Local Guidance, Global Impact: Gaussian-Reshaped Trust Region Unlocks Behavior Transitions

Bingxu Liu[1,4*] **Jiashun Liu**[1*] **Johan Obando-Ceron**[2,3] **Hao Wang**[5] **Runze Liu**[1]
Pablo Samuel Castro[2,3] **Aaron Courville**[2,3] **Ling Pan**[1†]
[1]Hong Kong University of Science and Technology
[2]Mila – Québec AI Institute [3]Université de Montréal
[4]Fudan University [5]City University of Hong Kong

Abstract

While Proximal Policy Optimization (PPO) demonstrates strong performance in stationary settings, we show that its standard optimization paradigm struggles in continual and non-stationary environments. The failure does not stem from insufficient model capacity or overly restrictive clipping. Instead, PPO performs persistent, directionally inefficient local updates, which indicates a lack of geometry-aware guidance for accumulating meaningful behavioral change and ultimately hindering transitions toward new behavior patterns. Although divergence-based regularization introduces partial geometric awareness, its monotonically increasing penalties implicitly discourage large policy deviations, even when such shifts are necessary for effective adaptation. To address this limitation, we propose **Gaussian Trust Region Policy Optimization** (`GTR`), which reshapes the trust region using a Gaussian kernel. The resulting constraint is bounded and non-monotonic, providing strong local stability while progressively relaxing under sustained high-advantage updates. To further improve robustness, we introduce a *Mixture Gaussian Anchor* that adapts to recent policy trajectories, reducing variance induced by stale references. `GTR` is architecture-agnostic and achieves strong performance across games, simulated robotic control, open-world exploration, and language model post-training. These results demonstrate that geometry-aware trust-region design can be a promising direction for robust reinforcement learning in complex non-stationary environments. Our code is available here.

1 Introduction

Proximal Policy Optimization (PPO) [Schulman et al., 2017], a standard deep reinforcement learning (RL) algorithm [Sutton et al., 1998], has achieved significant success across diverse domains, including robotic control [Hwangbo et al., 2019, Akkaya et al., 2019], game AI [Berner et al., 2019], and large foundation model post-training [Ouyang et al., 2022, Guo et al., 2025]. Its success largely stems from a simple yet effective principle: constraining policy updates to remain close to a reference policy via a clipping mechanism [Schulman et al., 2017], ensuring stable improvement under advantage-driven optimization [Kakade and Langford, 2002, Schulman et al., 2015]. This local-update paradigm is well-suited to traditional, stationary environments, where progress can often be made through incremental refinement around the current behavioral mode [Andrychowicz et al., 2021].

However, as modern RL systems move toward continual, non-stationary regimes [Wołczyk et al., 2021, Khetarpal et al., 2022] with evolving dynamics, this paradigm requires rethinking in the context of continual RL [Abel et al., 2023, Willi et al., 2024, Tang et al., 2025, Pan et al., 2025]. When tasks or required skills evolve over time, effective learning requires not only stable local improvement,

*Equal contribution.
†Correspondence to `lingpan@ust.hk`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

but also the ability to leave outdated behavioral modes and transition toward newly advantageous ones [Agarwal et al., 2021]. In such dynamic scenarios, the clipped objective may provide insufficient geometric preference among competing update directions, and we observe that standard PPO fails to achieve such effective transitions. As a result, the policy can undergo persistent but poorly directed local optimization and eventual collapse, as illustrated in Figure 1 (Top). This failure highlights a fundamental limitation of standard PPO in continual learning: *it lacks a mechanism for shaping local updates according to the distributional geometry induced by the policy, making advantage-driven optimization prone to unproductive drift rather than meaningful behavioral transition.*

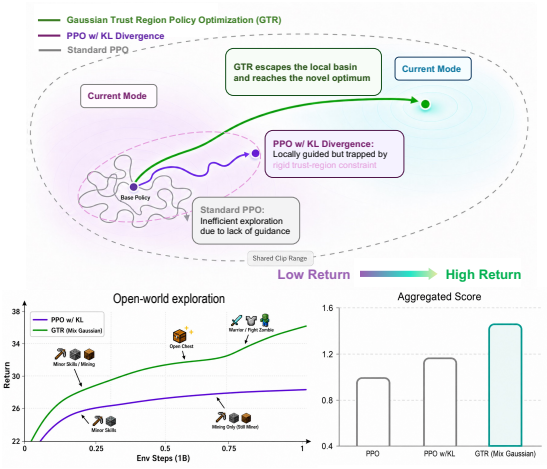

Figure 1: **Trust-region geometry shapes continual adaptation. Top:** Standard PPO can make persistent but ineffective local updates, while monotone divergence penalties can impede transitions to a new behavioral mode. **Bottom:** GTR enables behavioral transitions in open-world exploration and achieves the best aggregate performance across the evaluated benchmarks.

Intuitively, effective guidance should shape the local optimization geometry [Kakade, 2001] for distinguishing directions based on how sensitively policy distributions respond to parameter perturbations. This can be characterized by the Fisher Information Matrix (FIM): high-curvature directions induce large distributional shifts, while low-curvature directions correspond to relatively stable updates [Amari, 1998]. Divergence-based regularization introduces partial landscape awareness through its connection to the FIM, e.g., the second-order expansion of KL divergence recovers the Fisher metric [Schulman et al., 2015]. Empirically, we find that incorporating divergence-style penalties improves stability and alleviates inefficient exploration, suggesting that the absence of geometric guidance is a key source of adaptation failure in standard PPO. Nevertheless, most divergence penalties increase monotonically with policy deviation, implicitly assuming that larger shifts are undesirable [Nowozin et al., 2016]. While reasonable in stationary settings, this assumption becomes restrictive in continual RL, where optimal behaviors may lie far from the current policy [Lecarpentier et al., 2021]. Therefore, overly conservative regularization can hinder necessary mode transitions as shown in Figure 1 (Top), raising a key question: **can we design a trust-region regularization that is stiff enough to suppress noisy local drift, yet permissive enough to allow sustained high-advantage updates to escape the current behavioral mode?**

Idealized trust-region regularization should enforce stiffness around small deviations to stabilize incremental learning, while gradually relaxing as the policy is consistently driven toward high-advantage situations. This balance prevents unstable local updates while avoiding over-constraining exploration when strong advantage signals suggest departure from suboptimal behaviors. Guided by this principle, we introduce a Gaussian-shaped trust-region constraint, which induces localized stiffness near the reference policy and a rapidly diminishing penalty for larger deviations. This design also encodes a geometry-aware regularization: the Gaussian weighting modulates the influence of policy ratios in a distance-dependent manner, effectively reweighting gradient contributions and shaping the local optimization landscape. In contrast to monotonic divergence penalties, this Gaussian penalty allows the policy to preserve stability in the extreme proximal region while reducing resistance to sustained, high-advantage updates, thereby facilitating controlled exploration. However, as the reference policy becomes stale under a high-data-reuse setting, a fixed Gaussian anchor can be unstable. Therefore, we propose *Mixture Gaussian Anchor*, which models the constraint as a dynamically updated mixture over the recent policy snapshots, to reduce variance induced by outdated references while remaining responsive to recent high-advantage updates. We instantiate this framework as **Gaussian Trust Region Policy Optimization** (GTR).

GTR improves adaptability at the algorithmic level and generalizes seamlessly across diverse policy classes, i.e., MLPs [Mnih et al., 2015, Schwarzer et al., 2023, Obando Ceron et al., 2024], RNNs [Chung et al., 2014], SimBa [Lee et al., 2025], and Transformer-based [Vaswani et al., 2017] policies.

Extensive evaluations on robotic control [Todorov et al., 2012, Tassa et al., 2018], open-world exploration [Matthews et al., 2024], and language model post-training [Yu et al., 2025, Guo et al., 2025] demonstrate that GTR consistently achieves state-of-the-art performance, as shown in Figure 1, highlighting that principled constraint design could be a promising way for policy improvement. Our contributions are threefold:

1. We reveal that a key factor for the collapse of PPO in continual RL is that it exhibits persistent but directionally inefficient local updates, which indicates a lack of local optimization geometry.
2. We introduce GTR, a Gaussian-shaped, geometry-aware trust-region regularization that balances local stability and global adaptability in continual setups.
3. We extensively validate GTR across a wide range of benchmarks and various architectures. GTR consistently improves performance in non-stationary and continual learning settings.

2 Preliminaries

Proximal Policy Optimization Standard Proximal Policy Optimization (PPO) [Schulman et al., 2017] is a policy gradient method that improves training stability by constraining policy updates within a proximal region of a reference policy. Let π_θ denote the current policy and reference policy $\pi_{\theta_{\text{old}}}$ used for data collection. PPO optimizes a surrogate objective based on importance sampling: $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, where $r_t(\theta)$ is the likelihood ratio. The clipped objective is defined as: $L^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$, where $\hat{A}_t$ is an estimate of the advantage function and ϵ is a hyperparameter controlling the trust region size. However, the clipping mechanism does not explicitly account for the geometry of the policy space, as it constrains updates based solely on the likelihood ratio without incorporating curvature information.

Trust region regularization can be used as directional guidance We consider a regularized policy optimization objective of the standard form [Schulman et al., 2015]:

$$\max_\theta \quad \mathcal{L}(\theta) = \mathbb{E}\big[r_t(\theta)\hat{A}_t\big] - \lambda\, D(\pi_{\theta_{\text{old}}} \,\|\, \pi_\theta), \tag{1}$$

where $D(\cdot)$ is a divergence measure. Under a small update $\theta = \theta_{\text{old}} + \Delta\theta$, if D is the KL divergence, the objective admits a second-order approximation:

$$\mathcal{L}(\theta) \approx g^\top \Delta\theta - \tfrac{\lambda}{2}\Delta\theta^\top F \Delta\theta, \tag{2}$$

where $g = \mathbb{E}\big[\hat{A}_t \nabla_\theta \log \pi_\theta(a_t|s_t)\big], \quad F = \mathbb{E}\big[\nabla_\theta \log \pi_\theta \nabla_\theta \log \pi_\theta^\top\big]$ is the Fisher Information Matrix (FIM). The optimal update is given by: $\Delta\theta^* = \frac{1}{\lambda}F^{-1}g$. Compared to standard policy gradients $\Delta\theta \propto g$, trust-region regularization transforms the update into a geometry-aware form $F^{-1}g$, where the FIM induces anisotropic scaling across directions. In particular, directions associated with high curvature (large eigenvalues of F) are suppressed, while low-curvature directions are amplified.

This shows that trust-region regularization acts as a *directional filter*, guiding updates based on the local geometry of the policy space rather than purely on scalar advantage signals.

3 Trust Region Regularization Unlocks Behavior Transition

We first investigate the root cause of PPO's limited behavioral pattern transition capability using a controlled sequential training setup inspired by [Muppidi et al., 2024, Abbas et al., 2023]. Specifically, we construct a curriculum on the challenging starpilot environment from Procgen [Cobbe et al., 2020], where task difficulty increases progressively (Level 1→Level 2→Level 3). We train a single MLP policy using standard PPO across the full sequence (PPO-sequential), and compare it to a control setting where PPO is trained from scratch on each task independently (PPO-single task). As shown in Figure 2 (Bottom), PPO successfully solves each task in isolation. However, under sequential training, although the policy rapidly learns the first task coincides with PPO-single task, it experiences a sharp performance collapse when transitioning to subsequent tasks. This indicates that the failure is not due to task difficulty alone, but arises from the interaction between PPO's update mechanism and non-stationary task shifts.

Network capacity is not the bottleneck. To test whether the failure is caused by limited model expressiveness, we incorporate stochastic network perturbations (SnP) [Ash and Adams, 2020] to continually refresh network capacity (PPO-sequential-SnP) (Figure 2 (Middle)). Despite improved performance on the first task, the policy still collapses, suggesting that insufficient capacity is not the primary cause.

Larger clipping ranges fail to recover the policy. A natural hypothesis is that the clipping constraint restricts the policy to a region too small to capture new behaviors [Wang et al., 2019]. To evaluate this, we relax the clipping bound (PPO w/ relaxed clip) and also consider schedules where the clipping range grows over time (PPO w/ gradually relaxed clip). Neither modification recovers continual learning ability, indicating that clipping is not the root cause (Figure 2 (Bottom)).

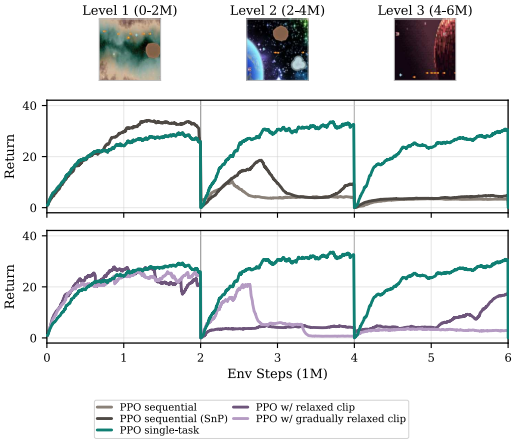

Figure 2: **PPO collapses under sequential task shifts. Top:** Procgen levels. **Middle:** Network perturbations do not prevent collapse. **Bottom:** Increasing the clipping range does not restore continual learning.

Failure arises from aimless local optimization. We further examine the dynamics preceding collapse. As shown in Figure 3 (Left), the policy shift ratio exhibits high-frequency fluctuations before failure, while gradient magnitudes remain significant. This suggests that the policy is actively updating and exploring in parameter space. However, these updates fail to produce consistent improvement, indicating that exploration is directionally inefficient despite being persistent.

Trust-region regularization is critical for effective adaptation. The above results indicate that PPO fails in continual settings not due to insufficient model capacity or overly restricted update magnitude, but because it lacks the navigation in complex local optimization landscapes. In particular, the absence of directional guidance leads to inefficient exploration, causing the accumulation of biased local updates and resulting in training collapse. This suggests that the core issue lies in how trust-region regularization shapes the update geometry, rather than in the scale of the update itself.

To test this hypothesis, we introduce divergence-based penalties, which provide partial landscape awareness through their connection to the Fisher Information Matrix (FIM). Unlike standard PPO, which relies solely on scalar advantage weighting, these penalties incorporate curvature information that differentiates update directions based on their induced distributional shifts. As shown in Figure 3 (Middle), both inverse KL [Tang and Munos, 2025, Shah et al., 2025] and Pearson χ^2 [Xie et al., 2024] penalties lead to immediate and consistent performance improvements over standard PPO in sequential Procgen training (Figure 3, right). These results provide empirical evidence that incorporating geometry-aware regularization significantly improves local optimization quality. By guiding updates toward directions that yield stable and meaningful policy changes, such regularization

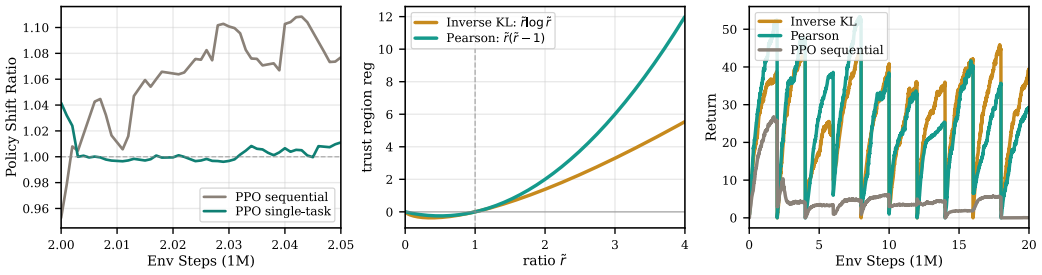

Figure 3: **Trust-region regularization improves continual adaptation. Left:** PPO exhibits larger policy shifts after task transitions. **Middle:** Penalty strength as a function of the policy ratio. **Right:** Mild inverse-KL and Pearson χ^2 penalties improve adaptation in sequential training (Levels 1–10).

enables iterative improvements to accumulate, resulting in gradual behavioral transitions and sustained performance in continual learning.

4 Reshape the Trust Region via Gaussian Kernel

As section 3 shows, incorporating divergence-based penalties can achieve behavior transition by providing geometry-aware guidance for local optimization. However, these approaches remain fundamentally limited. Most divergence-based penalties increase monotonically with policy deviation, implicitly enforcing a linear or unbounded growth of constraint strength as the policy moves away from the reference. In non-stationary environments, effective adaptation often requires transitioning toward behavior modes that are substantially different from the current policy. In such cases, monotonically increasing penalties can overly suppress exploration precisely when large, consistent updates are needed. This observation suggests that an effective trust-region constraint for continual learning should not only encode local geometry, but also relax its restriction in a controlled manner as the policy moves toward promising regions. In other words, the regularization should provide strong local stability while avoiding excessive resistance to sustained, high-quality updates far away from the current reference policy.

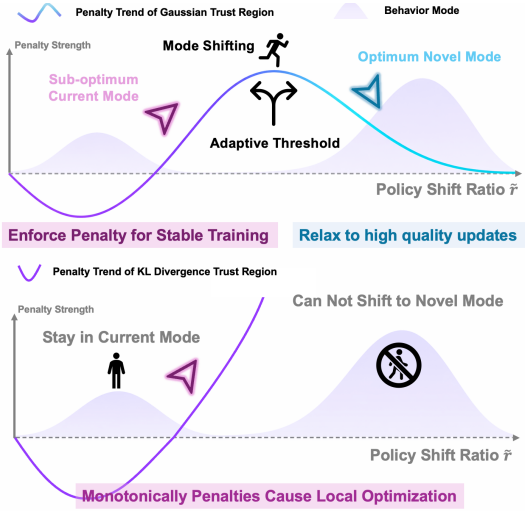

Figure 4: **Reshaping the trust region.** Unlike KL divergence, the Gaussian constraint preserves local stability while allowing large, high-advantage policy shifts.

To address this limitation, we propose to reshape the trust region using a Gaussian kernel. This formulation induces a bounded, non-monotonic constraint that preserves local geometry while allowing controlled exploration beyond the immediate neighborhood, enabling efficient navigation of complex optimization landscapes: (i) **Forgetting Mechanism:** $\lambda(\tilde{r}) \to 0$ as $\tilde{r} \to 0$. When an action is no longer sampled, no regularization gradient should be applied to it. (ii) **Unchanged Policy:** $\lambda(\tilde{r}) \to 0$ as $\tilde{r} \to 1$. No regularization is enforced when the policy remains unchanged. (iii) **Local Trust Region:** $\frac{\lambda(\tilde{r})}{\tilde{r}} \approx \pm\epsilon$ when $\tilde{r} = 1 \pm \epsilon$, preserving first-order trust-region behavior and endowing the regularization term with the functionality of a trust region, analogous to the ϵ-alignment [Xie et al., 2024]. (iv) **Vanishing Constraint at Scale:** $\lambda(\tilde{r}) \to 0$ as $\tilde{r} \gg 1$, allowing sustained, high-advantage updates to escape local modes. The complete derivation is deferred to subsection A.2.

Guided by these principles, we introduce a Gaussian decay kernel (as shown in Figure 4 (Top)), yielding $\lambda_{\text{Gaussian}}(\tilde{r}) = \tilde{r}(\tilde{r}-1)e^{-(\tilde{r}-1)^2}$, which satisfies localized stiffness near $\tilde{r} \approx 1$ while inducing rapidly vanishing gradients for large deviations. This behavior allows policy updates to remain stable in local regions while avoiding excessive penalization of large, high-advantage shifts. Remarkably, this formulation admits an elegant anti-derivative:

$$L_{\text{Gaussian}} = \frac{1}{2}(1 - e^{-(\tilde{r}-1)^2}) \tag{3}$$

This is a plug-and-play module that can be effortlessly integrated into standard PPO. We term this idealized approach which aligns with the theory as **Gaussian Trust Region Policy Optimization** (GTR), with the objective function formulated as follows:

$$L_{\text{GTR}} = \mathbb{E}_t \left\{ -\min\left[r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right] + \frac{\eta}{2}\left(1 - e^{-(r_t-1)^2}\right) \right\} \tag{4}$$

Sliding Trust Region via Gaussian Mixtures While the single Gaussian kernel improves the functional shape of the trust region, it does not consider another implicit issue: During multi-epoch updates, the trust-region anchor remains tied to the behavior policy, causing likelihood ratios to become increasingly unstable as the policy drifts away. This issue is further exacerbated under high

UTD regimes, where repeated reuse of stale data leads to biased updates and degraded adaptation [Gan et al., 2024, Li et al., 2025]. We reinterpret it as a consequence of a *rigid trust region*. The constraint remains anchored to an outdated policy and fails to account for the progressive shift of the policy distribution throughout optimization. To address this limitation, we propose a *Mixture Gaussian Anchor*, which generalizes the trust-region constraint by replacing the single reference policy with a mixture over recent policy snapshots. Concretely, we compute the trust-region penalty using multiple intermediate policies from different PPO epochs and average their contributions. This yields a trajectory-aware constraint that tracks the progressive shift of the policy distribution, mitigating the effect of stale references while preserving sensitivity to recent updates. Moreover, due to the vanishing tails of the Gaussian kernel, the resulting regularization is inherently bounded and avoids excessive variance. From a theoretical standpoint, this mixture Gaussian regularization directly shapes the policy update through its adaptive, smoothly decaying penalty:

$$L_{\text{MixGaussian}} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\pi_i(a \mid s)}{\mu(a \mid s)} \cdot \frac{1}{2}\left[1 - \exp\left(-\left(\frac{\pi_\theta(a \mid s)}{\pi_i(a \mid s)} - 1\right)^2\right)\right]\right) \tag{5}$$

where μ denotes the behavior policy, π_i represents the policy at the i-th epoch (with $\pi_1 = \mu$), and π_θ is the current policy. The final GTR objective is thus given by:

$$L_{\text{GTR}} = \mathbb{E}_t\left\{-\min\left[\frac{\pi_\theta(a \mid s)}{\mu(a \mid s)}\hat{A}_t, \text{clip}\left(\frac{\pi_\theta(a \mid s)}{\mu(a \mid s)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right] + \eta L_{\text{MixGaussian}}\right\} \tag{6}$$

Fisher Geometry along the Optimization Path. From an optimization perspective, we construct a trust region along *the path of policy improvement* [Dabney et al., 2021]. From a gradient perspective, the probability measures of reference policies are unified to the current parameterized policy π_θ via the Radon-Nikodym derivative (i.e. the importance weight). Concurrently, a Gaussian kernel is applied to these derivatives to smoothly penalize deviations, yielding:

$$\nabla_\theta L_{\text{MixGaussian}} = \mathbb{E}_{\pi_\theta}\left[\frac{1}{K}\sum_{i=1}^{K}\underbrace{\exp\left(-\left(\frac{\pi_\theta(a \mid s)}{\pi_i(a \mid s)} - 1\right)^2\right)}_{\text{Gaussian Kernel}} \cdot \underbrace{\left(\frac{\pi_\theta(a \mid s)}{\pi_i(a \mid s)} - 1\right)}_{\text{Trust Region Penalty}} \nabla_\theta \log \pi_\theta\right] \tag{7}$$

Our approach can thus be viewed as approximating the effective FIM via a Gaussian-weighted aggregation of multiple local FIMs along the optimization trajectory.

5 Experiments

We evaluate GTR across a diverse spectrum of architectures and tasks ranging in complexity, to determine whether continual learning ability is effectively recovered. Our evaluation spans three distinct domains: (i) Robotic Control: Evaluating continual learning performance and generalization across tasks. (ii) Open-World Exploration: Assessing adaptability under prolonged training with high data staleness. (iii) Language Reasoning Model Post-Training: Demonstrating the efficacy of GTR on large-scale models, highlighting the mode transition issue in such domain. (iv) Ablation Study: Investigating the robustness of GTR to relaxed clipping and examining whether adjusting KL's penalty strength can replicate the effect of GTR.

5.1 Robotic Control: Continual Adaptation with SimBa Policies

Experimental Setup We compare GTR with standard PPO, PPO with a KL-based trust-region constraint and a recent PPO variant, SPO [Xie et al., 2024]. All results are averaged over three random seeds, and each method is tuned with empirically optimal hyperparameters (details in Appendix B). We adopt the same SimBa [Lee et al., 2025] architecture (advanced scalable residual blocks with LayerNorm) to parameterize policies to ensure baseline performance and eliminate the performance degradation caused by network expressiveness. To comprehensively evaluate continual learning ability, we construct two types of benchmarks:

Related Tasks. This setting evaluates the ability to adapt between tasks with shared action and observation spaces (use the same embodied skeleton) but differing task goals, capturing smooth behavioral mode transitions. We build two difficult locomotion benchmarks via DeepMind Control Suite [Tassa et al., 2018]: (i) Walker: *walker stand→walk→run* and (ii) Dog: *dog stand→walk→run→trot*, repeated for two cycles.

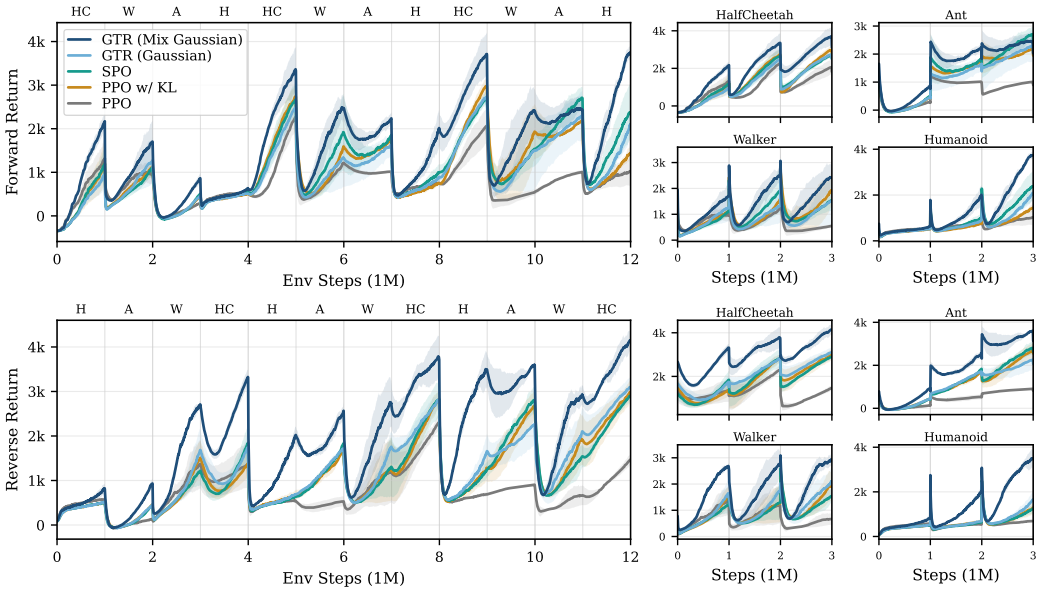

Figure 6: **Continual control across distinct tasks. Top:** Forward training order. **Bottom:** Reverse order. H: Humanoid; A: Ant; W: Walker; HC: HalfCheetah. GTR consistently achieves the best return across training cycles.

Results on Related Tasks. As shown in Figure 5 (first row), GTR consistently achieves the best performance across all benchmarks and outperforms strong baselines by up to 25% in peak performance. GTR exhibits substantially faster adaptation when switching tasks, as evidenced by the rapid rise in performance immediately after task transitions. This indicates more efficient behavioral mode switching. Further analysis of policy entropy (Figure 5 (second row)) shows that GTR maintains higher entropy during training. This suggests that the reshaped trust region preserves exploration capacity and enables efficient mode seeking in a huge solution space.

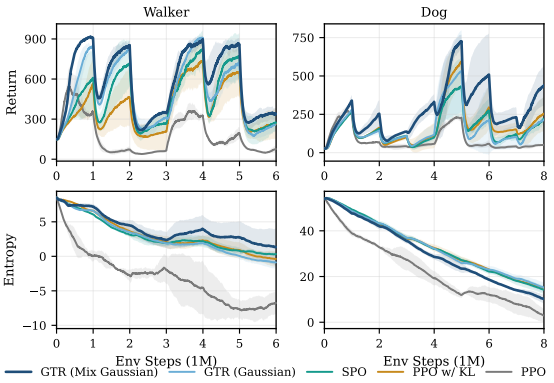

Figure 5: **Continual control with SimBa policies. Top:** Episodic return on Walker and Dog. **Bottom:** Policy entropy. Results are averaged over three seeds.

Distinct Tasks across Scenarios. This setting evaluates adaptation under significant distribution shifts, including differences in task logic, observation spaces, and action spaces. It poses a more challenging test of policy plasticity. We construct two benchmarks using MuJoCo-v5 tasks [Todorov et al., 2012], *HalfCheetah→Walker→Ant→Humanoid*, along with the reversed order, trained sequentially over three cycles.

Results on Distinct Tasks. When task differences become more significant, the benefits of GTR are further amplified. Comparing Figure 6 (left) with Figure 5, we observe that increased difficulty in behavior mode transitions leads to larger performance gaps between GTR and baseline methods. GTR consistently achieves superior performance under both forward and reverse training orders, demonstrating robustness to curriculum variations. This indicates that the method is less sensitive to task ordering and does not rely on favorable training sequences. A per-task analysis (Figure 6 (right)) further reveals that the mixture Gaussian trust region enables more precise modeling of behavioral boundaries. As a result, performance improves progressively across training cycles, with most tasks surpassing their previous best scores. Since state and action spaces are different across tasks, we employ task-specific input layers for both actors and critics and task-specific output heads for actors. Such layers account for a small fraction of the total parameters.

5.2 Open-World Exploration: Learning New Skills during Prolonged Training

Experimental Setup To evaluate whether GTR can achieve obvious behavior mode transition and stable training in prolonged training, we conduct an extremely long-horizon experiment in a challenging Minecraft-style open-world environment, Craftax [Matthews et al., 2024], which features a rich set of compositional subtasks and diverse behavioral patterns. We consider two architectures: (i) SimBa, an advanced residual network, following the official implementation and (ii) a GRU backbone with LayerNorm [Ba et al., 2016, Nauman et al., 2024], following the implementation in Matthews et al. [2024]. Agents are trained for 1 billion steps to test the ability of behavior transition.

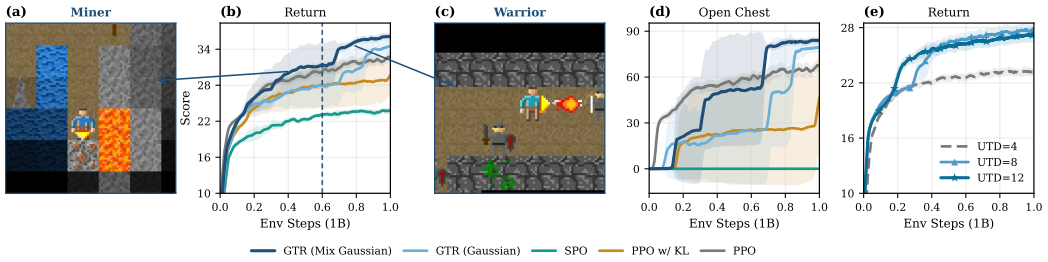

Figure 7: **Skill transitions and prolonged learning in Craftax. (a–c)** GRU-based PPO; arrows mark the transition from miner to warrior. **(d)** *Open-chest* achievement. **(e)** Update-to-data ratio ablation with SimBa. Results are averaged over three seeds.

Results In experiments with GRU-based PPO (Figure 7(b)), we identify *open-chest* (recorded in Figure 7(d)) as a critical milestone highly predictive of ultimate performance. Prior to mastering this skill, the agent defaults to a conservative mining strategy (termed the *old behavior*). Upon acquiring *open-chest*, the agent secures sufficient resources to hunt formidable monsters, effectively transitioning into a "warrior" (the *new behavior*). We observe an intriguing dynamic: while standard PPO discovers *open-chest* early in training, accumulated inefficiencies in local optimization leave it entrenched in the old pattern. SPO, hampered by its aggressive regularization against outliers, completely fails to transition into the new pattern, converging to a suboptimal policy. Conversely, although GTR learns *open-chest* slightly later in the training process, its exceptional learning capacity enables a rapid behavioral shift into the new mode, ultimately yielding the highest return.

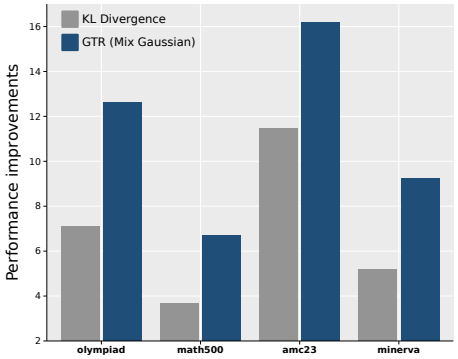

Figure 8: **Language model post-training.** Score improvement over the initial policy on four tasks, with GRPO as the backbone for both methods.

We further investigate performance under high update-to-data (UTD) regimes using SimBa-based PPO (as shown in Figure 7(e)). As we increase the number of PPO epochs from 4 to 12, GTR not only maintains stable training, but also learns *open-chest* significantly faster and converges to higher returns. This suggests that GTR effectively leverages high-UTD updates to accelerate adaptation.

5.3 Language Model Post-Training: Adapting across Reasoning Tasks

We adopt GRPO [Guo et al., 2025], a simplified variant of PPO widely used in LLM post-training, as the basis, where group normalization replaces GAE for advantage estimation. We set $UTD = 2$ (equivalent to PPO epochs), introducing mild off-policy effects due to data reuse. Policies are initialized from `Qwen3-1.7B-Base`. To ensure reproducibility and fairness, we exclusively use open-source datasets in Yu et al. [2025]. We evaluate performance using pass@4 on four challenging benchmarks: MATH-500 [Lightman et al., 2023], OlympiadBench [He et al., 2024], MinervaMath [Lewkowycz et al., 2022], and AMC 2023 [Xue et al., 2025]. See subsection B.1 for detailed settings.

Results As shown in Figure 8, GTR achieves the best performance across all benchmarks. We attribute this improvement to the implicit continual learning nature of LLM post-training. In each

training batch, prompts correspond to diverse reasoning prompts, requiring the policy to continuously adapt across heterogeneous objectives while maintaining previously acquired capabilities. As a result, optimization dynamics naturally involve frequent shifts between different behavioral patterns, and local directional guidance plays a critical role in achieving strong reasoning performance.

5.4 Ablation Study

Relaxed Clipping. As illustrated in Figure 9 (Left), after increasing the clip range of GRU-based PPO in Craftax, although the convergence slows down, it does not affect the final performance of GTR. In contrast, PPO experiences performance degradation. This demonstrates that the Gaussian trust region is more reliable to guide the local landscape.

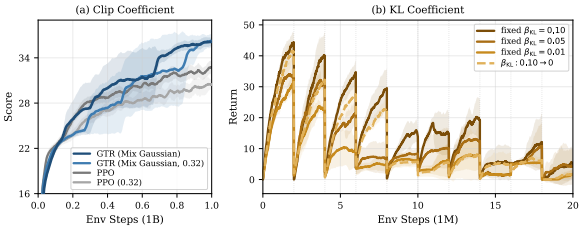

Figure 9: **Sensitivity to clipping and KL strength. Left:** GTR remains robust to a larger clipping range. **Right:** Adjusting the KL coefficient does not prevent PPO collapse.

KL Coefficient. As depicted in Figure 9 (Right), we evaluated three fixed KL coefficients alongside a linearly decaying KL schedule starting from 0.1. All configurations led to collapse in Procgen. This highlights the necessity of directly reshaping the trust region to unlock the continual learning ability.

6 Related Work

While stable dynamics and signals are theoretically crucial for algorithm convergence in reinforcement learning [Agarwal et al., 2021], numerous empirical practices in model-free RL have demonstrated the robustness of deep RL to such tasks, utilizing techniques like residual learning [Silver et al., 2018] and network ensembling [Ball et al., 2023]. Interestingly, these methods are entirely based on Q-learning [Watkins and Dayan, 1992]. Rather than contradicting foundational theories, these empirical techniques effectively stabilize the variance and mitigate non-stationarity, thereby satisfying the implicit requirements of classic stochastic approximation theory [Borkar and Borkar, 2008] and Robbins-Monro theorem [Robbins and Monro, 1951] in highly non-linear settings. Furthermore, the replay buffer mechanism in Q-learning naturally aligns with classic continual learning algorithms, such as EWC [Kirkpatrick et al., 2017] and PackNet [Mallya and Lazebnik, 2018], which has directly shaped the mainstream research trajectory of continual RL in recent years.

Compared to off-policy counterparts, research on continual learning based on PPO remains exceptionally scarce. Prior mainstream research has undergone a paradigm shift from learning dynamics [Lyle et al., 2022, Tang and Munos, 2023], generalization and interference [Bengio et al., 2020], towards plasticity [Dohare et al., 2024]. The latter specifically emphasizes the plasticity of neural networks, and existing methods have explored this from a diverse array of perspectives: plasticity injection [Nikishin et al., 2023], loss landscapes [Lyle et al., 2023, Castanyer et al., 2025, Pasand et al., 2026], regularization [Lyle et al., 2024b, Obando-Ceron et al., 2026], network weights [Elsayed et al., 2024], effective learning rates [Lyle et al., 2024a], dormant neurons [Sokar et al., 2023], the lottery ticket hypothesis [Graesser et al., 2022, Ceron et al., 2024, Liu et al., 2025a], gradient geometry [Liu et al., 2025b], and Neural Tangent Kernel (NTK) theory [Tang et al., 2025]. While fruitful, these approaches often blur the mechanistic distinctions between policy gradient methods and other RL algorithms, leaving the inherent continual learning potential of PPO itself largely overlooked.

Although our foundational work builds upon vanilla PPO, numerous exceptional variants have been derived from PPO at the application level, which may harbor even greater potential for continual learning. PPO-EWMA [Hilton et al., 2022] investigates asynchronous settings under limited computational resources, PPG [Cobbe et al., 2021] explores representation-level sharing and data efficiency, and GRPO [Guo et al., 2025] has successfully enabled the RL training of reasoning models. Additionally, the data collection pipeline [Mayor et al., 2025] itself warrants further consideration. We leave the exploration of these promising directions for future work.

7 Conclusion

In this work, we investigated the limitations of PPO in continual and non-stationary learning settings. Through systematic empirical analysis, we found that policy collapse is associated less with insufficient model capacity or clip range and more with ineffective local exploration. Without geometry-aware guidance, policy updates may become directionally misaligned, leading to the accumulation of biased local updates that hinder global adaptation. While divergence-based constraints introduce partial geometric structure, their monotonically increasing penalties may restrict transitions toward emerging behavior modes.

To address this limitation, we proposed `GTR`, a trust-region formulation based on a Gaussian kernel that balances local stability with controlled exploration. By further introducing a mixture-based anchor that adapts to evolving policies, `GTR` enables efficient local optimization while facilitating gradual and reliable behavioral transitions. Extensive experiments across diverse domains show that `GTR` consistently improves performance over PPO baselines, supporting the importance of trust-region geometry in non-stationary learning. We hope this work motivates further research on trust-region design as a foundation for more robust and adaptive reinforcement learning.

Limitations. First, due to computational constraints, we do not evaluate `GTR` in large-scale language model training or across varying LLM sizes; extending geometry-aware trust-region methods to such high-compute regimes remains an important direction for future work. Second, although the Gaussian-shaped trust-region formulation is empirically effective and well motivated, its design remains primarily grounded in empirical observations and intuition rather than a complete theoretical characterization. Developing a deeper theoretical understanding of geometry-aware and non-monotonic regularization is an important avenue for future research.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on lifelong learning agents*, pages 620–636. PMLR, 2023.

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407, 2023.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894, 2020.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. In *International Conference on Machine Learning*, pages 767–777. PMLR, 2020.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Roger Creus Castanyer, Johan Obando-Ceron, Lu Li, Pierre-Luc Bacon, Glen Berseth, Aaron Courville, and Pablo Samuel Castro. Stable gradients for stable learning at scale in deep reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Vqj65VeDOu.

Johan Samir Obando Ceron, Aaron Courville, and Pablo Samuel Castro. In value-based deep reinforcement learning, a pruned network is a good network. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=seo9V9QRZp.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020.

Karl W Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *International Conference on Machine Learning*, pages 2020–2027. PMLR, 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7160–7168, 2021.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Mohamed Elsayed, Qingfeng Lan, Clare Lyle, and A. Rupam Mahmood. Weight clipping for deep continual and reinforcement learning. In *Reinforcement Learning Conference*, 2024. URL https://openreview.net/forum?id=9Vagp8vnPa.

Yaozhong Gan, Renye Yan, Xiaoyang Tan, Zhe Wu, and Junliang Xing. Transductive off-policy proximal policy optimization. *arXiv preprint arXiv:2406.03894*, 2024.

Laura Graesser, Utku Evci, Erich Elsen, and Pablo Samuel Castro. The state of sparse training in deep reinforcement learning, 2022. URL https://arxiv.org/abs/2206.10369.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, 2024.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization. *Advances in Neural Information Processing Systems*, 35:17086–17098, 2022.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science robotics*, 4(26):eaau5872, 2019.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L Littman. Lipschitz lifelong reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8270–8278, 2021.

Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R. Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jXLiDKsuDo.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Xiaocan Li, Shiliang Wu, and Zheng Shen. A-3po: Accelerating asynchronous llm training with staleness-aware proximal policy approximation. *arXiv preprint arXiv:2512.06547*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Jiashun Liu, Johan Samir Obando Ceron, Aaron Courville, and Ling Pan. Neuroplastic expansion in deep reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=20qZK2T7fa.

Jiashun Liu, Zihao Wu, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Ling Pan. Measure gradients, not activations! enhancing neuronal activity in deep reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=FjNHmO39pp.

Clare Lyle, Mark Rowland, Will Dabney, Marta Kwiatkowska, and Yarin Gal. Learning dynamics and generalization in deep reinforcement learning. In *International conference on machine learning*, pages 14560–14581. PMLR, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, pages 23190–23211. PMLR, 2023.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado van Hasselt, Razvan Pascanu, and Will Dabney. Normalization and effective learning rates in reinforcement learning. *Advances in Neural Information Processing Systems*, 37:106440–106473, 2024a.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado Van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. *arXiv preprint arXiv:2402.18762*, 2024b.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, and Jakob Nicolaus Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hg4wXlrQCV.

Walter Mayor, Johan Obando-Ceron, Aaron Courville, and Pablo Samuel Castro. The impact of on-policy parallelized data collection on deep reinforcement learning networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=cnqyzuZhSo.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K.R. Zentner, Ryan Julian, J K Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-world+: An improved, standardized, RL benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=1de3azE606.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Aneesh Muppidi, Zhiyu Zhang, and Heng Yang. Pick up the pace: A parameter-free optimizer for lifelong reinforcement learning. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

Michal Nauman, Michał Bortkiewicz, Piotr Miłoś, Tomasz Trzcinski, Mateusz Ostaszewski, and Marek Cygan. Overestimation, overfitting, and plasticity in actor-critic: the bitter lesson of reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=5vZzmCeTYu.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and André Barreto. Deep reinforcement learning with plasticity injection. *Advances in Neural Information Processing Systems*, 36:37142–37159, 2023.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Johan Obando-Ceron, Walter Mayor, Samuel Lavoie, Scott Fujimoto, Aaron Courville, and Pablo Samuel Castro. Simplicial embeddings improve sample efficiency in actor–critic agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mCpq1GCKxA.

Johan Samir Obando Ceron, Ghada Sokar, Timon Willi, Clare Lyle, Jesse Farebrother, Jakob Nicolaus Foerster, Gintare Karolina Dziugaite, Doina Precup, and Pablo Samuel Castro. Mixtures of experts unlock parameter scaling for deep RL. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38520–38540. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/obando-ceron24b.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Chaofan Pan, Xin Yang, Yanhua Li, Wei Wei, Tianrui Li, Bo An, and Jiye Liang. A survey of continual reinforcement learning. *arXiv preprint arXiv:2506.21872*, 2025.

Ali Saheb Pasand, Johan Obando-Ceron, Aaron Courville, Pouya Bashivan, and Pablo Samuel Castro. Stable deep reinforcement learning via isotropic gaussian representations, 2026. URL https://arxiv.org/abs/2602.19373.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

Max Schwarzer, Johan Samir Obando Ceron, Aaron Courville, Marc G Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level Atari with human-level efficiency. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30365–30380. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/schwarzer23a.html.

Vedant Shah, Johan Obando-Ceron, Vineet Jain, Brian Bartoldson, Bhavya Kailkhura, Sarthak Mittal, Glen Berseth, Pablo Samuel Castro, Yoshua Bengio, Nikolay Malkin, et al. A comedy of estimators: On kl regularization in rl training of llms. *arXiv preprint arXiv:2512.21852*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pages 32145–32168. PMLR, 2023.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EkoFXfSauv.

Yunhao Tang and Rémi Munos. Towards a better understanding of representation dynamics under td-learning. In *International Conference on Machine Learning*, pages 33720–33738. PMLR, 2023.

Yunhao Tang and Rémi Munos. On a few pitfalls in kl divergence gradient estimation for rl. *arXiv preprint arXiv:2506.09477*, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

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*, pages 5026–5033. IEEE, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Timon Willi, Johan Samir Obando Ceron, Jakob Nicolaus Foerster, Gintare Karolina Dziugaite, and Pablo Samuel Castro. Mixture of experts in a mixture of RL settings. In *Reinforcement Learning Conference*, 2024. URL https://openreview.net/forum?id=5FFO6RlOEm.

Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28496–28510, 2021.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. *arXiv preprint arXiv:2401.16025*, 2024.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

A Extended Preliminaries and Derivations

A.1 Divergence and Fisher Information

Let $\pi_\theta(a \mid s)$ be a differentiable policy, and consider a local perturbation $\theta' = \theta + \Delta\theta$. The forward KL divergence between π_θ and $\pi_{\theta'}$ is defined as:

$$D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta'}) = \mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta}\left[\log \frac{\pi_\theta(a \mid s)}{\pi_{\theta'}(a \mid s)}\right]. \tag{8}$$

Proposition 1 (Second-Order Expansion of KL and FIM). For sufficiently small $\Delta\theta$, the KL divergence admits the following expansion:

$$D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta+\Delta\theta}) = \tfrac{1}{2}\,\Delta\theta^\top F(\theta)\,\Delta\theta + o(\|\Delta\theta\|^2), \tag{9}$$

where

$$F(\theta) = \mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)^\top\right] \tag{10}$$

is the Fisher Information Matrix (FIM).

Proof (Sketch). Let $\ell(\theta') := \log \pi_{\theta'}(a \mid s)$. Applying a second-order Taylor expansion of $\ell(\theta')$ around θ, we obtain:

$$\ell(\theta+\Delta\theta) = \ell(\theta) + \nabla_\theta \ell(\theta)^\top \Delta\theta + \tfrac{1}{2}\Delta\theta^\top \nabla_\theta^2 \ell(\theta)\,\Delta\theta + o(\|\Delta\theta\|^2). \tag{11}$$

Substituting into the KL definition:

$$D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta+\Delta\theta}) = \mathbb{E}\big[\ell(\theta) - \ell(\theta+\Delta\theta)\big]. \tag{12}$$

Taking expectations:

- The first-order term vanishes:

$$\mathbb{E}_{a\sim\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a \mid s)\big] = 0. \tag{13}$$

- The second-order term becomes:

$$-\tfrac{1}{2}\Delta\theta^\top \mathbb{E}\big[\nabla_\theta^2 \log \pi_\theta(a \mid s)\big]\Delta\theta. \tag{14}$$

Using the identity:

$$F(\theta) = -\mathbb{E}\big[\nabla_\theta^2 \log \pi_\theta(a \mid s)\big], \tag{15}$$

we obtain:

$$D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta+\Delta\theta}) = \tfrac{1}{2}\Delta\theta^\top F(\theta)\,\Delta\theta + o(\|\Delta\theta\|^2). \tag{16}$$

□

Remarks. KL divergence locally induces a quadratic form defined by the FIM, thereby providing a geometry-aware metric over the policy space. Consequently, KL-based trust-region methods implicitly constrain updates under the Fisher geometry for small policy perturbations.

A.2 Derivation of Score Function Penalties

Let $\tilde{r} = \frac{\pi_\theta(a|s)}{\pi_{\mathrm{old}}(a|s)}$ denote the probability ratio. Using the identity $\nabla_\theta \tilde{r} = \tilde{r}\nabla_\theta \log \pi_\theta$, a general gradient penalty of the form $\nabla_\theta L = \mathbb{E}_{\pi_{\mathrm{old}}}[\lambda(\tilde{r})\nabla_\theta \log \pi_\theta]$ corresponds to a pointwise penalty function $f(\tilde{r})$ satisfying:

$$\mathbb{E}_{\pi_{\mathrm{old}}}\left[f'(\tilde{r})\nabla_\theta \tilde{r}\right] = \mathbb{E}_{\pi_{\mathrm{old}}}\left[\underbrace{\tilde{r} f'(\tilde{r})}_{:=\lambda(\tilde{r})}\nabla_\theta \log \pi_\theta\right]. \tag{17}$$

This yields the general integration rule:

$$f(\tilde{r}) = \int \frac{\lambda(\tilde{r})}{\tilde{r}}\, d\tilde{r}. \tag{18}$$

Applying this rule, we derive three specific penalty objectives based on the choice of the score function multiplier $\lambda(\tilde{r})$:

Inverse KL divergence: Let $\lambda(\tilde{r}) = \tilde{r}\log\tilde{r}$. Integrating yields:

$$f(\tilde{r}) = \int \frac{\tilde{r}\log\tilde{r}}{\tilde{r}}\, d\tilde{r} = \int \log\tilde{r}\, d\tilde{r} = \tilde{r}\log\tilde{r} - \tilde{r} + 1. \tag{19}$$

Pearson χ^2-divergence: Let $\lambda(\tilde{r}) = \tilde{r}(\tilde{r}-1)$. Integrating yields:

$$f(\tilde{r}) = \int \frac{\tilde{r}(\tilde{r}-1)}{\tilde{r}}\, d\tilde{r} = \int (\tilde{r}-1)\, d\tilde{r} = \frac{1}{2}(\tilde{r}-1)^2. \tag{20}$$

Gaussian penalty: Let $\lambda(\tilde{r}) = \tilde{r}(\tilde{r}-1)\exp(-(\tilde{r}-1)^2)$. Integrating yields:

$$f(\tilde{r}) = \int (\tilde{r}-1)\exp\left(-(\tilde{r}-1)^2\right)\, d\tilde{r} = \frac{1}{2}\left[1 - \exp\left(-(\tilde{r}-1)^2\right)\right]. \tag{21}$$

B Hyperparameter Settings

B.1 LLM Post-training

To ensure consistency with prior work, we fix the global batch size to 1024 and the maximum response length to 8192 tokens. We use VERL [Sheng et al., 2025] as the infrastructure. The learning rate is set to 1×10^{-6}. We adopt GRPO [Guo et al., 2025]—a simplified PPO variant widely used in LLM post-training—as the optimization backbone, where group normalization replaces GAE for advantage estimation. For text generation, we use nucleus sampling with $\text{top}_p = 0.99$, $\text{top}_k = 100$, and temperature 0.6. Each task was trained for 500 steps. We set the update-to-data ratio to $UTD = 2$ (equivalent to PPO epochs), introducing mild off-policy effects due to data reuse. The policy is initialized from `Qwen3-1.7B-Base`. We exclusively use open-source datasets. We follow the training data from Yu et al. [2025], which is known to expose clear performance gaps across algorithms on long-tail reasoning tasks. We evaluate performance using pass@4 on four challenging benchmarks: MATH-500 [Lightman et al., 2023], OlympiadBench [He et al., 2024], MinervaMath [Lewkowycz et al., 2022], and AMC 2023 [Xue et al., 2025].

B.2 Robotic Control and Open-World Exploration

For tasks in these two domains, we employ SimBa and RNN architectures, SimBa [Lee et al., 2025] is a residual neural network tailored specifically for RL that incorporates a simplicity bias. Its efficacy has been well-validated in the original paper across environments such as DeepMind Control Suite [Tassa et al., 2018], MuJoCo [Todorov et al., 2012], and Craftax [Matthews et al., 2024]. For RNN baseline, we adopt Recurrent-PPO [Akkaya et al., 2019, Berner et al., 2019], a widely used PPO variant. To minimize the confounding effects of network plasticity [Lyle et al., 2024b, Nauman et al., 2024], we introduce LayerNorm into the official implementation [Matthews et al., 2024].

Furthermore, to explicitly evaluate the model's continual learning capabilities, we maintain a constant learning rate throughout training rather than applying learning rate annealing. Unlike the LLM setup, we employ advantage-weighted penalty, formulated as $\eta = \frac{|A|}{\epsilon}$. Empirically, this dynamic formulation yields superior performance on control tasks compared to utilizing a fixed penalty coefficient [Xie et al., 2024]. Aside from these specific modifications, we strictly adhere to the official hyperparameter configurations. All experiments were conducted on NVIDIA GeForce RTX 3090 GPU and completed within 12 hours.

Table 1: SimBaBlock structure

Step	Layer Configuration
1. Norm	LayerNorm(H)
2. FC + Act	Linear($H, 4H$), ReLU
3. FC	Linear($4H, H$)
4. Residual	Output = Step 3 Output + Block Input

Table 2: SimBa actor-critic network architectures

Layer	Actor Network	Critic Network
Fully Connected	Linear($d_{\text{obs}}, 128$)	Linear($d_{\text{obs}}, 512$)
SimBaBlock	$1\times$ SimBaBlock	$2\times$ SimBaBlock
LayerNorm	LayerNorm(128)	LayerNorm(512)
Output Head	Linear($128, d_{\text{act}}$)	Linear($512, 1$)
Policy Std	Learnable log std	–

Table 3: Layer-normalized GRU cell

Step	Layer Configuration
1. Gates	Linear($x, 2H$) + Linear($h, 2H$)
2. Gate Norm	LayerNorm($2H$), sigmoid
3. Candidate	Linear(x, H) + Linear($r \odot h, H$)
4. Candidate Norm	LayerNorm(H), tanh
5. Hidden Update	$h' = (1 - z) \odot \tilde{h} + z \odot h$

Table 4: LN-RNN actor-critic architecture for Craftax, $H = 512$ by default

Layer	Actor Network	Critic Network
Input Projection	Linear(d_{obs}, H), LayerNorm(H), ReLU	
Recurrent Core	LayerNormGRUCell(H)	
Fully Connected	Linear(H, H), LayerNorm, ReLU	Linear(H, H), LayerNorm, ReLU
Fully Connected	Linear(H, H), LayerNorm, ReLU	Linear(H, H), LayerNorm, ReLU
Output Head	Linear(H, d_{act})	Linear($H, 1$)
Policy	Categorical logits	–

Table 5: PPO hyperparameter settings across the three domains

Hyperparameter	DMC	MuJoCo-v5	Craftax
Learning Rate	3×10^{-4}	3×10^{-4}	2×10^{-4}
Optimizer	Adam	Adam	Adam
Optimizer ϵ	1×10^{-5}	1×10^{-5}	1×10^{-5}
Discount Factor (γ)	0.99	0.99	0.99
GAE Parameter (λ)	0.95	0.95	0.8
Number of Environments	1	1	1024
Rollout Length	2048	2048	64
Batch Size	2048	2048	65536
Minibatches	32	32	8
Minibatch Size	64	64	8192
Update Epochs	10	10	4
PPO Clip Coefficient	0.2	0.2	0.2
Value Loss Clipping	True	True	True
Value Loss Coefficient	0.5	0.5	0.5
Entropy Coefficient	0.0	0.0	0.01
Max Gradient Norm	0.5	0.5	1.0
Advantage Normalization	True	True	True
Learning Rate Annealing	False	False	False
Training Budget	1×10^6 per task	1×10^6 per task	1×10^9 total

C Additional Experiments

Additionally, we conducted experiments on MetaWorld+ [McLean et al., 2025] and Procgen [Cobbe et al., 2020]. For MetaWorld+, we utilized tanh-activated MLP on *door open→close* and *faucet open→close*, repeated for three cycles. For Procgen, we utilized the original IMPALA-CNN [Cobbe et al., 2020] on starpilot. As these environments are relatively straightforward, we include the results in this section for completeness and reference.

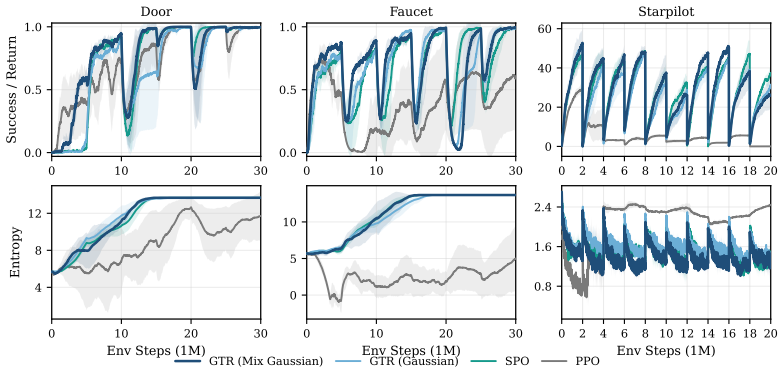

Figure 10: Aside from vanilla PPO, all variants demonstrate superior continual learnability.

D Broader Impacts

Our algorithm has several potential broader impacts, the most notable being the improvement of computational sustainability and the subsequent reduction in energy costs. We do not foresee any direct negative social impacts arising from this work.

E The Use of Large Language Models

In this paper, LLMs are only used to polish the writing of some paragraphs to improve clarity and grammar. The key ideas, theoretical analysis, method design, figures, and experimental results are all from the human authors' contributions, and LLMs do not constitute a core component of the proposed methodology.

F Connect with Student Authors

Bingxu Liu (Core Contributor) bxliu124@gmail.com
Jiashun Liu (Project Leader) ljshasdream@gmail.com
Johan Obando-Ceron jobando0730@gmail.com
Hao Wang hao.wang@my.cityu.edu.hk
Runze Liu 1036433215@qq.com

G Licenses

The licenses of used assets in this paper are listed as follows:

- GSM8K: MIT License
- MATH: MIT License
- AMC23: Apache-2.0 License
- Minerva Math: No License
- OlympiadBench: MIT License
- τ-bench: MIT License
- GPQA: MIT License
- ARC: CC BY-SA 4.0
- MMLU-Pro: Apache-2.0 License
- verl: Apache-2.0 License
- slime: Apache-2.0 License
- LlamaFactory: Apache-2.0 License
- AESL: Apache-2.0 License
- DFT: Apache-2.0 License
- CurioSFT: Apache-2.0 License

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See section 1 and section 5 for our contribution and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See our limitation at section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See section 4 and Appendix A for the theory and its proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix B for our experiment settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use data that's totally open-sourced, see Appendix B for details.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See hyperparameters in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See section 5 we use 3 seeds with standard deviations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix B for details

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that this research conforms to all its guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix D for details.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This research focuses on fundamental algorithmic improvements in reinforcement learning. The models and data used (e.g., standard benchmarks) do not pose a high risk of misuse as defined by the guidelines.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix G for details.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released the core implementation of our method, including the model architecture and training scripts, as part of the supplementary material. See abstract for details.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects, personal data, or any experiments requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for language editing and proofreading of the manuscript, which does not impact the core methodology, scientific rigor, or originality of the research as per the conference guidelines.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.