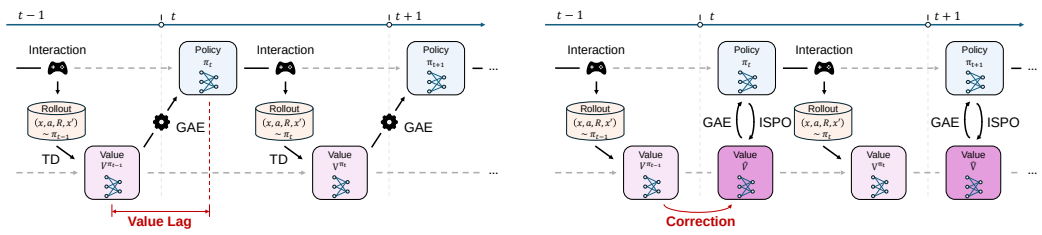
ISPO: CORRECTING POLICY-INDUCED VALUE LAG IN PPO VIA IMPORTANCE-SAMPLED POLICY SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy actor–critic RL, represented by PPO, trains the critic on rollouts from the policy it just replaced. After each actor update, the true value function moves while the critic still fits the previous one, leaving a **policy-induced value lag**. Our analysis shows that this lag biases every advantage estimate the actor consumes and, when large, violates the small-update assumption that PPO relies on. The lag is bounded by the product of effective horizon, action-wise residual dispersion, and policy-update size, identifying long-horizon, action-sensitive control as its worst case. Prior work addresses this through trajectory-level reweighting (Retrace, V-trace), which compounds importance-ratio variance, or policy-conditioned values (PeVFA), which add parameters and training costs. We argue neither is necessary: PPO constrains policy updates, making the value shift nearly first-order and estimable directly from the existing rollout. Motivated by this, we propose **I**mportance-**S**ampled **P**olicy-Shift **O**ptimization (**ISPO**). We construct a **PINN-style evaluation residual** per transition based on the policy evaluation PDE, and reweight it with a **single-step importance ratio**, avoiding cumulative off-policy weights on the whole trajectory. The resulting shift corrects the critic target after each actor update toward the updated policy's value function. We evaluate ISPO on an analytic LQR system, standard MuJoCo benchmarks, and an embodied manipulation task. Experiments show that ISPO achieves higher final return while adding no environment interaction, no learnable parameters, and no cumulative importance weights, improving return by 10% to 40% over vanilla PPO and matching or exceeding other baselines. Here is Project Page

1 INTRODUCTION



(a) **Standard on-policy RL** (b) **ISPO (Ours)**

Figure 1: **Standard on-policy RL vs. our ISPO**. *(a)* On-policy RL uses critic V_ϕ to optimize π_t, while V_ϕ itself fits rollout from π_{t-1}. *(b)* ISPO directly utilizes the policy evaluation PDE to perform advance estimation of V^π, without re-interacting with the environment.

Deep Reinforcement Learning (RL) has been widely applied across various fields, driving remarkable advances. Among these methods, on-policy RL, particularly Proximal Policy Optimization (PPO) (Schulman et al., 2017), has shown significant advantages in robustness and convergence, notably reducing the difficulty of parameter tuning and reward function design in high-dimensional state spaces. As RL is applied to increasingly challenging tasks, such as LLM post-training (Ziegler et al., 2019; Ouyang et al., 2022; Shao et al., 2024) and embodied AI fine-tuning (Zhang et al., 2025; Li et al., 2026), these methods are becoming increasingly important.

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

However, exploring the state space and finding optimal policies remains non-trivial, especially in large-scale, complex tasks. For on-policy RL, a primary challenge is value-function lag, which has been discussed (Borkar & Konda, 1997; Tang et al., 2022). Specifically, continuous updates of the actor π_t cause the true value function V^{π_t} to shift, while the critic still fits rollouts from π_{t-1}, as shown in Figure 1 *(a)*, which introduces bias into policy updates and slows down actor convergence. PPO employs importance sampling for Generalized Advantage Estimation (GAE), which partially alleviates this issue. However, this estimation assumes that the new and old policies do not diverge significantly in their value. When the value is sensitive to policy changes, this correction can introduce considerable instability. PPO consequently clips updates, although this slows convergence in complex scenarios such as long-horizon embodied RL with precise operations.

A more straightforward idea is to directly adjust the critic toward V^{π_t} rather than $V^{\pi_{t-1}}$. In recent years, several studies have explored this direction, such as applying the off-policy trajectory correction methods (Munos et al., 2016; Espeholt et al., 2018; Luo et al., 2020), conditioning value functions on policy representations (Faccio et al., 2021; Tang et al., 2022; Bohlinger & Peters, 2025), modeling value-improvement paths across historical policies (Dabney et al., 2021), or modifying actor-critic optimization to improve value tracking (Wang et al., 2025). However, these methods either require learning additional information or are restricted to specific application domains. This motivates the problem studied in this work:

> *How can one directly use trajectories rolled out by π_{t-1} to obtain a correction of V_ϕ toward V^{π_t} rather than $V^{\pi_{t-1}}$, without additional rollout or learnable parameters?*

We formulate the value function as a policy-evaluation equation and interpret policy-induced value lag through the lens of Partial Differential Equation (PDE) solving. We analyze the potential effects of the intrinsic update mechanism in on-policy RL. Based on this perspective, we propose **I**mportance-**S**ampled **P**olicy-Shift **O**ptimization (**ISPO**), a plug-and-play enhancement to PPO. Specifically, we add a first-order local correction term to the critic loss using a Physics-Informed Neural Network (PINN)-style residual (Raissi et al., 2019) derived from the policy evaluation PDE. With importance sampling, the critic directly approximates V^{π_t} of the current policy π_t rather than $V^{\pi_{t-1}}$ of the previous policy, as shown in Figure 1 *(b)*. Unlike other off-policy correction methods that directly apply importance sampling to global trajectories, ISPO mitigates the accumulated importance weights in multi-step returns and reduces estimation variance from trajectory sampling. Our contributions are summarized as:

1. We rethink policy update-induced value lag through the policy evaluation PDE, showing how horizon, residual dispersion, and policy shift govern its magnitude, how policy sensitivity affects its gradient, and how it affects long-horizon complex tasks.
2. We apply a first-order value shift estimator using local importance sampling, avoiding full trajectory reweighting while estimating the correction directly from existing rollouts.
3. We integrate this correction into PPO to steer the critic toward V^{π_t} rather than $V^{\pi_{t-1}}$, without additional environment interaction or learnable parameters.

Experiments across standard MuJoCo benchmarks and complex embodied manipulation tasks demonstrate the proposed method's competitive gains in complex scenarios.

2 Related Work

2.1 Critic Bias Analysis for On-Policy RL

In the actor–critic paradigm, the bias between an evolving policy and its value estimate is widely discussed. Traditionally, this issue is mitigated by updating the critic faster than the actor (Borkar & Konda, 1997). Recent studies have analyzed policy optimization via Taylor expansion (Tang et al., 2020) and critic tracking under Markovian sampling and policy drift across various settings (Chen & Zhao, 2023; Chen et al., 2024; Chen & Zhao, 2025; Kumar et al., 2025; Chen et al., 2026). Zorba et al. (2026) further relates critic error to value-function drift and timescale separation, while Gaur et al. (2024) bounds the last-iterate optimality gap of actor–critic by the critic estimation error. Collectively, these studies show that the critic must track a moving, policy-dependent value target,

and that insufficient tracking can impair actor updates and convergence. Nevertheless, they mainly focus on critic training errors, with limited analysis of mechanisms inherent to on-policy RL.

Another line of prior work mitigates critic lag by improving value estimation or learning values across policies, particularly for policy-gradient and PPO convergence. These include modeling value-improvement paths (Dabney et al., 2021), conditioning values on policy representations (Faccio et al., 2021; Tang et al., 2022), accelerating critic tracking with extra value updates (Cobbe et al., 2021; Wang et al., 2025), and learning relative value differences for PPO (Höftmann et al., 2026). These methods improve value tracking through additional optimization or alternative representations, whereas we directly correct the value estimate from the existing rollout without an explicit policy-conditioned model. Off-policy methods such as Retrace (Munos et al., 2016), V-trace (Espeholt et al., 2018), and IMPACT (Luo et al., 2020) avoid full-trajectory ratio variance by truncating per-step importance weights, but still reweight multi-step returns through products of truncated ratios to estimate target-policy values from off-policy data.

2.2 Physics-Informed Regularization of Value Function

Some studies have framed RL as optimal control and characterized it through Hamilton—Jacobi—Bellman (HJB) equations. Early work established the continuous-time HJB and viscosity-solution foundations of value learning (Doya, 2000; Munos, 2000), while subsequent studies derived HJB-based formulations for policy evaluation, actor–critic learning, exploration, action-value functions, and return distributions (Wang et al., 2020; Kim & Yang, 2020; Jia & Zhou, 2022a;b; Wiltzer et al., 2022). A complementary line connects HJ reachability with RL by constructing specialized Bellman objectives for safety, reach-avoid, and multi-objective control tasks (Fisac et al., 2019; Hsu et al., 2021; So et al., 2024; Sharpless et al., 2026). These works demonstrate that optimal control theory provides a general framework for interpreting RL objectives and algorithms.

There are other works that directly incorporate the HJB equation or its variants into RL objectives as critic residual constraints (Mukherjee & Liu, 2023), extending this idea to offline RL with local first-order dynamics (Lien et al., 2024) and goal-conditioned learning with Eikonal constraints (Giammarino et al., 2025; Giammarino & Qureshi, 2026). Combined with PINNs, these methods inject optimal-control priors into value learning to improve critic Out-Of-Distribution (OOD) generalization. In contrast, we use the PDE residual to construct an importance-sampled value correction for mitigating value-function lag using only single-step importance ratios.

3 Preliminaries

We introduce the PPO formulation and the continuous-time policy-evaluation PDE used in our analysis. Consider a Markov Decision Process (MDP) with state $x \in \mathcal{X}$, action $a \in \mathcal{A}$, and discount factor γ. Let $\pi_\theta(a \mid x)$ denote the stochastic policy, and $V_\phi(x)$ denote the expected discounted return from state x under policy π_θ. PPO updates the policy using trajectories collected by the old policy $\pi_{\theta_{\text{old}}}$. At each MDP step s, define the probability ratio

$$\rho_s(\theta) = \frac{\pi_\theta(a_s \mid x_s)}{\pi_{\theta_{\text{old}}}(a_s \mid x_s)}, \tag{1}$$

where (x_s, a_s) denotes the state-action pair. PPO maximizes the clipped surrogate objective

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_s \left[\min \left(\rho_s(\theta)\hat{A}_s, \text{clip}(\rho_s(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_s \right) \right], \tag{2}$$

where $\hat{A}_s$ is the estimated advantage and $\epsilon > 0$ controls the policy update size. Meanwhile, a critic V_ϕ parameterized by ϕ is trained to approximate the value function of the policy, which is trained by

$$\mathcal{L}_{\text{TD}}(\phi) = \frac{1}{2}\mathbb{E}_{x_s \sim \mathcal{D}_{t-1}} \left[\left(V_\phi(x_s) - \text{sg}\left[\hat{A}_s + V_{\phi_{\text{old}}}(x_s) \right] \right)^2 \right], \tag{3}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

Prior work has shown that Temporal Difference (TD) admits an HJB equation in the continuous-time limit (Doya, 2000; Munos, 2000). Assume that the underlying physical system is modeled as a continuous-time deterministic flow with a differentiable value function. Consider a discount rate

$\beta > 0$ related to the discrete-time discount factor γ through $\gamma = e^{-\beta \bar{s}}$, where $\bar{s}$ denotes the time interval between two consecutive MDP steps. The optimal value $V^\star$ satisfies the HJB equation

$$\beta V^\star(x) = \max_{a \in \mathcal{A}} \left[r(x,a) + \nabla V^\star(x)^\top f(x,a) \right], \tag{4}$$

where $r(x,a)$ is the instantaneous reward rate, and $f(x,a)$ represents the system dynamics. For on-policy methods, the value function corresponds to a specific policy rather than the globally optimal action, and the HJB equation reduces to the corresponding policy evaluation PDE

$$\beta V^\pi(x) = \mathbb{E}_{a \sim \pi(\cdot|x)} \left[r(x,a) + \nabla V^\pi(x)^\top f(x,a) \right], \tag{5}$$

where V^π denotes the value function under a given policy π. Therefore, this equation provides a continuous-time characterization of the target value function approximated by the PPO critic.

4 Theoretical Analysis

In this section, we characterize the value-function lag induced by policy updates. Prior works (Gaur et al., 2024; Zorba et al., 2026) study value lag caused by approximation error when the critic is underconverged, which differs from our setting. Here, we assume that the critic V_ϕ accurately represents the ideal value function V^π of policy π, thereby isolating the policy-update-induced effect. All mathematical proofs of the lemmas and propositions in this section are provided in Appendix B.

4.1 Understanding Value Lag by Policy Evaluation PDE

We first quantify the policy lag from the perspective of the policy evaluation PDE. At each training step t (distinct from the previous MDP step s), consider the current learned policy π_t and the behavior policy π_{t-1} used to collect the rollout. We define the policy-induced value shift as

$$\Delta_t(x) := V^{\pi_t}(x) - V^{\pi_{t-1}}(x). \tag{6}$$

This quantity measures how far the target value function moves after updating the policy from π_{t-1} to π_t. Substitute this equation into Equation 5 in Section 3:

$$\begin{aligned} \beta \left(V^{\pi_{t-1}}(x) + \Delta_t(x) \right) &= \mathbb{E}_{a \sim \pi_t(\cdot|x)} \left[r(x,a) + \nabla \left(V^{\pi_{t-1}}(x) + \Delta_t(x) \right)^\top f(x,a) \right] \\ &= \mathbb{E}_{a \sim \pi_t(\cdot|x)} \left[r(x,a) + \nabla V^{\pi_{t-1}}(x)^\top f(x,a) + \nabla \Delta_t(x)^\top f(x,a) \right], \end{aligned} \tag{7}$$

and rearranging the terms gives

$$\beta \Delta_t(x) = \mathbb{E}_{a \sim \pi_t(\cdot|x)} \left[\nabla \Delta_t(x)^\top f(x,a) \right] + g_t(x), \tag{8}$$

where we write the source term as

$$g_t(x) := \mathbb{E}_{a \sim \pi_t(\cdot|x)} \left[r(x,a) + \nabla V^{\pi_{t-1}}(x)^\top f(x,a) - \beta V^{\pi_{t-1}}(x) \right]. \tag{9}$$

This indicates that the policy-induced value shift itself also satisfies a PDE similar to Equation 5. Notably, its source term g_t is determined by evaluating old value function $V^{\pi_{t-1}}$ under new policy π_t. Therefore, it can measure the value mismatch caused by policy change, because if $\pi_t = \pi_{t-1}$, then $g_t(x) = 0$ by Equation 5. For further analysis, we write the policy evaluation residual as

$$\delta_{t-1}(x,a) := r(x,a) + \nabla V^{\pi_{t-1}}(x)^\top f(x,a) - \beta V^{\pi_{t-1}}(x), \tag{10}$$

and from Equation 1, let

$$\rho_t(x,a) := \frac{\pi_t(a \mid x)}{\pi_{t-1}(a \mid x)}. \tag{11}$$

Then we establish the following lemma:

Lemma 4.1 (Policy Update Induces Local Evaluation Mismatch)**.** *Assume that $\pi_t(\cdot \mid x)$ is absolutely continuous with respect to $\pi_{t-1}(\cdot \mid x)$, and that $\delta_{t-1}(x,a)$ and $\rho_t(x,a)$ have finite second moments under $\pi_{t-1}(\cdot \mid x)$. Then*

$$g_t(x) = \mathrm{Cov}_{a \sim \pi_{t-1}(\cdot|x)} \left(\rho_t(x,a), \delta_{t-1}(x,a) \right), \tag{12}$$

That is, the value of the source term of Equation 8 is determined by both the interaction between the action-wise residual and the policy update.

Equation 12 provides an interpretation of how the actor update changes the target value function. If the actor increases the probability of actions with positive δ_{t-1} and decreases the probability of actions with negative δ_{t-1}, the covariance becomes positive, and hence $g_t > 0$. Therefore, even when $V^{\pi_{t-1}}$ exactly satisfies the evaluation equation of the old policy, changing the policy generally makes the same value function inconsistent with the evaluation equation of the new policy.

4.2 From Local Evaluation Mismatch to Value Lag

We further analyze how g_t specifically exerts its effect on Δ_t. Based on Equation 8, we have a continuous-time version of the classic performance difference lemma (Kakade & Langford, 2002):

Lemma 4.2 (Continuous-Time Performance Difference). *For $\beta > 0$, assume sufficient regularity, discounted source integrability, and $e^{-\beta T}\mathbb{E}_{\pi_t}[\Delta_t(x_T) \mid x_0 = x] \to 0$ as $T \to \infty$. Then*

$$\Delta_t(x) = \mathbb{E}_{\pi_t}\left[\int_0^\infty e^{-\beta\tau} g_t(x_\tau)\, d\tau \,\middle|\, x_0 = x\right], \tag{13}$$

where τ denotes continuous time, and x_τ denotes the state reached at time τ under policy π_t.

Lemma 4.2 provides two interpretations, as shown in Figure 2:

***(a)* Systematic Underestimation.** By Lemma 4.1, if $g_t(x) \geq 0$ over reachable states, then $\Delta_t(x) \geq 0$. Thus, when policy updates favor actions with positive residuals, the old value function systematically underestimates the current policy's value.

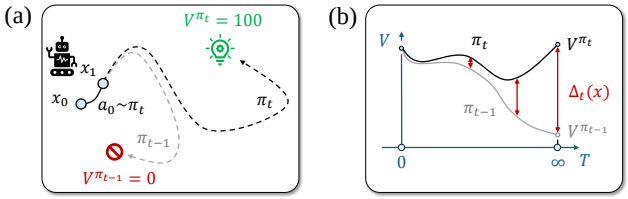


Figure 2: **Two interpretations** of Equation 13.

***(b)* Long-Horizon Accumulation.** The value shift at x depends on both local and future source terms along reachable trajectories. Hence, even small local mismatches can accumulate over a long horizon and produce substantial estimation error.

4.3 Magnitude and Gradient of Policy-Induced Value Lag

We now identify the factors governing value lag and distinguish their effects on its state gradient.

Proposition 4.3 (Magnitude Bound of Value Lag). *Under the conditions of Lemmas 4.1 and 4.2,*

$$\|\Delta_t\|_\infty \leq \underbrace{\frac{1}{\beta}}_{\text{horizon}} \sup_x \Big\{ \underbrace{\mathrm{Std}_{a\sim\pi_{t-1}(\cdot|x)}[\delta_{t-1}(x,a)]}_{\text{residual dispersion}} \times \underbrace{\sqrt{\chi^2\left(\pi_t(\cdot \mid x) \| \pi_{t-1}(\cdot \mid x)\right)}}_{\text{policy update size}} \Big\}. \tag{14}$$

That is, this magnitude is jointly controlled by the effective horizon, the dispersion of the action-wise residual, and the magnitude of the policy update.

Proposition 4.4 (Gradient Bound of Value Lag). *Under the conditions of Lemmas 4.1 and 4.2 and the assumptions in Appendix B.4, let $\lambda < \beta$ bound the expansion of the current-policy flow Φ_τ, then*

$$\|\nabla\Delta_t\|_\infty \leq \frac{1}{\beta - \lambda}\|\nabla g_t\|_\infty. \tag{15}$$

That is, the gradient bound depends on closed-loop expansion through $1/(\beta - \lambda)$.

Proposition 4.3 implies that value lag causes pronounced biases in **long-horizon tasks**, **large action-wise residual dispersion**, and when **policy shifts are rapid**, while Proposition 4.4 further shows that the bound on the lag gradient depends on how quickly trajectories starting from similar states move apart over time, as Figure 2 (a).

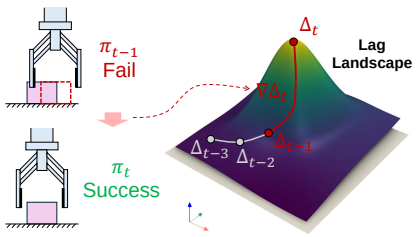


Figure 3: **Moderate policy change** can increase **substantial value deviation** in action-sensitive tasks.

These two bounds are particularly relevant in action-sensitive domains, such as embodied RL with long horizons and precise manipulation, where even nearby actions can lead to substantially different state transitions and long-term outcomes. In particular, $\mathrm{Std}_{a\sim\pi_{t-1}(\cdot|x)}[\delta_{t-1}(x,a)]$ in Equation 14 captures how actions at the same state differentially affect the local policy-evaluation residual. This means even a moderate policy update can induce a large evaluation mismatch when it systematically shifts

probability toward actions with different residuals, as shown in Figure 3. From Proposition 4.4, these mismatches produce a larger policy-induced value shift in value space. This value deviation may violate PPO's small-update assumption, introducing substantial variance into importance sampling and impairing policy convergence, as discussed in Appendix C.

5 METHODOLOGY

We introduce ISPO, which directly estimates the first-order policy-induced value shift using the existing rollout and constructs a corrected value target for the updated policy. Section 4 shows that the policy update first induces a local evaluation mismatch and that these local mismatches accumulate into the overall value shift. For PPO, however, the policy update is intentionally constrained to remain small. This leads to the following corollary:

Corollary 5.1 (First-Order Value Change under Policy Updates)**.** *Under the conditions of Lemmas 4.1 and 4.2, let $\bar{\theta}$ denote the policy-parameter increment, with $\pi_{t-1} = \pi_\theta$ and $\pi_t = \pi_{\theta+\bar{\theta}}$. Suppose that the policy and its value function are twice continuously differentiable in θ, and differentiation can be interchanged with the discounted expectation. Then*

$$\Delta_t(x) = (\bar{\theta})^\top \nabla_\theta V^{\pi_\theta}(x) + \mathcal{O}(\|\bar{\theta}\|^2). \tag{16}$$

The leading term is the directional derivative of the value function along the policy update.

This result suggests that, under the small policy updates of PPO, the local source term is already first-order in the policy change, while trajectory distribution corrections affect propagation only at second order. Therefore, the leading correction can be estimated from an existing π_{t-1} rollout without fully importance-reweighting trajectories toward π_t, such as Espeholt et al. (2018).

5.1 FIRST-ORDER VALUE-SHIFT APPROXIMATION

Consider a rollout $\mathcal{D}_{t-1} = \{(x_i, a_i, R_i, x_i')\}_{i=1}^N$ under π_{t-1}, where R_i is the accumulated reward over $\bar{s}$. We follow Mukherjee & Liu (2023) and approximate dynamics via forward differences, i.e., $f(x,a) \approx (x' - x)/\bar{s}$, avoiding explicit system solutions or world-model learning. Keep the critic $V_{\phi_{\text{old}}}$ from before the current PPO iteration fixed as the reference value function. From Equation 10, we construct a PINN-style policy-evaluation PDE residual from each observed transition as

$$\widehat{\delta}_{t-1,s} = \frac{R_s}{\bar{s}} + \nabla_x V_{\phi_{\text{old}}}(x_s)^\top \frac{x_s' - x_s}{\bar{s}} - \beta V_{\phi_{\text{old}}}(x_s). \tag{17}$$

According to Lemma 4.1, we estimate its sample-wise contribution by

$$\widehat{g}_s = [\text{clip}\,(\rho_s, 1-\epsilon_v, 1+\epsilon_v) - 1]\,\widehat{\delta}_{t-1,s}, \tag{18}$$

where $\epsilon_v > 0$ controls the truncation of the importance ratio, and subtracting 1 can remove dependence on any action-independent bias in $V_{\phi_{\text{old}}}$, since $\mathbb{E}_{a\sim\pi_{t-1}}[\rho_s - 1] = 0$.

From Lemma 4.2, we estimate the value shift through a backward recursion over the existing rollout:

$$\widehat{\Delta}_s = \bar{s}\,\widehat{g}_s + \gamma \widehat{\Delta}_{s+1}, \qquad \widehat{\Delta}_M = 0, \tag{19}$$

where M denotes the terminal step index of a trajectory. For small policy updates, $\widehat{\Delta}_s$ approximates the first-order value change in Corollary 5.1 by accumulating local residual corrections along the existing rollout. This recursion resembles vanilla GAE, but each local correction uses only a single-step importance ratio rather than a cumulative trajectory ratio. Since the local correction is already first order in the policy change, replacing the behavior-state distribution with that of π_t affects the estimate only at second order. Thus, Equation 19 can propagate these corrections along the π_{t-1} trajectory while retaining the leading term and avoiding variance amplification from cumulative importance weights. A detailed analysis is provided in Appendix D.

We then directly add the estimated value shift to the standard TD loss. Unlike prior PINN-style RL work (Mukherjee & Liu, 2023; Lien et al., 2024; Giammarino & Qureshi, 2026), we do not optimize the PDE residual directly, avoiding inherent PINN training challenges such as gradient pathology. Instead, we treat this residual merely as a correction term to the TD loss. Let $\hat{A}_s^{\text{old}}$ denote the GAE computed from $\mathcal{D}_{t-1}$ before the current policy update. The ISPO critic objective is

$$\mathcal{L}_{\text{ISPO}}(\phi) = \frac{1}{2}\mathbb{E}_{x_s\sim\mathcal{D}_{t-1}}\left[\left(V_\phi(x_s) - \text{sg}\left[\hat{A}_s^{\text{old}} + V_{\phi_{\text{old}}}(x_s) + \alpha\widehat{\Delta}_s\right]\right)^2\right], \tag{20}$$

where the correction scale α controls the magnitude of the correction. In this loss, $\hat{A}_s^{\text{old}} + V_{\phi_{\text{old}}}(x_s)$ retains the empirical value target of standard PPO, while $\widehat{\Delta}_s$ corrects this target toward the updated policy. In particular, the smaller the policy update, the smaller the correction. When the policy remains unchanged, $\rho_s = 1$ and $\widehat{\Delta}_s = 0$, so the correction vanishes.

Algorithm 1: Importance-Sampled Policy-Shift Optimization

Require: Behavior policy π_{t-1}, critic V_ϕ, PPO epochs K, actor updates m, critic updates n, critic learning rate η_ϕ

Collect rollout $\mathcal{D}_{t-1} = \{(x_s, a_s, R_s, x'_s)\}$ using π_{t-1}
Compute returns and GAE targets $\hat{A}_s^{\text{old}}$ from $\mathcal{D}_{t-1}$
Compute the local residuals $\widehat{\delta}_{t-1,s}$ using Equation 17
for $\nu = 1, \ldots, K$ **do**
  Update actor by standard PPO loss $\mathcal{L}_{\text{PPO}}(\theta)$ for m times
  Recompute the current policy ratio ρ_s
  Compute $\widehat{g}_s$ using Equation 18
  Compute $\widehat{\Delta}_s$ by the backward recursion in Equation 19
  for $j = 1, \ldots, n$ **do**
    Construct the corrected critic objective $\mathcal{L}_{\text{ISPO}}(\phi)$
    Update the critic by $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}_{\text{ISPO}}(\phi)$
  end for
  Recompute GAE using the corrected critic V_ϕ
end for

5.2 Integration into PPO

At each iteration, we collect a rollout under π_{t-1} and compute the standard return and GAE targets, together with the local residuals in Equation 17 using the fixed critic $V_{\phi_{\text{old}}}$. After each PPO actor update epoch, we recompute the policy ratios ρ_s, estimate the local source terms via Equation 18, and obtain the value shifts through Equation 19. We then update the critic using Equation 20, followed by GAE recomputation for the next actor update. Algorithm 1 summarizes the procedure, with orange steps for ISPO additions and blue steps for modifications to standard PPO.

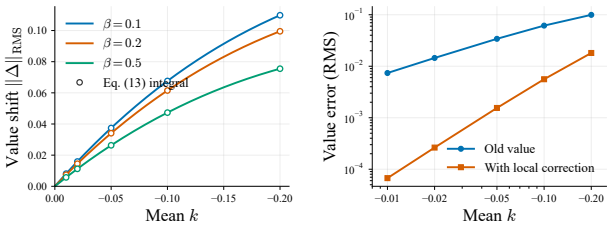
6 Experiments and Results

Our experiments address the following research questions: *RQ1.* Does value lag really exist? *RQ2.* Does our ISPO truly mitigate it, and does ISPO consistently yield performance improvements? *RQ3.* Do the performance gains align with theoretically predicted scenarios? *RQ4.* Under which scenarios is our method not expected to yield performance improvements? *RQ5.* Do the performance gains originate from importance sampling, or from the GAE refresh itself?

6.1 Mechanism Verification

We first conduct a controlled toy Linear–Quadratic Regulator (LQR) experiment to examine how policy updates change the value function. We consider the scalar continuous-time system $f(x, a) = -x + a$ with $r(x, a) = -(x^2 + a^2)$, and Gaussian policies $\pi_{k,v}(a \mid x) = \mathcal{N}(kx, v)$ with mean kx and variance v. The analytical solution of the value is

$$V^{k,v}(x) = -\frac{(1+k^2)x^2}{\beta + 2 - 2k} - \frac{v}{\beta}. \quad (21)$$

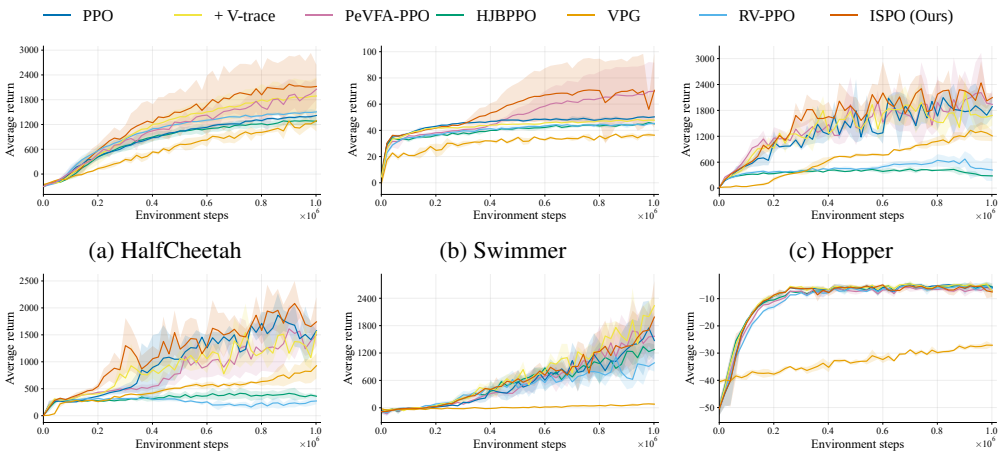


(a) **Policy-induced value lag** (b) **First-order correction**

Figure 4: **Mechanism verification**. *(a)* Analytical value lag and its reconstruction via Equation 13. *(b)* Value estimation error before and after local correction.

We fix the variance at $v = 0.25$ and vary the mean $k \in \{0, -0.01, -0.02, -0.05, -0.1, -0.2\}$ and discount rate $\beta \in \{0.1, 0.2, 0.5\}$, with the analytical value fixed at $k_{\text{old}} = 0$. This illustrates the discussion in Section 4: the policy $\pi_{k,v}$ improves while the value function $V^{k,v}$ fails to track the change, causing value lag. Appendix E.1 provides more details and the derivation.

RQ1: Mechanism Verification. Figure 4 *(a)* shows that policy updates induce a nonzero value lag even when the frozen critic represents the old policy, separating this effect from the critic approximation error. The lag increases with policy-update magnitude and is larger for smaller discount rates β. Numerically accumulating the local evaluation mismatch along new-policy trajectories closely matches the value shift, supporting Equation 13. Figure 4 *(b)* shows that accumulating the local correction along old-policy trajectories substantially reduces value estimation error, with greater improvement for smaller updates, consistent with the first-order approximation underlying ISPO.



(a) HalfCheetah (b) Swimmer (c) Hopper

(d) Walker2D (e) Ant (f) Reacher

Figure 5: **Average return curve during the training process with 1M environment steps**. Solid curves show mean return across 5 random seeds, and shaded regions indicate ± 1 standard deviation.

Table 1: **Performance on 1M environment steps** (mean ± 1 std.). **Bold** indicates the highest mean return in each task, while the improvement is calculated based on the comparison with vanilla PPO.

	PPO (2017)	+ V-trace (2018)	PeVFA-PPO (2022)	HJBPPO (2023)	VPG (2025)	RV-PPO (2026)	ISPO (Ours)	Improvement
HalfCheetah	1516 ± 53	1889 ± 448	2054 ± 259	1278 ± 106	1291 ± 259	1505 ± 46	$\mathbf{2121 \pm 523}$	+39.91%
Swimmer	50.3 ± 0.6	47.7 ± 1.0	69.8 ± 21.9	45.0 ± 0.1	36.3 ± 0.6	45.3 ± 0.7	$\mathbf{70.5 \pm 28.5}$	+40.16%
Hopper	1889 ± 241	1685 ± 238	1940 ± 317	283 ± 13	1209 ± 123	419 ± 254	$\mathbf{2139 \pm 94}$	+13.23%
Walker2D	1580 ± 15	1559 ± 67	1500 ± 214	362 ± 26	926 ± 236	273 ± 61	$\mathbf{1741 \pm 458}$	+10.21%
Ant	1473 ± 421	$\mathbf{2236 \pm 199}$	1567 ± 462	1278 ± 110	79 ± 27	984 ± 188	1987 ± 792	+34.89%
Reacher	$\mathbf{-5.7 \pm 1.4}$	-7.2 ± 2.6	-7.0 ± 0.8	-5.8 ± 0.7	-27.1 ± 0.5	-6.2 ± 1.1	-7.5 ± 1.9	-

6.2 Main Experiment

We consider MuJoCo environments (Todorov et al., 2012) and train using the Gymnasium framework (Towers et al., 2025). We select six benchmark tasks: `HalfCheetah`, `Swimmer`, `Hopper`, `Walker2D`, `Ant`, and `Reacher`. We compare our method against six baselines: standard PPO (Schulman et al., 2017), PPO + V-trace (Espeholt et al., 2018) (PPO with off-policy correction), PeVFA-PPO (Tang et al., 2022) (policy-conditioned value generalization), HJBPPO (Mukherjee & Liu, 2023) (HJB/PINN-based critic regularization), VPG-repeat-50 (Wang et al., 2025) (enhanced value estimation through additional critic updates), and RV-PPO (Höftmann et al., 2026) (relative value-difference learning). We run experiments on a workstation with an Intel Core Ultra-9 285K processor and an NVIDIA GeForce RTX-5090 GPU. For each environment, we average evaluations over 5 random seeds. More details are in Appendix E.2.

RQ2: Performance Comparison. Figure 5 and Table 1 show that ISPO improves PPO across several MuJoCo tasks, with particularly clear gains on `HalfCheetah` and `Swimmer`. On `HalfCheetah`, its advantage persists throughout the latter half of training; on `Swimmer`, ISPO continues to improve after PPO has largely plateaued. ISPO also achieves competitive returns on `Hopper`, `Walker2D` and `Ant` without introducing new learnable parameters. This advantage suggests that it improves synchronization between policy updates and value feedback, allowing gains in sustained locomotion and sequential coordination to be reflected more promptly in advantage estimates. For example, in the long-horizon `Swimmer` task, the agent must continuously coordinate body movements to generate propulsion through fluid interaction, while small policy changes can substantially affect propulsive forces over the horizon. Similarly, in `Hopper`, small changes in joint torques can cause agent falls and sharply reduce cumulative returns. These examples illustrate the sensitivity of the value to policy changes discussed in Figure 3.

Similar to Tang et al. (2022), we present the critic value calibration on `HalfCheetah` in Figure 6. Each point represents the relationship between the critic's predicted value and the ground-truth Mo-

nte Carlo rollout for a given initial state. Consequently, points closer to the diagonal indicate that the critic estimates the true value more accurately. This result directly confirms the systematic underestimation in vanilla PPO and validates the effectiveness of our correction method.

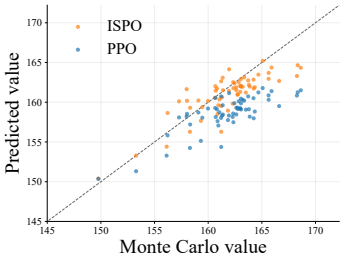


Figure 6: **Critic calibration**.

To evaluate the stability of our method under complex policies, we conduct our experiments in an embodied manipulation scenario using RLinf (Yu et al., 2026), a scalable and flexible RL infrastructure for robotics. We evaluate GR00T N1.5 (Bjorck et al., 2025) on the LIBERO benchmark (Liu et al., 2023) (including the `Spatial`, `Object`, `Goal`, and `Long` task suites). As training is confined strictly to the flow matching action expert, we adopt the noise injection approach from Flow-SDE (Liu et al., 2025). More details are in Appendix E.3.

RQ3: Complex Embodied Tasks. Table 2 presents the final success rate across the LIBERO suites, demonstrates that our ISPO outperforms standard PPO across the majority of configurations. As RL methods already reach performance saturation on LIBERO, we use this benchmark mainly for in-distribution validation, confirming that our method achieves further improvements on long-horizon tasks relative to standard baselines without significant trade-offs.

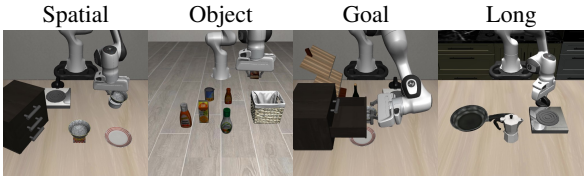


Figure 7: **Illustration of embodied tasks.**

Table 2: **Success rate (%) on embodied tasks**.

	Few-Shot SFT	Flow-SDE-PPO	Flow-SDE-GRPO	ISPO (Ours)
Spatial	41.4%	96.6%	93.0%	**98.8%**
Object	58.6%	**100.0%**	98.6%	99.4%
Goal	48.2%	93.8%	86.4%	**95.5%**
Long	61.9%	95.6%	91.7%	**97.2%**

RQ4: Limitation Analysis. On the `Reacher` in Figure 5, ISPO and most baselines show similar learning curves and final returns. Since its reward is directly determined by the Euclidean distance between the fingertip and the target, the value function is less policy-sensitive, leaving limited room for correction when policy-induced value lag is small (Proposition 4.3). In this regime, PPO's own importance sampling may already sufficiently correct the policy mismatch, while ISPO does not anticipate further improvement, which is consistent with our discussion.

6.3 Ablation Study

We conduct a parameter sweep over $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 2.5, 5, 10\}$ on the `Ant` benchmark, where $\alpha = 0$ corresponds to completely disabling the shift in Equation 20 to isolate the contribution of the GAE refresh. All other hyperparameters remain consistent with the main experiment.

RQ5: Ablation Study. Table 3 examines the correction scale α on `Ant`, where introducing a modest correction with $\alpha = 0.25$ raises the mean return to 1987, the highest among the tested configurations. This result suggests that relying solely on GAE refreshes is insufficient to boost performance, whereas appropriate correction can better approximate V^{π_t}. However, increasing α too much generally degrades performance because it may introduce excessive estimated noise for the value loss, consistent with the local nature of our correction.

Table 3: **Ablation study on `Ant`**.

Config	**Return** (Mean ±1 Std.)
PPO	1473 ± 421
ISPO with $\alpha = 0$	1694 ± 288
$\alpha = 0.25$	$\mathbf{1987 \pm 792}$
$\alpha = 0.5$	1576 ± 339
$\alpha = 0.75$	1770 ± 152
$\alpha = 1$	1334 ± 249
$\alpha = 2.5$	630 ± 29
$\alpha = 5$	835 ± 90
$\alpha = 10$	715 ± 107

7 Conclusion

We characterize policy-induced value lag in on-policy RL through a policy evaluation PDE and propose ISPO to correct the critic toward the updated policy using local importance sampling. Experiments show that ISPO improves performance without additional environment interaction or learnable parameters. Future work includes analyzing value lag in viscosity-solution settings such as contact-rich manipulation and investigating how to improve ISPO for such scenarios.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks with required disclosure. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design research methodology or experiments, implement methods, or interpret results, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing author-written text to improve grammar, clarity, and readability. We have reviewed all AI-assisted work. The authors manually checked all AI-assisted edits to ensure they did not alter the technical content or scientific meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We take several steps to facilitate reproducibility. All assumptions and complete proofs of the theoretical results are provided in the appendix. We also report detailed experimental settings in the appendix, including the analytical LQR setup, environment configurations, network architectures, training hyperparameters, baseline-specific settings, evaluation protocols, and random-seed specifications. For the embodied manipulation experiments, we report the benchmark, model, training configuration, and evaluation protocol used. We evaluate all reported results over multiple independent random seeds whenever applicable. In addition, the anonymous supplementary repository provides the ISPO implementation, baseline configurations, and scripts needed to reproduce the main experiments, ablations, tables, and figures in this paper.

REFERENCES

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Conference on Robot Learning (CoRL)*, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Robotics: Science and Systems XXI (RSS)*, 2025b.

Nico Bohlinger and Jan Peters. Massively scaling explicit policy-conditioned value functions. *arXiv preprint arXiv:2502.11949*, 2025.

Vivek S. Borkar and Vijaymohan R. Konda. The actor–critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543, 1997.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and

Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Robotics: Science and Systems XIX (RSS)*, 2023.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. In *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Xuyang Chen and Lin Zhao. On the convergence of continuous single-timescale actor-critic. In *42nd International Conference on Machine Learning (ICML)*, 2025.

Xuyang Chen, Jingliang Duan, and Lin Zhao. Global optimality of single-timescale actor-critic under continuous state-action space: A study on linear quadratic regulator. In *33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Xuyang Chen, Fengzhuo Zhang, Keyu Yan, and Lin Zhao. Finite-time analysis of actor-critic methods with deep neural network approximation. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Karl W Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *38th International Conference on Machine Learning (ICML)*, 2021.

Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. In *AAAI 35th Conference on Artificial Intelligence (AAAI)*, 2021.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1): 219–245, 2000.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *35th International Conference on Machine Learning (ICML)*, 2018.

Francesco Faccio, Louis Kirsch, and Jürgen Schmidhuber. Parameter-based value functions. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Jaime F. Fisac, Neil F. Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J. Tomlin. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (Last iterate) of actor-critic under Markovian sampling with neural network parametrization. In *41st International Conference on Machine Learning (ICML)*, 2024.

Vittorio Giammarino and Ahmed H. Qureshi. Goal reaching with eikonal-constrained hierarchical quasimetric reinforcement learning. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Vittorio Giammarino, Ruiqi Ni, and Ahmed Qureshi. Physics-informed value learner for offline goal-conditioned reinforcement learning. In *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Marc Höftmann, Jan Robine, and Stefan Harmeling. Relative value learning. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J. Tomlin, and Jaime F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *Robotics: Science and Systems XVII (RSS)*, 2021.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022a.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022b.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *19th International Conference on Machine Learning (ICML)*, 2002.

Jeongho Kim and Insoon Yang. Hamilton-jacobi-bellman equations for q-learning in continuous time. In *2nd Conference on Learning for Dynamics and Control (L4DC)*, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Conference on Robot Learning (CoRL)*, 2025.

Navdeep Kumar, Priyank Agrawal, Giorgia Ramponi, Kfir Y. Levy, and Shie Mannor. On the convergence of single-timescale actor-critic. In *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. SimpleVLA-RL: Scaling VLA training via reinforcement learning. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Yun-Hsuan Lien, Ping-Chun Hsieh, Tzu-Mao Li, and Yu-Shuen Wang. Enhancing value function estimation through first-order state-action dynamics in offline reinforcement learning. In *41st International Conference on Machine Learning (ICML)*, 2024.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di ZHANG, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online rl. In *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Michael Luo, Jiahao Yao, Richard Liaw, Eric Liang, and Ion Stoica. IMPACT: Importance weighted asynchronous architectures with clipped target networks. In *8th International Conference on Learning Representations (ICLR)*, 2020.

Amartya Mukherjee and Jun Liu. Bridging physics-informed neural networks with reinforcement learning: Hamilton-jacobi-bellman proximal policy optimization (HJBPPO). In *ICML 2023 Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299, 2000.

Remi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *30th Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

William Sharpless, Dylan Hirsch, Sander Tonkens, Nikhil Shinde, and Sylvia Herbert. Dual-objective reinforcement learning with novel hamilton-jacobi-bellman formulations. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Oswin So, Cheng Ge, and Chuchu Fan. Solving minimum-cost reach avoid using reinforcement learning. In *38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Hongyao Tang, Zhaopeng Meng, Jianye Hao, Chen Chen, Daniel Graves, Dong Li, Changmin Yu, Hangyu Mao, Wulong Liu, Yaodong Yang, Wenyuan Tao, and Li Wang. What about inputting policy in value function: Policy representation and policy-extended value function approximator. In *AAAI 36th Conference on Artificial Intelligence (AAAI)*, 2022.

Yunhao Tang, Michal Valko, and Remi Munos. Taylor expansion policy optimization. In *37th International Conference on Machine Learning (ICML)*, 2020.

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

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198): 1–34, 2020.

Tao Wang, Ruipeng Zhang, and Sicun Gao. Improving value estimation critically enhances vanilla policy gradient. In *42nd International Conference on Machine Learning (ICML)*, 2025.

Harley Wiltzer, David Meger, and Marc G. Bellemare. Distributional Hamilton–Jacobi–Bellman equations for continuous-time reinforcement learning. In *39th International Conference on Machine Learning (ICML)*, 2022.

Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, Quanlu Zhang, Yongji Wu, Chunyang Zhu, Junhao Hu, Zixiao Huang, Mingjie Wei, Yuqing Xie, Ke Yang, Bo Dai, Zhexuan Xu, Jiakun Du, Xiangyuan Wang, Xu Fu, Letong Shi, Zhihao Liu, Kang Chen, Weilin Liu, Gang Liu, Boxun Li, Jianlei Yang, Zhi Yang, Guohao Dai, and Yu Wang. RLinf: Flexible and efficient Large-Scale reinforcement learning via Macro-to-Micro flow transformation. In *20th USENIX Symposium on Operating Systems Design and Implementation (OSDI 26)*, 2026.

Hongyin Zhang, Zifeng Zhuang, Han Zhao, Pengxiang Ding, Hongchao Lu, and Donglin Wang. ReinboT: Amplifying robot visual-language manipulation with reinforcement learning. In *42nd International Conference on Machine Learning (ICML)*, 2025.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Conference on Robot Learning (CoRL)*, 2023.

Denis Zorba, David Šiška, and Łukasz Szpruch. Convergence of an actor-critic gradient flow for entropy-regularised MDPs in general spaces. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Contents

A Notation **15**

B Theoretical Proofs **16**

C Implications of Value Lag for Advantage Estimation **18**

D Approximation and Variance Analysis **19**

E Experimental Setup **20**

F Additional Experiments and Results **22**

A NOTATION

Table 4: Summary of the main notation used throughout this paper.

Symbol	Description	Symbol	Description
$x \in \mathcal{X}$	State and state space.	$a \in \mathcal{A}$	Action and action space.
s	Discrete MDP step index.	t	Policy-update iteration index.
τ	Continuous-time variable.	$\bar{s}$	Time interval between two consecutive MDP steps.
γ	Discrete-time discount factor.	β	Continuous-time discount rate, with $\gamma = e^{-\beta \bar{s}}$.
$r(x, a)$	Instantaneous reward rate.	$f(x, a)$	Continuous-time system dynamics.
$\Phi_\tau(x)$	State flow under the current policy π_t.	λ	Uniform growth bound of the current-policy flow.
R_s	Accumulated reward over one interval $\bar{s}$.	$\mathcal{D}_{t-1}$	Rollout collected under the behavior policy π_{t-1}.
π_θ	Stochastic policy parameterized by θ.	π_t, π_{t-1}	Current and behavior policies at iteration t.
$\bar{\theta}$	Policy-parameter increment from π_{t-1} to π_t.	$\chi^2(\pi_t \Vert \pi_{t-1})$	Chi-squared divergence measuring policy shift.
V^π	Exact value function under policy π.	$V^\star$	Optimal value function.
V_ϕ	Critic parameterized by ϕ.	$V_{\phi_{\text{old}}}$	Frozen reference critic before the current PPO iteration.
$\nabla_x V$	Gradient of the value function with respect to state.	$\Delta_t(x)$	Policy-induced value shift $V^{\pi_t}(x) - V^{\pi_{t-1}}(x)$.
$\delta_{t-1}(x, a)$	Action-wise policy-evaluation residual of $V^{\pi_{t-1}}$.	$g_t(x)$	Local source term induced by the policy update.
$\widehat{\delta}_{t-1,s}$	Sample-based policy-evaluation PDE residual.	$\widehat{g}_s$	Sample estimate of the local source contribution.
$\widehat{\Delta}_s$	Estimated first-order value shift at step s.	α	Scale of the ISPO value-shift correction.
ϵ_v	ISPO importance-ratio clipping threshold.	$\text{sg}[\cdot]$	Stop-gradient operator.
ρ_s	Single-step importance ratio used by PPO/ISPO.	ϵ	PPO policy-ratio clipping threshold.
$\hat{A}_s$	Estimated advantage at step s.	$\hat{A}_s^{\text{old}}$	GAE computed before the current policy update.
$\mathcal{L}_{\text{PPO}}$	PPO clipped actor objective.	$\mathcal{L}_{\text{ISPO}}$	ISPO-corrected critic objective.
K	Number of PPO epochs per rollout.	m, n	Number of actor and critic updates, respectively.
η_ϕ	Critic learning rate.	M	Terminal step index of a rollout.

B THEORETICAL PROOFS

B.1 PROOF OF LEMMA 4.1

Assumption B.1 (Absolute Continuity)**.** For the state x under consideration, the current policy is absolutely continuous with respect to the behavior policy:

$$\pi_t(\cdot \mid x) \ll \pi_{t-1}(\cdot \mid x). \tag{22}$$

Assumption B.2 (Finite Second Moments)**.** For the state x under consideration, the policy-evaluation residual δ and importance ratio ρ have finite second moments under the behavior policy:

$$\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}[\delta_{t-1}(x,a)^2] < \infty, \qquad \mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}[\rho_t(x,a)^2] < \infty. \tag{23}$$

Lemma B.3 (Policy Update Induces Local Evaluation Mismatch)**.** *See Lemma 4.1.*

Proof. Since the value $V^{\pi_{t-1}}$ satisfies Equation 5, the residual defined in Equation 10 has a zero expectation under π_{t-1}:

$$\begin{aligned}\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)\right] &= \mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[r(x,a) + \nabla V^{\pi_{t-1}}(x)^\top f(x,a)\right] - \beta V^{\pi_{t-1}}(x) \\ &= 0.\end{aligned} \tag{24}$$

Meanwhile, absolute continuity and the definition of ρ_t give

$$\begin{aligned}\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\rho_t(x,a)\right] &= \int_{\mathcal{A}} \rho_t(x,a)\,\pi_{t-1}(da \mid x) \\ &= 1.\end{aligned} \tag{25}$$

Using importance sampling, the source term in Equation 9 can therefore be rewritten as

$$\begin{aligned} g_t(x) &= \mathbb{E}_{a\sim\pi_t(\cdot|x)}\left[\delta_{t-1}(x,a)\right] \\ &= \mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\rho_t(x,a)\delta_{t-1}(x,a)\right] \\ &= \mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\left(\rho_t(x,a)-1\right)\delta_{t-1}(x,a)\right] \\ &= \mathrm{Cov}_{a\sim\pi_{t-1}(\cdot|x)}\left(\rho_t(x,a), \delta_{t-1}(x,a)\right), \end{aligned} \tag{26}$$

which proves Equation 12. □

B.2 PROOF OF LEMMA 4.2

Assumption B.4 (Dynamics and Regularity)**.** For the state x under consideration, the state process x_τ under π_t, with $x_0 = x$, is well defined for all $\tau \geq 0$ and has generator given by the policy evaluation operator

$$\mathcal{N}^{\pi_t}[h](x) := \mathbb{E}_{a\sim\pi_t(\cdot|x)}\left[\nabla h(x)^\top f(x,a)\right] \tag{27}$$

for sufficiently smooth functions h. The value shift Δ_t is sufficiently smooth for this operator to apply, and the regularity conditions allow differentiation under the expectation and interchange of expectation and time integration on every finite time interval.

Assumption B.5 (Discounting and Vanishing Terminal Term)**.** For the state x under consideration, the discount rate is positive, and the expected discounted terminal value shift vanishes:

$$\beta > 0, \qquad \lim_{T\to\infty} \mathbb{E}_{\pi_t}\left[e^{-\beta T}\Delta_t(x_T)\middle|\, x_0 = x\right] = 0. \tag{28}$$

Boundedness of Δ_t is sufficient for this condition, but is not required.

Assumption B.6 (Discounted Source Integrability)**.** For the state x under consideration, the source term g_t has a finite expected discounted absolute integral along the current-policy trajectory:

$$\mathbb{E}_{\pi_t}\left[\int_0^\infty e^{-\beta\tau}|g_t(x_\tau)|\,d\tau \middle| x_0 = x\right] < \infty. \tag{29}$$

Since $\beta > 0$, this condition holds when g_t is bounded.

Lemma B.7 (Continuous-Time Performance Difference)**.** *See Lemma 4.2.*

Proof. Using the policy evaluation operator in Assumption B.4, Equation 8 gives

$$\mathcal{N}^{\pi_t}[\Delta_t](x) - \beta\Delta_t(x) = -g_t(x). \tag{30}$$

Consider the expected discounted value shift

$$F(\tau) := \mathbb{E}_{\pi_t}\left[e^{-\beta\tau}\Delta_t(x_\tau)\middle|\, x_0 = x\right]. \tag{31}$$

Under Assumption B.4, differentiating with respect to τ and using Equation 30 gives

$$\begin{aligned}\frac{dF(\tau)}{d\tau} &= \mathbb{E}_{\pi_t}\left[e^{-\beta\tau}\left(\mathcal{N}^{\pi_t}[\Delta_t](x_\tau) - \beta\Delta_t(x_\tau)\right)\middle|\, x_0 = x\right] \\ &= -\mathbb{E}_{\pi_t}\left[e^{-\beta\tau}g_t(x_\tau)\middle|\, x_0 = x\right].\end{aligned} \tag{32}$$

Integrating both sides from 0 to T and rearranging the terms yields

$$\Delta_t(x) = \mathbb{E}_{\pi_t}\left[\int_0^T e^{-\beta\tau}g_t(x_\tau)\,d\tau + e^{-\beta T}\Delta_t(x_T)\middle|\, x_0 = x\right]. \tag{33}$$

By Assumption B.5, the terminal term satisfies

$$\mathbb{E}_{\pi_t}\left[e^{-\beta T}\Delta_t(x_T)\middle|\, x_0 = x\right] \longrightarrow 0 \qquad \text{as } T \to \infty. \tag{34}$$

Moreover, Assumption B.6 allows us to pass to the limit in the integral by dominated convergence. Taking $T \to \infty$ in Equation 33, we obtain

$$\Delta_t(x) = \mathbb{E}_{\pi_t}\left[\int_0^\infty e^{-\beta\tau}g_t(x_\tau)\,d\tau\middle|\, x_0 = x\right], \tag{35}$$

which proves Equation 13. □

B.3 Proof of Proposition 4.3

Proposition B.8 (Magnitude Bound of Value Lag)**.** *See Proposition 4.3.*

Proof. We combine the local source-term estimate from Lemma 4.1 with the trajectory representation in Lemma 4.2. Throughout this proof, we assume both lemmas hold at every state x. From Equation 26, the Cauchy–Schwarz inequality gives

$$\begin{aligned}|g_t(x)| &= \left|\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[(\rho_t(x,a) - 1)\,\delta_{t-1}(x,a)\right]\right| \\ &\le \sqrt{\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[(\rho_t(x,a) - 1)^2\right]}\sqrt{\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)^2\right]}.\end{aligned} \tag{36}$$

The first factor is determined by the χ^2 divergence,

$$\chi^2\left(\pi_t(\cdot\mid x)\|\pi_{t-1}(\cdot\mid x)\right) := \mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[(\rho_t(x,a) - 1)^2\right]. \tag{37}$$

Meanwhile, the zero-mean identity in Equation 24 implies

$$\sqrt{\mathbb{E}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)^2\right]} = \mathrm{Std}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)\right]. \tag{38}$$

Substituting these identities into Equation 36 gives

$$|g_t(x)| \le \mathrm{Std}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)\right]\sqrt{\chi^2\left(\pi_t(\cdot\mid x)\|\pi_{t-1}(\cdot\mid x)\right)}. \tag{39}$$

If the supremum of the right-hand side over x is infinite, the claimed bound is immediate. Otherwise, g_t is bounded. Using Lemma 4.2, we then obtain

$$\begin{aligned}|\Delta_t(x)| &\le \mathbb{E}_{\pi_t}\left[\int_0^\infty e^{-\beta\tau}|g_t(x_\tau)|\,d\tau\middle|\, x_0 = x\right] \\ &\le \|g_t\|_\infty\int_0^\infty e^{-\beta\tau}\,d\tau = \frac{1}{\beta}\|g_t\|_\infty.\end{aligned} \tag{40}$$

Taking the supremum over x and applying Equation 39 yields

$$\|\Delta_t\|_\infty \le \frac{1}{\beta}\sup_x\left\{\mathrm{Std}_{a\sim\pi_{t-1}(\cdot|x)}\left[\delta_{t-1}(x,a)\right]\sqrt{\chi^2\left(\pi_t(\cdot\mid x)\|\pi_{t-1}(\cdot\mid x)\right)}\right\}, \tag{41}$$

which proves Equation 14. □

B.4 Proof of Proposition 4.4

All vector norms are Euclidean, and matrix norms are the induced operator norms.

Assumption B.9 (Differentiable Flow and Uniform Growth Bound)**.** The current-policy dynamics $\dot{x} = f_{\pi_t}(x)$, where $f_{\pi_t}(x) = \mathbb{E}_{a\sim\pi_t(\cdot|x)}[f(x,a)]$, admit a global flow $\Phi_\tau(x)$ that is continuously differentiable with respect to the initial state x. There exists a constant $\lambda < \beta$ such that

$$\|D_x\Phi_\tau(x)\| \le e^{\lambda\tau}, \qquad \text{for all } x \in \mathcal{X} \text{ and } \tau \ge 0. \tag{42}$$

Assumption B.10 (Differentiable Source and Bounded Gradient)**.** The source term g_t is continuously differentiable with respect to the state and has a bounded gradient:

$$g_t \in C^1(\mathcal{X}), \qquad \|\nabla g_t\|_\infty < \infty. \tag{43}$$

Proposition B.11 (Gradient Bound of Value Lag)**.** *See Proposition 4.4.*

Proof. Under the first-order evaluation operator in Equation 5 and the flow in Assumption B.9, Lemma 4.2 gives the characteristic representation

$$\Delta_t(x) = \int_0^\infty e^{-\beta\tau} g_t(\Phi_\tau(x))\, d\tau, \qquad J_\tau(x) := D\Phi_\tau(x). \tag{44}$$

By Assumptions B.9 and B.10, the differentiated integrand is bounded uniformly by $e^{-(\beta-\lambda)\tau}\|\nabla g_t\|_\infty$. Since $\beta > \lambda$, this majorant is integrable, allowing differentiation under the integral:

$$\nabla\Delta_t(x) = \int_0^\infty e^{-\beta\tau} J_\tau(x)^\top \nabla g_t(\Phi_\tau(x))\, d\tau. \tag{45}$$

Consequently,

$$\|\nabla\Delta_t\|_\infty \le \|\nabla g_t\|_\infty \int_0^\infty e^{-(\beta-\lambda)\tau}\, d\tau = \frac{\|\nabla g_t\|_\infty}{\beta-\lambda}. \tag{46}$$

For sufficient regularity, write $r_\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)} r(x,a)$, $b_t = f_{\pi_t} - f_{\pi_{t-1}}$, and $q_t = r_{\pi_t} - r_{\pi_{t-1}}$. The old policy evaluation equation yields

$$g_t = q_t + (\nabla V^{\pi_{t-1}})^\top b_t, \tag{47}$$

$$\nabla g_t = \nabla q_t + (Db_t)^\top \nabla V^{\pi_{t-1}} + \nabla^2 V^{\pi_{t-1}} b_t. \tag{48}$$

Thus, $V^{\pi_{t-1}} \in C^2$ and $r_\pi, f_\pi \in C^1$ for both policies suffice for the continuous differentiability of g_t, the boundedness of its gradient is separately required by Assumption B.10.

Boundedness of g_t itself is not required here, since the discounted source integrability and terminal condition are supplied by the assumptions of Lemma 4.2. Here λ is a uniform flow-growth bound, which is stronger than an asymptotic Lyapunov exponent alone. When $\beta \le \lambda$, the value bound remains valid for bounded g_t, but this argument provides no finite gradient bound. □

C Implications of Value Lag for Advantage Estimation

We examine how policy-induced value lag enters the feedback used by PPO. Following Section 4, assume that the frozen critic exactly represents $V^{\pi_{t-1}}$, and compare the TD quantities obtained from the old and current value functions on the same rollout. For an observed transition (x_s, a_s, R_s, x'_s), define the one-step TD residual evaluated using a value function U as $d_s[U] := R_s + \gamma U(x'_s) - U(x_s)$. Then

$$d_s[V^{\pi_{t-1}}] - d_s[V^{\pi_t}] = \Delta_t(x_s) - \gamma\Delta_t(x'_s). \tag{49}$$

This identity holds transition by transition. On the same trajectory, the difference between the corresponding GAE estimates is the discounted, trace-weighted sum of these TD differences. Thus, policy-induced value lag changes the advantage feedback to the actor network even when the critic perfectly fits the old policy.

PPO importance ratios reweight the supplied advantage estimates without updating the values used for bootstrapping. Thus, even when policy ratios remain close to one, the critic may provide inaccurate value feedback for the updated policy. Proposition 4.3 shows how long horizons, large action-wise residual dispersion, and substantial policy changes amplify the bound on this discrepancy. This

motivates tracking value functions across successive policies. Gaur et al. (2024) and Wang et al. (2025) further show the importance of accurate value estimates for policy-gradient performance, providing sufficient motivation to address this lag.

For the PDE residual, gradient lag additionally changes relative residuals across actions through $-\nabla\Delta_t(x)^\top[f(x,a) - f(x,a')]$. Proposition 4.4 therefore connects closed-loop sensitivity to the action-dependent evaluation feedback used by ISPO. These observations motivate correcting the critic toward the updated policy and then refreshing the advantage estimates.

D Approximation and Variance Analysis

We quantify the error from propagating corrections along behavior trajectories and the sampling variance of the resulting estimator. Fix the policy pair (π_{t-1}, π_t) and the reference critic, assume absolute continuity as in Assumption B.1 at every visited state, and consider estimators without ratio clipping. Fix a starting step s and let $H = M - s \geq 1$ be the number of remaining transitions, where M is the terminal step index in Equation 19. The time step is $\bar{s} > 0$, with $\gamma = e^{-\beta\bar{s}} \in (0, 1)$.

Error from local trajectory correction. For $0 \leq l < H$, write $\rho_{s+l} := \rho_t(x_{s+l}, a_{s+l}) = \pi_t(a_{s+l} \mid x_{s+l})/\pi_{t-1}(a_{s+l} \mid x_{s+l})$, $Z_l = (\rho_{s+l} - 1)\widehat{\delta}_{t-1,s+l}$, and $W_{s:l} = \prod_{j=0}^{l} \rho_{s+j}$, with $W_{s:-1} = 1$. Here l is an offset from the fixed starting step s, and $\widehat{\delta}_{t-1,s+l}$ is the residual in Equation 17. Denote the full trajectory correction by $\widehat{\Delta}_s^{\mathrm{full}}$ and the local correction without clipping by $\widehat{\Delta}_s^{\mathrm{loc}}$:

$$\widehat{\Delta}_s^{\mathrm{full}} = \bar{s} \sum_{l=0}^{H-1} \gamma^l W_{s:l-1} Z_l, \tag{50}$$

$$\widehat{\Delta}_s^{\mathrm{loc}} = \bar{s} \sum_{l=0}^{H-1} \gamma^l Z_l. \tag{51}$$

The local estimator is the unrolled form of Equation 19 with ratio clipping removed, while the superscript loc distinguishes it from the clipped estimator $\widehat{\Delta}_s$ used in the method. With exact local residuals, $W_{s:l-1}$ corrects the distribution of the state at which each source contribution is evaluated.

Assume uniformly that $|\rho_t(x,a) - 1| \leq \eta_\rho$ with $0 \leq \eta_\rho < 1$, and $|\widehat{\delta}_{t-1,s+l}| \leq C_\delta$, with C_δ independent of η_ρ. Since $|W_{s:l-1} - 1| \leq l\eta_\rho(1+\eta_\rho)^{l-1}$ for $l \geq 1$, subtracting the two estimators gives

$$\left|\widehat{\Delta}_s^{\mathrm{full}} - \widehat{\Delta}_s^{\mathrm{loc}}\right| \leq \bar{s} C_\delta \eta_\rho^2 \sum_{l=1}^{H-1} l\gamma^l(1+\eta_\rho)^{l-1} = \mathcal{O}(\eta_\rho^2). \tag{52}$$

The order holds for fixed H, γ, and $\bar{s}$, and the same bound applies to the difference in expected corrections under π_{t-1} conditional on x_s. The explicit sum quantifies how the approximation error depends on the rollout horizon.

Effect of closed-loop expansion. A continuous-time counterpart further identifies the dynamical factor in this error. Under the conditions of Proposition 4.4, assume a global old-policy flow Ψ_τ satisfying $\dot{\Psi}_\tau(x) = f_{\pi_{t-1}}(\Psi_\tau(x))$ and $\Psi_0(x) = x$, where $f_\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)}[f(x,a)]$. Let $b_t = f_{\pi_t} - f_{\pi_{t-1}}$ be bounded. The time variable τ is continuous, and λ bounds the current-policy flow Φ_τ as in Assumption B.9. For each initial state x, also assume

$$\int_0^\infty e^{-\beta\tau}|g_t(\Psi_\tau(x))|\,d\tau < \infty, \qquad \lim_{T\to\infty} e^{-\beta T}\Delta_t(\Psi_T(x)) = 0. \tag{53}$$

These conditions concern the old-policy flow, while the corresponding conditions in Lemma 4.2 concern the current policy. Define $\widetilde{\Delta}_t(x) = \int_0^\infty e^{-\beta\tau} g_t(\Psi_\tau(x))\,d\tau$ as the population correction along the old-policy flow. Integrating $\beta\Delta_t - \nabla\Delta_t^\top f_{\pi_{t-1}} = g_t + \nabla\Delta_t^\top b_t$ along Ψ_τ yields

$$\Delta_t(x) - \widetilde{\Delta}_t(x) = \int_0^\infty e^{-\beta\tau} b_t(\Psi_\tau(x))^\top \nabla\Delta_t(\Psi_\tau(x))\,d\tau. \tag{54}$$

For a smooth policy update of size η_{pol}, suppose $\|b_t\|_\infty \leq B\eta_{\mathrm{pol}}$ and $\|\nabla g_t\|_\infty \leq G\eta_{\mathrm{pol}}$, with B, G independent of η_{pol}. Here η_{pol} is the perturbation scale, distinct from the ratio-deviation bound η_ρ

and the critic learning rate η_ϕ. Proposition 4.4 then gives

$$\|\Delta_t - \widetilde{\Delta}_t\|_\infty \le \frac{\|b_t\|_\infty \|\nabla g_t\|_\infty}{\beta(\beta-\lambda)} \le \frac{BG}{\beta(\beta-\lambda)}\eta_{\mathrm{pol}}^2. \tag{55}$$

For fixed $\beta > 0$ and a gap $\beta - \lambda$ bounded away from zero, the error remains second order. The factor $1/(\beta-\lambda)$ quantifies how the closed-loop expansion amplifies its coefficient. This complements the local policy approximations of Kakade & Langford (2002) and Tang et al. (2020) by making the dependence on the dynamics explicit. While Tang et al. (2020) formulate policy expansions using Q-values or advantages, ISPO directly constructs the value-shift correction from transition-based policy-evaluation PDE residuals and uses it to update the PPO critic target after each actor update.

Sampling variance. Retain the residual bound C_δ. All moments below are under π_{t-1}. Let $\mathcal{F}_{s+j}$ denote the history through state x_{s+j}, before sampling a_{s+j}; it includes the starting state x_s. Assume $\mathbb{E}_{\pi_{t-1}}[\rho_{s+j}^2 \mid \mathcal{F}_{s+j}] \le \kappa$ uniformly, with $\kappa \ge 1$. Repeated conditioning and $\mathbb{E}_{\pi_{t-1}}[\rho_{s+j} \mid \mathcal{F}_{s+j}] = 1$ give

$$\mathbb{E}_{\pi_{t-1}}[W_{s:l}^2 \mid x_s] \le \kappa^{l+1}, \qquad \mathbb{E}_{\pi_{t-1}}[Z_l^2 \mid x_s] \le C_\delta^2(\kappa - 1). \tag{56}$$

Applying Cauchy–Schwarz to the cross terms in the squared local correction without clipping yields

$$\mathrm{Var}_{\pi_{t-1}}(\widehat{\Delta}_s^{\mathrm{loc}} \mid x_s) \le \mathbb{E}_{\pi_{t-1}}[(\widehat{\Delta}_s^{\mathrm{loc}})^2 \mid x_s] \le \bar{s}^{\,2} C_\delta^2(\kappa - 1)\left(\frac{1-\gamma^H}{1-\gamma}\right)^2. \tag{57}$$

This bound allows dependence across time steps. It depends on the local ratio of the second moment and the discounted horizon, without the cumulative-weight factor κ^{l+1}. For uniformly bounded residuals, the variance bound vanishes as κ approaches one.

E Experimental Setup

E.1 Mechanism Verification

We provide the derivation of the analytical value function used in Section 6.1. Consider the scalar dynamics and reward

$$f(x,a) = -x + a, \qquad r(x,a) = -(x^2 + a^2), \tag{58}$$

under the Gaussian policy

$$\pi_{k,v}(a \mid x) = \mathcal{N}(kx, v). \tag{59}$$

Therefore,

$$\mathbb{E}[a \mid x] = kx, \qquad \mathbb{E}[a^2 \mid x] = k^2x^2 + v. \tag{60}$$

Substituting these quantities into the policy evaluation PDE in Equation 5 gives

$$\beta V^{k,v}(x) = -(1+k^2)x^2 - v + (k-1)x\frac{dV^{k,v}(x)}{dx}. \tag{61}$$

Since both the dynamics and reward are linear–quadratic, we seek a quadratic value function

$$V^{k,v}(x) = -Px^2 - C, \tag{62}$$

where P and C are constants. Then

$$\frac{dV^{k,v}(x)}{dx} = -2Px. \tag{63}$$

Substituting this expression into the evaluation equation yields

$$-\beta P x^2 - \beta C = -(1+k^2)x^2 - v - 2P(k-1)x^2. \tag{64}$$

Matching the coefficients of x^2 and the constant terms gives

$$P = \frac{1+k^2}{\beta + 2 - 2k}, \qquad C = \frac{v}{\beta}. \tag{65}$$

Hence,

$$V^{k,v}(x) = -\frac{(1+k^2)x^2}{\beta + 2 - 2k} - \frac{v}{\beta}, \tag{66}$$

which gives Equation 21. The discounted value is finite provided $\beta + 2 - 2k > 0$, which holds for all parameter settings used in our experiments.

E.2 Main Experiment: MuJoCo

Table 5 summarizes all hyperparameters used in our MuJoCo experiments. Unless otherwise specified, we use the same hyperparameters across all environments and baselines to ensure a fair comparison. We follow Gymnasium's default reward specifications and termination criteria across all tasks. We train each method using 5 independent random seeds and report the mean return and one standard deviation across seeds. The results in Table 1 are computed using the final checkpoint.

Table 5: Hyperparameters used in the MuJoCo experiments.

Hyperparameter	**Value**
Total environment steps	1×10^6
Rollout length	2,048
Minibatch size	256
Actor hidden layers	$[64, 64]$
Critic hidden layers	$[64, 64]$
Activation function	`Tanh`
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clip coefficient ϵ	0.2
Actor learning rate	3×10^{-4}
Critic learning rate	1×10^{-3}
PPO epochs K	10
Evaluation interval	20,000 steps
Evaluation episodes	10
Number of random seeds	5

For baseline-specific hyperparameters, we adopt the configurations recommended in their original papers. Moreover, for the VPG-repeat-50 baseline, a core claim of its original work (Wang et al., 2025) is that VPG achieves performance competitive with PPO when paired with multi-step critic updates. We therefore follow the author-recommended setup by applying 50 separate full-batch critic updates, rather than using the configuration listed in Table 5.

For ISPO, Table 6 shows the correction scales α and ratio-clipping threshold ϵ_v we used in each environment.

Table 6: α and ϵ_v used in the MuJoCo experiments.

Environment	α	$\epsilon_{\mathbf{v}}$
HalfCheetah	0.50	0.2
Swimmer	3.00	0.2
Hopper	1.00	0.2
Walker2D	1.00	0.2
Ant	0.25	0.2
Reacher	1.00	0.2

E.3 Main Experiment: Embodied Manipulation

Current robotic policy architectures in academia generally fall into two categories (Zhong et al., 2025). The first directly adopts the native autoregressive Transformer paradigm of LLMs (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025), formatting robot actions into reserved special tokens. This discrete tokenization scheme inherently introduces a decoupling from continuous robot action spaces. The second category employs a representation backbone coupled with a dedicated action head (Black et al., 2025b;a; Bjorck et al., 2025), structurally ensuring temporal alignment. We select GR00T N1.5 as our foundation model because it features a clean architecture belonging to the second paradigm, with demonstrated reliability across diverse open-source benchmarks.

GR00T N1.5 (Bjorck et al., 2025) is a cross-embodiment VLA model from NVIDIA's open-source GR00T family, retaining N1's core architecture while upgrading its vision-language model (Chen et al., 2025). Trained on mixed datasets that incorporate synthetic data, N1.5 is reported by NVIDIA to deliver superior performance, though these architectural modifications inevitably increase the complexity of post-training. Training this policy needs high computational costs. Thus, we employed $8\times$ NVIDIA H200 GPUs. Following the experimental parameters detailed in Table 7 , the actual training required roughly 171 GPU hours.

Table 7: Default hyperparameters for GR00T N1.5 on the four LIBERO suites.

Hyperparameter	Spatial	Object	Goal	Long
Train epochs	800	800	800	800
Batch size	1024	1024	1024	1024
Update epochs	4	1	1	1
Actor lr	5×10^{-6}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Value lr	1×10^{-4}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Interaction steps	240	240	320	480
Parallel environments	64	64	64	64
Rollout epochs	8	8	8	8
Action chunk C	5	5	5	10
Denoise steps	4	4	4	4
Noise level σ	0.3	0.3	0.3	0.3

F Additional Experiments and Results

F.1 Runtime Comparison

We compare ISPO and PPO on `HalfCheetah`, running each method once in the same environment with one PyTorch thread. Both methods use an identical setup as the main experiment. Timings include rollout collection and all training updates, but exclude initialization, evaluation, and logging. We also apply several engineering optimizations to ISPO-specific operations to reduce incremental overhead, including skipping the unused GAE refresh after the final actor epoch and performing the value-shift recursion in NumPy on the CPU to reduce GPU kernel-launch overhead. These operations minimize the computational overhead introduced by ISPO while keeping the baseline PPO cost unchanged, ensuring a fair comparison.

Table 8: Runtime on `HalfCheetah`.

Component	PPO (s)	ISPO (s)	Ratio
Rollout collection	84.21	79.63	0.95×
Training updates	42.89	51.71	1.21×
Total	127.10	131.34	1.03×

As shown in Table 8, ISPO increases total training time by 3.3% in this run. The additional cost lies in the update stage, which includes state-gradient computation, value-target correction, and GAE refresh. These results describe fixed-budget runtime for this implementation.

F.2 Comparison with TD Residual

Regarding our formulated PDE residual term, a natural question arises as to whether one can directly employ TD to approximate Equation 17. The PDE residual directly follows the value-shift analysis and uses the local value gradient to capture how actions affect value evolution. Compared with the finite-step TD residual, this local linearization may reduce sensitivity to inaccurate critic curvature, provided that the local gradient is reliable. We therefore compare the PDE residual in Equation 17

with its TD counterpart on `HalfCheetah`. Specifically, we construct an ISPO variant, denoted as ISPO-TD, by replacing the local residual with

$$\widehat{\delta}^{\mathrm{TD}}_{t-1,s} = \frac{R_s + \gamma V_{\phi_{\mathrm{old}}}(x'_s) - V_{\phi_{\mathrm{old}}}(x_s)}{\bar{s}}. \tag{67}$$

The factor $1/\bar{s}$ matches the scale of Equation 17. Both variants use the same frozen reference critic $V_{\phi_{\mathrm{old}}}$, importance-ratio clipping, value-shift recursion, corrected critic objective, and GAE refresh. Thus, this comparison isolates the choice of local residual. Both methods use an identical setup as the main experiment in Appendix E.2, with 1×10^6 environment steps, $\alpha = 0.5$, and $\epsilon_v = 0.2$. We use the same 5 random seeds and report the mean return and one standard deviation across seeds at the final checkpoint.

Table 9: Comparison of local residuals on `HalfCheetah`.

Method	**Return** (Mean $\pm$ 1 Std.)
ISPO-TD	1987 ± 617
ISPO-PDE	2121 ± 523

As shown in Table 9, ISPO-PDE achieves a higher mean return than ISPO-TD on `HalfCheetah` (2121 ± 523 vs. 1987 ± 617), with a relative improvement of 6.7% and lower variability across seeds. These results suggest a modest empirical advantage of the PDE residual in this setting.