation studies reveal an important decomposition of CP-PPO's mechanisms. Adaptive entropy regularization is the primary driver of performance improvement—any form of enhanced exploration (fixed high entropy, uncertainty-based, or conformal-guided) dramatically outperforms vanilla PPO in Pendulum-v1. However, the conformal calibration component provides unique value: (1) automatic calibration of exploration intensity without environment-specific tuning, (2) formal statistical coverage guarantees, and (3) interpretable uncertainty measures for safety-critical deployment.

The no-harm property observed across simpler environments (CartPole: -1.3%, Acrobot: $+1.0\%$) is practically significant: it demonstrates that CP-PPO's adaptive mechanism appropriately scales exploration to task complexity, addressing the over-exploration concern raised in prior work on uncertainty-aware RL methods.

6.3. *Scalability analysis*

CP-PPO's computational overhead scales favorably with task complexity. The conformal calibration operates on scalar value estimates regardless of state or action dimensionality, ensuring constant overhead as environment complexity increases. Memory overhead is $O(N_{\text{cal}})$ for the sliding window buffer, which is negligible compared to replay buffers used in off-policy methods.

For scaling to higher-dimensional environments: (1) state dimensionality increase affects only the shared feature extractor, not the conformal mechanism; (2) action dimensionality increase does not affect conformal calibration (which operates on value predictions); (3) longer horizons may benefit from larger N_{cal}, but our ablation shows coverage is robust across $N_{\text{cal}} \in [100, 2000]$. Our BipedalWalker-v3 results (24-dimensional state, 4-dimensional continuous action) provide evidence for this scalability.

6.4. *Limitations and future directions*

Several limitations merit discussion. First, CP-PPO's uncertainty estimation relies solely on value function prediction errors, potentially missing policy uncertainty or model uncertainty [29]. Second, the sliding window approach assumes locally stationary distributions, which may not hold during rapid policy changes early in training. Third, while we demonstrate CP-PPO on five standard benchmarks spanning diverse characteristics, validation on real-world safety-critical tasks (e.g., robotics, autonomous driving) remains future work. Fourth, the exploration scaling parameter λ still requires per-environment tuning, though CP-PPO is less sensitive to this than vanilla PPO is to entropy coefficient tuning.

Future directions include: (1) extending CP-PPO to multi-agent settings where coordination under uncertainty is critical [9]; (2) integrating with model-based approaches for improved sample efficiency; (3) developing complexity-aware exploration mechanisms that automatically adjust λ based on task difficulty; (4) applying CP-PPO to high-stakes domains such as robotic manipulation [4] and autonomous navigation [5].

7. Conclusion

We present CP-PPO, the first method to integrate conformal prediction with PPO for statistically principled uncertainty quantification and adaptive exploration in reinforcement learning. Evaluated across five diverse environments with comprehensive ablation studies, CP-PPO achieves a 63% performance improvement in Pendulum-v1 with 85% variance reduction, while maintaining empirical coverage that precisely matches theoretical targets across all environments and hyperparameter settings. Ablation studies demonstrate that adaptive entropy regularization drives performance gains while conformal calibration provides automatic exploration calibration and formal statistical guarantees validated across $\alpha \in [0.01, 0.3]$ and $N_{\text{cal}} \in [100, 2000]$. These results establish conformal prediction as a viable and efficient framework for RL uncertainty quantification, offering practical guidance for safe deployment in safety-critical applications.

Data availability statement

The source code and experimental data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools (Claude, ChatGPT) to improve language and readability of selected sections. All AI-generated suggestions were carefully reviewed, verified, and edited by the authors. The authors take full responsibility for the content of the manuscript.

Acknowledgments

This work was supported by the Ningbo Municipal Science and Technology Innovation 2025 Major Project (Grant No. 2025Z126).

Authors' contribution

Conceptualization, B.Z. and G.C.; methodology, B.Z.; software, B.Z.; validation, B.Z. and G.C.; formal analysis, B.Z.; investigation, B.Z.; resources, B.Z. and W.F.; data curation, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, B.Z., G.C. and W.F.; visualization, B.Z.; supervision, B.Z.; project administration, B.Z.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, *et al*. Human-level control through deep reinforcement learning. *Nature* 2015, 518(7540):529–533.

[2] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016, 529(7587):484–489.

[3] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, PMLR*, Stockholm, Sweden, July 10–15, 2018, pp. 1861–1870.

[4] Levine S, Finn C, Darrell T, Abbeel P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* 2016, 17(39):1–40.

[5] Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, *et al*. Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans. Intell. Transp. Syst.* 2021, 23(6):4909–4926.

[6] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, *et al.* Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, New York, USA, June 19–24, 2016, pp. 1928–1937.

[7] Ghavamzadeh M, Mannor S, Pineau J, Tamar A. Bayesian reinforcement learning: a survey. *Found. Trends Mach. Learn.* 2015, 8(5–6):359–483.

[8] Moerland TM, Broekens J, Plaat A, Jonker CM. Model-based reinforcement learning: a survey. *Found. Trends Mach. Learn.* 2023, 16(1):1–118.

[9] Shalev-Shwartz S, Shammah S, Shashua A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv* 2016, arXiv:1610.03295.

[10] Gottesman O, Johansson F, Komorowski J, Faisal A, Sontag D, *et al.* Guidelines for reinforcement learning in healthcare. *Nat. Med.* 2019, 25(1):16–18.

[11] Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. *ACM Comput. Surv.* 2021, 55(1):1–36.

[12] Xiao J, Pisutsin P, Feroskhan M. Toward collaborative multitarget search and navigation with attention-enhanced local observation. *Adv. Intell. Syst.* 2024, 6(6):2300761.

[13] Xiao J, Pisutsin P, Feroskhan M. Collaborative target search with a visual drone swarm: an adaptive curriculum embedded multistage reinforcement learning approach. *IEEE Trans. Neural Netw. Learn. Syst.* 2025, 36(1):313–327.

[14] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In *Proceedings of ICML 2015 (International Conference on Machine Learning 2015)*, Lille, France, June 6–11, 2015, pp. 1613–1622.

[15] Dearden R, Friedman N, Russell S. Bayesian Q-learning. In *Proceedings of AAAI 1998 (American Association for Artificial Intelligence Conference)*, Madison, USA, July 26–30, 1998, pp. 761–768.

[16] Buckman J, Hafner D, Tucker G, Brevdo E, Lee H. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Proceedings of Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, Canada, December 3–8, 2018, pp. 8234–8244.

[17] Chen R, Sidor S, Abbeel P, Schulman J. UCB exploration via Q-ensembles. *arXiv* 2017, arXiv:1706.01502.

[18] Osband I, Blundell C, Pritzel A, Van Roy B. Deep exploration via bootstrapped DQN. In *Proceedings of Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, December 5–10, 2016, pp. 4026–4034.

[19] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation. *arXiv* 2016, arXiv:1506.02157.

[20] Romano Y, Patterson E, Candès E. Conformalized quantile regression. In *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, December 8–14, 2019, pp. 3543–3553.

[21] Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World*, 1st ed. Berlin: Springer, 2005.

[22] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. *arXiv* 2017, arXiv:1707.06347.

[23] Strens MJA. A Bayesian framework for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, Stanford, USA, June 29–July 2, 2000, pp. 943–950.

[24] Zhang W, Zhou D, Li L, Gu Q. Neural Thompson sampling. *arXiv* 2021, arXiv:2010.00827.

[25] Osband I, Aslanides J, Cassirer A. Randomized prior functions for deep RL. In *Proceedings of Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, Canada, December 3–8, 2018.

[26] Bellemare MG, Dabney W, Munos R. A distributional perspective on reinforcement learning. In *Proceedings of ICML 2017 (International Conference on Machine Learning 2017)*, Sydney, Australia, August 6–11, 2017, pp. 449–458.

[27] Dabney W, Ostrovski G, Silver D, Munos R. Implicit quantile networks for distributional RL. In *Proceedings of ICML 2018 (International Conference on Machine Learning 2018)*, Stockholm, Sweden, July 10–15, 2018, pp. 1096–1105.

[28] Dabney W, Rowland M, Bellemare MG, Munos R. Distributional reinforcement learning with quantile regression. In *Proceedings of AAAI 2018 (Thirty-Second AAAI Conference on Artificial Intelligence)*, New Orleans, USA, February 2–7, 2018.

[29] Osband I, Wen Z, Asghari SM, Dwaracherla V, Ibrahimi M, *et al*. Epistemic neural networks. In *Proceedings of Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, USA, December 10–16, 2023, pp. 2795–2823.

[30] Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv* 2021, arXiv:2107.07511.

[31] Shafer G, Vovk V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* 2008, 9(3):371–421.

[32] Papadopoulos H, Vovk V, Gammerman A. Conformal prediction with neural networks. In *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, October 29–31, 2007, pp. 388–395.

[33] Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Predictive inference with the jackknife+. *Ann. Stat.* 2021, 49(1):486–507.

[34] Gibbs I, Candès E. Adaptive conformal inference under distribution shift. In *Proceedings of Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, Virtual, December 6–14, 2021, pp. 1660–1672.

[35] Zhang Y, Shi C, Luo S. Conformal off-policy prediction. *arXiv* 2023, arXiv:2206.06711.

[36] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* 1992, 8(3–4):229–256.

[37] Thrun SB. Efficient exploration in reinforcement learning. 1992. Available: https://dl.acm.org/doi/abs/10.5555/865072 (accessed on 15 April 2026).

[38] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 2002, 47(2–3):235–256.

[39] Lattimore T, Szepesvári C. *Bandit Algorithms*, 1st ed. Cambridge: Cambridge University Press, 2020.

[40] Lee k, Laskin M, Srinivas A, Abbeel P. SUNRISE: ensemble learning in deep RL. In *Proceedings of the 38th International Conference on Machine Learning, PMLR*, Virtual, July 18–21, 2021, pp. 6131–6141.

[41] Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning, PMLR* , Sydney, Australia, August 6–11, 2017, pp. 2778–2787.

[42] Burda Y, Edwards H, Storkey A, Klimov O. Exploration by random network distillation. *arXiv* 2018, arXiv:1810.12894.

[43] Stadie BC, Levine S, Abbeel P. Incentivizing exploration in RL with deep predictive models. *arXiv* 2015, arXiv:1507.00814.

[44] Hafez MB, Weber C, Kerzel M, Wermter S. Curious meta-controller: adaptive alternation between model-based and model-free control. *arXiv* 2019, arXiv:1905.01718.

[45] Sekar R, Rybkin O, Daniilidis K, Abbeel P, Hafner D, *et al*. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning, PMLR*, Virtual, July 13–18, 2020, pp. 8583–8592.

[46] Dean V, Tulsiani S, Gupta A. See, hear, explore: curiosity via audio-visual association. In *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Virtual, December 6–12, 2020, pp. 14961–14972.

[47] Islam R, Seraj R, Bacon PL, Precup D. Marginalized state distribution entropy regularization in policy optimization. *arXiv* 2019, arXiv:1906.05064.

[48] Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge: MIT Press, 2018.

[49] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, Lille, France, July 6–11, 2015, pp. 1889–1897.

[50] Schulman J, Moritz P, Levine S, Jordan M, Abbeel P. High-dimensional continuous control using generalized advantage estimation. *arXiv* 2016, arXiv:1506.02438.

[51] Kakade S, Langford J. Approximately optimal approximate reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, Australia, July 8–12, 2002, pp. 267–274.

[52] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, *et al*. OpenAI gym. *arXiv* 2016, arXiv:1606.01540.

[53] Henderson P, Islam R, Bachman P, Pineau J, Precup D, *et al*. Deep reinforcement learning that matters. In *Proceedings of AAAI Conference on Artificial Intelligence*, New Orleans, USA, February 2–7, 2018.