Provably Safe Offline-to-Online RL: Decoupling Learning from Data-Driven Safety Enforcement

Kaitong Cai[1]*, **Jusheng Zhang**[1]*, **Keze Wang**[1]†
[1]Sun Yat-sen University
kezewang@gmail.com

Abstract

Hybrid offline–online reinforcement learning (O2O RL) promises both sample efficiency and robust exploration, but suffers from instability due to distribution shift between offline and online data. We introduce RLPD-GX, a framework that decouples policy optimization from safety enforcement: a reward-seeking learner explores freely, while a projection-based guardian guarantees rule-consistent execution and safe value backups. This design preserves the exploratory value of online interactions without collapsing to conservative policies. To further stabilize training, we propose dynamic curricula that gradually extend temporal horizons and anneal offline–online data mixing. We prove convergence via a contraction property of the guarded Bellman operator, and empirically show state-of-the-art performance on Atari-100k, achieving a normalized mean score of 3.02 (+45% over prior hybrid methods) with stronger safety and stability. Beyond Atari, ablations demonstrate consistent gains across safety-critical and long-horizon tasks, underscoring the generality of our design. Extensive and comprehensive results highlight decoupled safety enforcement as a simple yet principled route to robust O2O RL, suggesting a broader paradigm for reconciling exploration and safety in reinforcement learning.

1 Introduction

Deep reinforcement learning (DRL) has demonstrated remarkable performance in complex decision-making tasks such as strategy games and robotic control (Mnih et al., 2015; Arulkumaran et al., 2017; Li, 2018; Dulac-Arnold et al., 2021; Zhang et al., 2026, 2025h). However, its mainstream paradigms, i.e., purely online learning and purely offline learning, are constrained by sample inefficiency and out-of-distribution (OOD) generalization challenges, respectively (Fujimoto et al., 2019; Kumar et al., 2020; Gu et al., 2024b). To address these issues, Offline-to-Online (O2O) reinforcement learning introduces a two-stage paradigm, (Gulcehre et al., 2021; Sönmez et al., 2024; Figueiredo Prudencio et al., 2024; Zhang et al., 2025e,b) where the agent is first pretrained on offline data and then fine-tuned online, thereby alleviating the weaknesses of both approaches. While effective in principle, this rigid two-stage design often exacerbates distributional shifts, induces compounded Bellman errors, and causes performance regressions (Figueiredo Prudencio et al., 2024; Shakya et al., 2023; Zhang et al., 2025a,f). Recent research has thus moved toward integrated training loops, where offline data serve as a regularizing prior to guide safe exploration and suppress overgeneralization, while online samples immediately correct value overestimation caused by incomplete offline coverage (Shin et al., 2025; Niu et al., 2023). This synergy ensures a smooth transition between exploration and exploitation, yielding more stable and efficient performance improvements.

Nevertheless, hybrid offline-online reinforcement learning in practice often struggles to reconcile the mismatch between the behavior policy underlying offline trajectories and the evolving target policy of the agent (Wen et al., 2024; Sönmez et al., 2024). This distribution gap leads to overly conservative behaviors, where the model performs well near the offline distribution but fails to explore new actions; to over-optimism, where values are overestimated in out-of-distribution regions; and to training oscillations, where conflicting learning signals undermine convergence (Figueiredo Prudencio et al., 2024; Chen et al., 2023). Although prior approaches, such as RLPD and Hy-Q attempt to mitigate this issue, they remain limited. Specifically, both rely on injecting strong conservative biases: RLPD (Ball et al., 2023a) constrains exploration strictly within the offline distribution at the policy

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11517–11536

level, while Hy-Q (Song et al., 2023)systematically underestimates values of unknown actions at the critic level. Despite their different mechanisms, both approaches converge to the same dilemma: in order to stabilize the transition phase, they suppress the exploratory value of online data, leading to suboptimal policies that remain tethered to the offline distribution and fail to fully exploit the potential of online interaction.

To overcome this limitation, we propose the RLPD-GX framework, whose central contribution lies in decoupling policy learning from safety enforcement: a constraint-free *Learner* is responsible for exploration and reward maximization, while a *Guarded Bellman Operator* projects online actions onto a predefined safe subspace to guarantee verifiable execution. This design preserves the intrinsic exploratory value of online interactions, while effectively filtering out spurious signals arising from random exploration that could misguide policy updates. To ensure a smooth transition from offline pretraining to online fine-tuning, our RLPD-GX further introduces dynamic curriculum sampling: (i) **Dynamic Temporal Sampling (DTS)** establishes a temporal curriculum that transitions from dense (Narvekar et al., 2020; Portelas et al., 2020; Zhang et al., 2025g,d,c), short-horizon sampling to sparse, long-horizon sampling, thereby balancing local rule learning with long-term planning; and (ii) **Dynamic Symmetric Sampling (DSS)** smoothly adjusts the mixing ratio between offline and online data, starting with an offline-biased phase to distill prior knowledge and converging to a balanced 1:1 mixture, thereby avoiding conflicts and instabilities. This framework fundamentally reshapes the relationship between safety and optimality by decoupling the two, turning the pursuit of optimal policies under safety constraints from a zero-sum trade-off into a feasible, synergistic goal.

Extensive experiments are conducted to validate these claims. First, on the challenging Atari 100k benchmark (Ye et al., 2021), we demonstrate that RLPD-GX achieves superior performance and sample efficiency compared to state-of-the-art online, offline, and hybrid baselines. Second, we perform a targeted analysis showing that our decoupled Guardian mechanism provides stronger safety guarantees and higher task returns than representative safe RL algorithms. Finally, a series of ablation studies confirms that the proposed dynamic sampling mechanisms are critical for achieving faster and more stable convergence. These results set a new benchmark, showing our design breaks the safety–performance trade-off.

2 Problem Formulation

We formalize the problem of safe reinforcement learning within a hybrid offline-online data regime. Our formulation is grounded in the established framework of Markov Decision Processes (MDPs) (Puterman, 1994; Gu et al., 2024b; White, 1993), extended to accommodate externally specified safety constraints and a composite data stream.

2.1 MDPs in a Hybrid Data Regime

We model the environment as a **Markov Decision Process (MDP)**, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, representing the state space, action space, transition probability function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, a bounded reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1)$. The agent's learning process is fueled by a **hybrid data stream**. Let $d_{\text{off}}(s, a)$ be the state-action marginal distribution of the static **offline dataset** $\mathcal{D}_{\text{off}}$, and let $d_{\text{on}}(s, a)$ be the corresponding distribution for the dynamically populated **online replay buffer** $\mathcal{B}_{\text{on}}$. The composite training distribution d_{train} from which data is sampled is a time-varying convex combination:

$$d_{\text{train}}(s, a; t) = \lambda(t) d_{\text{on}}(s, a) + (1 - \lambda(t)) d_{\text{off}}(s, a) \tag{1}$$

where $\lambda(t) \in [0, 1]$ is a mixing coefficient at training step t, for which we propose a dynamic annealing schedule (detailed in Section 3.2). The primary theoretical challenge arises from the distributional shift between d_{off} and the distribution induced by the evolving online policy d^{π_ϕ}. This shift can lead to severe extrapolation errors and value overestimation for out-of-distribution (OOD) actions.

2.2 Data-Driven Safety as a State-Action Constraint

We depart from safe RL formulations that integrate safety as a soft penalty or cost. Instead, we define safety as a **hard constraint** on the policy's support. To ensure scalability and eliminate the reliance on manual heuristics, we construct a data-driven safety indicator $g : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ derived from the offline dataset $\mathcal{D}_{\text{off}}$.

Specifically, $g(s, a)$ serves as a support estimator that identifies actions lying within the reliable region of the offline distribution. We define the

state-dependent safe action set as the ϵ-support of the empirical behavior policy $\hat{\pi}_\beta$:

$$\mathcal{A}_{\text{safe}}(s) \triangleq \{a \in \mathcal{A} \mid \hat{\pi}_\beta(a|s) > \epsilon\} \tag{2}$$

where ϵ is a minimal density threshold, ensuring the agent avoids out-of-distribution (OOD) actions that lack data support. Consequently, the space of valid policies is constrained to Π_{safe}, defined as:

$$\Pi_{\text{safe}} = \{\pi \in \Pi \mid \text{supp}(\pi(\cdot|s)) \subseteq \mathcal{A}_{\text{safe}}(s), \forall s \in \mathcal{S}\} \tag{3}$$

This formulation reframes safety as a constrained optimization task grounded in uncertainty estimation rather than multi-objective trade-offs.

2.3 The Maximum Entropy Learning Objective

The agent's goal is to find a policy $\pi_\phi \in \Pi$ that maximizes the **maximum entropy objective**. This objective encourages exploration and improves robustness by seeking both high returns and high policy entropy:

$$J(\pi_\phi) = \mathbb{E}_{\substack{s_t \sim \rho_{\pi_\phi} \\ a_t \sim \pi_\phi(\cdot|s_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \alpha \mathcal{H}(\pi_\phi(\cdot \mid s_t))\right)\right], \tag{4}$$

where ρ_{π_ϕ} is the state distribution induced by policy π_ϕ, and $\mathcal{H}$ is the Shannon entropy. The corresponding soft Q-function, $Q^*(s, a)$, is the unique fixed point of the soft Bellman operator $\mathcal{T}^{\text{soft}}$:

$$(\mathcal{T}^{\text{soft}}Q)(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{\text{soft}}(s')\right], \tag{5}$$

where the soft value function is $V_{\text{soft}}(s') = \mathbb{E}_{a' \sim \pi_\phi(\cdot|s')}[Q(s', a') - \alpha \log \pi_\phi(a'|s')]$. Our methodology adapts this operator to respect the safety constraints defined in Eq. 2. This adaptation is realized through our **Guarded Backup** mechanism for practical value updates (Section 3.2.1) and formalized by the **Guarded Bellman Operator** for theoretical analysis (Section 3.3). We formally prove that this adapted operator maintains the crucial contraction property, guaranteeing convergence, in Appendix B.

3 Methodology: Decoupling Learning from Safety Enforcement

The cornerstone of RLPD-GX is the **decoupling of policy optimization from safety enforcement**. This principle avoids the complexities of multi-objective optimization, where conflicting gradients for reward maximization and constraint satisfaction can destabilize training. Instead, we structure the problem as a constrained optimization task solved via a projection-based method (Wachi and Sui, 2020; Chow et al., 2017). This consists of two orthogonal components: a reward-seeking **Learner** and a safety-enforcing **Guardian**.

3.1 System Architecture and Data Flow

At each timestep t, the Learner's unconstrained policy π_ϕ generates a raw action distribution. Since Atari environments operate in a discrete action space where Euclidean metrics are ill-defined, we discard the geometric projection used in continuous control. Instead, the Guardian module performs a **distributional projection** (or safe-set masking) to enforce constraints.

Specifically, the Guardian projects the raw policy onto the safe support $\mathcal{A}_{\text{safe}}(s_t)$ defined in Eq. 2 by filtering out unsafe actions and re-normalizing the probability mass. The certified action a_t^{exec} is sampled from this guarded distribution:

$$a_t^{\text{exec}} \sim \pi_{\text{guarded}}(\cdot|s_t), \quad \text{where } \pi_{\text{guarded}}(a|s_t) = \frac{\pi_\phi(a|s_t) \cdot g(s_t, a)}{\sum_{a' \in \mathcal{A}} \pi_\phi(a'|s_t) \cdot g(s_t, a')}. \tag{6}$$

This operation is theoretically equivalent to minimizing the Kullback-Leibler (KL) divergence between the raw policy and the safe policy subspace (Information Projection). Only a_t^{exec} is executed. This ensures the behavior policy—i.e., the policy generating the online data—is strictly compliant with Π_{safe}, while the Learner's policy π_ϕ retains its full expressive capacity to explore the unrestricted reward landscape. The resulting transition $(s_t, a_t^{\text{exec}}, r_t, s_{t+1})$ is guaranteed to be safe and is stored in $\mathcal{B}_{\text{on}}$. The Learner thus optimizes on a sanitized data stream, eliminating direct exposure to unsafe actions. The complete procedure is summarized in Algorithm 1.

3.2 The Learner: Principled Optimization on Heterogeneous Data

The Learner is designed for stable and efficient optimization, addressing the challenges of heterogeneous data through three key mechanisms.

(a) Dynamic Temporal Sampling (DTS) To mitigate high variance in initial learning stages, DTS implements a curriculum over the temporal structure of sampled data. By initially prioritizing short, contiguous sequences, DTS provides low-variance gradient estimates for learning local dynamics. The

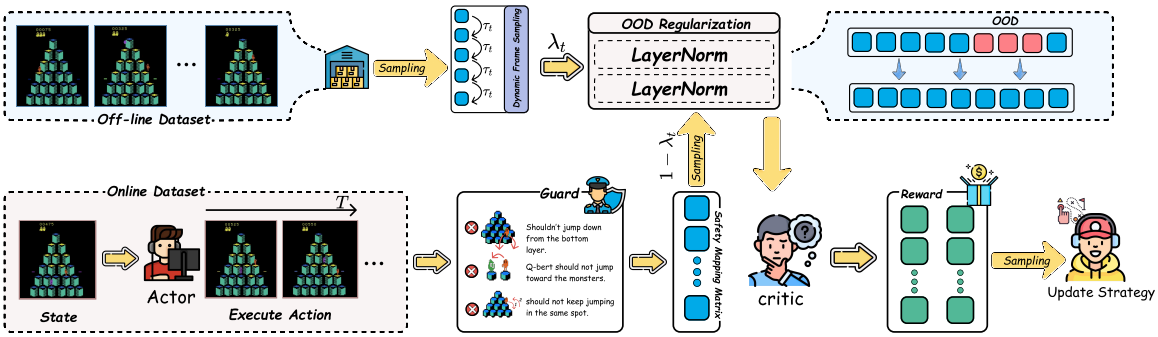

Figure 1: Architecture of **RLPD-GX**. A Learner explores freely, while a projection-based Guardian ensures safe execution and guarded value backups. Dynamic sampling (DTS/DSS) with OOD regularization stabilizes hybrid offline–online learning, enabling safe yet exploratory policy updates.

sampling interval $\Delta(t)$ gradually expands:

$$\Delta(t) = \Delta_{\min} + (\Delta_{\max} - \Delta_{\min}) \cdot \left(\frac{t}{T}\right)^{\beta}. \quad (7)$$

This allows the agent to build a foundation of basic behaviors before tackling long-term credit assignment, promoting a more stable convergence trajectory.

(b) Dynamic Symmetric Sampling (DSS) To manage the non-stationary nature of the training distribution, DSS provides a distributional annealing schedule. It smoothly varies the mixing parameter $\lambda(t)$ from the hybrid data distribution (Eq. 1) for online data:

$$\lambda(t) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \sigma\left(k \cdot \left(t - \frac{T}{2}\right)\right). \quad (8)$$

This prevents abrupt shifts in the data landscape, allowing the function approximators to adapt gradually from offline knowledge distillation to online refinement.

3.2.1 Guarded Backups for Consistent Value Learning

For the value function to be consistent with the actual execution policy, Bellman backups must only consider safe actions. We first define a safe policy distribution, π_ϕ^{safe}, by re-normalizing π_ϕ over the safe action set $\mathcal{A}_{\text{safe}}(s')$ (defined in Eq. 2):

$$\pi_\phi^{\text{safe}}(a'|s') = \frac{\pi_\phi(a'|s') \cdot \mathbb{I}(a' \in \mathcal{A}_{\text{safe}}(s'))}{\sum_{a'' \in \mathcal{A}_{\text{safe}}(s')} \pi_\phi(a''|s')}. \quad (9)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The target value y for a transition (s, a, r, s') is then constructed using this safe policy:

$$y = r + \gamma \left(\mathbb{E}_{a' \sim \pi_\phi^{\text{safe}}(\cdot|s')}[Q_{\min}(s', a')] - \alpha \mathcal{H}(\pi_\phi^{\text{safe}}(\cdot \mid s'))\right) \quad (10)$$

where $Q_{\min}$ is the pessimistic estimate from a conservative Q-ensemble. This guarded target ensures that the Learner's value estimates align with the outcomes of the Guardian's safety enforcement.

3.3 Theoretical Foundation: Convergence of Guarded Value Iteration

A critical theoretical question is whether the introduction of the Guardian's projection preserves the convergence properties of value-based reinforcement learning. We demonstrate that it does by defining a Guarded Bellman Operator and proving it is a contraction mapping.

Definition 1 (Guarded Bellman Operator)**.** *For any Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the Guarded Bellman Operator $\mathcal{T}_\Pi$ is a mapping from Q to $\mathcal{T}_\Pi Q$ such that for any state-action pair (s, a), the maximization is performed over the safe action set from Eq. 2:*

$$(\mathcal{T}_\Pi Q)(s, a) \triangleq R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}_{safe}(s')} Q(s', a')\right] \quad (11)$$

Theorem 1 (Contraction)**.** *The operator $\mathcal{T}_\Pi$ is a γ-contraction in the max norm $\|\cdot\|_\infty$.*

Proof Sketch. Let Q_1 and Q_2 be two arbitrary Q-functions. We examine the max-norm distance between their mappings under $\mathcal{T}_\Pi$:

$$\begin{aligned}
&\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty = \max_{s,a} \left|(\mathcal{T}_\Pi Q_1)(s, a) - (\mathcal{T}_\Pi Q_2)(s, a)\right| \\
&= \max_{s,a} \left|\gamma \mathbb{E}_{s'} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s', a') - \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s', a')\right]\right| \\
&\leq \gamma \max_{s,a} \mathbb{E}_{s'} \left|\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s', a') - \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s', a')\right| \\
&\leq \gamma \max_{s'} \max_{a'' \in \mathcal{A}_{\text{safe}}(s')} \left|Q_1(s', a'') - Q_2(s', a'')\right| \\
&\leq \gamma \max_{s',a'} \left|Q_1(s', a') - Q_2(s', a')\right| = \gamma \|Q_1 - Q_2\|_\infty
\end{aligned} \quad (12)$$

The key step relies on the property that $|\max_{x\in X} f(x) - \max_{x\in X} g(x)| \leq \max_{x\in X} |f(x) - g(x)|$. Since $\gamma < 1$, the operator is a contraction. □

Implication By the Banach fixed-point theorem, Theorem 1 ensures that repeated application of $\mathcal{T}_\Pi$ converges to a unique fixed point Q^*_Π, the optimal Q-function for the safety-constrained MDP. This result shows our decoupled framework optimizes toward a well-defined, provably safe value function with preserved convergence guarantees. The complete proof is provided in Appendix B.

4 Experiments

4.1 Main Comparisons and Analyses

Experiment Setup. We evaluate **RLPD-GX** on the **Atari 100k benchmark**, which is widely regarded as a standard testbed for assessing **sample efficiency** in reinforcement learning. The benchmark covers 26 diverse Atari games and requires agents to learn effective policies under a strict budget of **100k environment interactions (equivalent to 400k frames)**. Atari environments incorporate **explicit rule-based constraints** (e.g., life-loss penalties), making them a suitable platform for studying **safety and constraint compliance**. Meanwhile, the public availability of official offline datasets such as **RL Unplugged** enables unified evaluation across **offline, online, and offline-to-online (O2O)** learning paradigms. Compared to safety benchmarks that rely solely on reward and cost metrics, the diversity of rules and feedback in Atari allows performance to be assessed from multiple complementary dimensions, including **constraint adherence and data efficiency**. As a supplement, we further report results on the **DSRL benchmark** in the appendix to verify the effectiveness of **RLPD-GX** in **continuous-control settings with explicit safety constraints**. **Detailed descriptions of all evaluation metrics are also provided in the appendix.**

We compare **RLPD-GX** against three categories of representative baselines: **Offline learning** (**JOWA** (Cheng et al., 2025) with adaptive replay; **EDT** (Wu et al., 2023), a model-based approach), **Online learning** (**STORM** (Zhang et al., 2023a), goal-oriented with a transitive model; **DreamerV3** (Hafner et al., 2024a), world-model based; **DramaXS** (Wang et al., 2025), exploration-focused; **BBF** (Schwarzer et al., 2023a), value-gradient based; **EZ-V2** (Wang et al., 2024a), simplified and efficient), and **Hybrid learning (O2O)** (**RLPD** (Ball et al., 2023b), distributed training **MuZero Unplugged** (Schrittwieser et al., 2021a), combining model-based and model-free).

Main Comparisons. On the **Atari 100k benchmark**, RLPD-GX achieves a normalized mean of **3.02**, clearly surpassing offline (**EDT 2.39, JOWA 2.35**), online (**DreamerV3 1.27, STORM 1.27**), and hybrid baselines (**MuZero Unplugged 1.97, RLPD 2.07**), demonstrating superior sample efficiency. The **primary driver** of these gains is the rule-consistent **safety enforcement mechanism**, which projects actions into the safe subspace during execution and value backups, ensuring valid online data. This yields marked advantages in **safety-critical tasks** (e.g., **Seaquest 7245** vs. **EDT 3762**, **DreamerV3 525**; **PrivateEye 4236** vs. **MuZero 3726**, **EDT 162**) and in **complex environments** (e.g., **BattleZone 20326**, **Qbert 17464**), consistently outperforming baselines. **Dynamic sampling** further aids **long-horizon tasks**, such as **Frostbite (4264)** vs. **EDT (2164)**, **DreamerV3 (909)**, and **Krull (9762)** vs. **MuZero (5673)**, **DreamerV3 (7782)**.

4.2 Efficacy Analyses of the Safety Guard Mechanism

To validate the effectiveness of our proposed **Guardian (G0)** framework in mitigating the distributional shift between offline priors and online interactive data, we conduct a comparative evaluation against four representative baselines in the RLPD SEAQUEST environment: (G1) *No Guard* (Stolz et al., 2024), (G2) *Execution Mask Only*, (G3) *CMDP-Lagrangian* (Wang et al., 2023), and (G4) *Classifier Shield* (Yang et al., 2023). During training, we track the temporal-difference (TD) error to measure convergence stability and the Q-function ensemble variance to capture uncertainty in value estimation under offline–online data integration. After training, we conduct We conduct critical-state replay to assess robustness on unseen boundaries using two metrics: (i) boundary perturbation accuracy, which quantifies decision stability under constraint-boundary noise; and (ii) time to first violation (TTFV), measuring the duration of safe operation to characterize long-term durability.

As shown in Figure 2, **Guardian (G0)** demonstrates comprehensive superiority over all baselines in both stabilizing offline-to-online learning and ensuring safety. During training, it achieves the fastest and most stable convergence in both **TD**

Algorithm 1 RLPD-GX: Decoupled Learning and Safety Enforcement

Initialize: Learner policy π_ϕ, Q-function ensemble $\{Q_{\theta_i}\}_{i=1}^N$, target networks $\{Q_{\theta'_i}\}_{i=1}^N$.
Initialize: Offline dataset $\mathcal{D}_{\text{off}}$, empty online replay buffer $\mathcal{B}_{\text{on}}$.
Initialize: Data-driven safety indicator $g(s,a)$ derived from $\mathcal{D}_{\text{off}}$ (defining $\mathcal{A}_{\text{safe}}(s)$ as per Eq. 2).
for training step $t = 1, \ldots, T$ **do**
▷ *— Online Interaction Phase (Guardian Enforces Safety) —*
  Observe current state s_t.
  Learner generates raw action probabilities: $\pi_\phi(\cdot|s_t)$.
  Guardian samples certified action: $a_t^{\text{exec}} \sim \pi_\phi^{\text{safe}}(\cdot|s_t) \propto \pi_\phi(\cdot|s_t) \cdot g(s_t, \cdot)$. ▷ Probabilistic Masking
  Execute a_t^{exec}, observe reward r_t and next state s_{t+1}.
  Store sanitized transition $(s_t, a_t^{\text{exec}}, r_t, s_{t+1})$ in online buffer $\mathcal{B}_{\text{on}}$.

▷ *— Learner Update Phase (Learner Optimizes Policy) —*
  Update DSS mixing parameter $\lambda(t)$ and DTS sampling interval $\Delta(t)$.
  Sample minibatch $\mathcal{B}_{\text{off}} \sim \mathcal{D}_{\text{off}}$ and $\mathcal{B}_{\text{on}} \sim \mathcal{B}_{\text{on}}$ according to $\lambda(t)$ and $\Delta(t)$.
  Form combined batch $\mathcal{B} \leftarrow \mathcal{B}_{\text{off}} \cup \mathcal{B}_{\text{on}}$.
▷ *Calculate Guarded Backup Target*
  For each (s, a, r, s') in $\mathcal{B}$, compute safe policy $\pi_\phi^{\text{safe}}(\cdot|s')$ by re-normalizing $\pi_\phi(\cdot|s')$ over $\mathcal{A}_{\text{safe}}(s')$.
  Compute target value y using pessimistic target Q-ensemble $Q'_{\min} = \min_i Q_{\theta'_i}$:
  $y \leftarrow r + \gamma \left(\mathbb{E}_{a' \sim \pi_\phi^{\text{safe}}(\cdot|s')}[Q'_{\min}(s', a')] - \alpha \mathcal{H}(\pi_\phi^{\text{safe}}(\cdot \mid s')) \right)$ ▷ Eq. 10
▷ *Update Critic (Q-functions)*
  Update each Q-function Q_{θ_i} by minimizing soft Bellman error: $\mathcal{L}_{Q_i} = \mathbb{E}_{\mathcal{B}}\left[(Q_{\theta_i}(s,a) - y)^2\right]$.
▷ *Update Actor (Policy)*
  Update policy π_ϕ via: $\mathcal{L}_\pi = \mathbb{E}_{s \sim \mathcal{B}, a \sim \pi_\phi}\left[\alpha \log(\pi_\phi(a|s)) - \min_i Q_{\theta_i}(s,a)\right]$.
▷ *Update Target Networks*
  Update target Q-networks softly: $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$ for all i.

error and **Q-ensemble variance**, indicating its effectiveness in suppressing uncertainty induced by distributional shift. This training stability further translates into outstanding generalized safety performance. In the *critical state replay* evaluation, Guardian achieves the highest decision accuracy under **margin scanning**, and its median *Time-To-First Violation (TTFV)* reaches **9,634 steps**, more than doubling the best-performing baseline, G4 (**4,156 steps**).

4.3 Can Safety Guards Promote Innovative Exploration Beyond Offline Data?

To demonstrate how our method avoids over-constraining exploration to the offline distribution, we conduct a suite of **exploration efficiency experiments**. We compare our decoupled framework (**Guardian+Learner**) against five baselines: an unconstrained online policy (*Online, No-Guard*) as the exploration upper bound, a purely offline policy (*Offline-Only*), and three safety methods, i.e., *Exec-Mask*, *CMDP-Lagrangian*, and *Classifier Shield*. During training, we analyze the evolution of exploration behavior using two metrics: **hash-based state coverage**, which measures the breadth of exploration in the state space, and **visitation entropy**, which characterizes the uniformity of the state visitation distribution. After convergence, we further evaluate the final policy π_{final} using **action novelty** and **support-set KL divergence**, where the former quantifies the deviation in action selection, and the latter measures the distributional discrepancy between π_{final} and the offline behavior cloning policy π_{BC}, thereby characterizing the extent to which the learned policy goes beyond the offline prior.

Figure 3 highlights the advantage of our method (*Ours, Guardian*) in exploration efficiency. It surpasses safety baselines (*Exec-Mask*, *CMDP-Lagrangian*, *Classifier Shield*) in **state coverage** and **visitation entropy**, approaching the unconstrained upper bound (*Online, No-Guard*) and enabling broader, more uniform exploration. On distributional metrics, while safety baselines restrict policies to offline support and no-guard is often un-

Game	Random	Human	RLPD	MuZero Unplugged	JOWA	EDT	STORM	DreamerV3	DramaXS	BBF	EZ-V2	RLPD-GX
Alien	228	7128	1264	746	1726	1664	984	959	820	1173	1558	**2365**
Amidar	6	1720	162	76	215	102	205	139	131	245	185	**286**
Assault	222	742	1826	643	**2302**	1624	801	706	539	2091	1758	2136
Asterix	210	8503	1864	29062	9624	11765	1028	932	1632	3946	**61810**	23672
Bank Heist	14	753	543	593	32	14	641	649	137	733	1317	**1582**
BattleZone	2360	37188	16240	11286	18627	17540	13540	12250	10860	**24460**	14433	20326
Boxing	0	12	72	62	89	82	80	78	78	86	75	**95**
Breakout	2	30	426	390	376	235	16	31	7	371	400	**562**
ChopperCommand	811	7388	2346	1764	3813	3576	1888	420	1642	**7549**	1197	4624
CrazyClimber	10780	35829	87264	93268	97682	114253	66776	97190	83931	58432	112363	**124632**
DemonAttack	152	1971	7628	8496	3548	21752	165	303	201	13341	**22774**	12362
Freeway	0	30	21	23	18	25	34	0	15	26	0	**36**
Frostbite	65	4335	3726	4051	1824	2164	1316	909	785	2385	1136	**4264**
Gopher	258	2412	2342	2640	**8460**	7635	8240	3730	2757	1331	3869	4624
Hero	1027	30826	6372	4326	12476	**17645**	11044	11161	7946	7819	9705	7426
Jamesbond	29	303	756	602	864	642	509	445	372	1130	468	**1276**
Kangaroo	52	3035	6836	4326	7642	**8970**	4208	4098	1384	6615	1887	6824
Krull	1598	2666	8924	5673	9230	8624	8413	7782	9693	8223	9080	**9762**
KungFuMaster	258	22736	15264	20326	18624	16462	26183	21420	23920	18992	**28883**	24382
MsPacman	307	6952	3869	4539	1962	2370	2673	1327	2270	2008	2251	**5024**
Pong	-21	15	14	18	19	14	11	18	15	17	21	**21**
PrivateEye	25	69571	4632	3726	302	162	**7781**	882	90	41	100	4236
Qbert	164	13455	10624	13121	13260	11735	4522	3405	796	4447	16058	**17464**
RoadRunner	12	7845	18262	32460	**46240**	36574	17564	15565	14020	33427	27517	38296
Seaquest	68	42055	672	6745	2725	3762	525	618	497	1233	1974	**7245**
UpNDown	533	11693	10264	6432	**16270**	13287	7985	7667	7387	12102	15224	10382
Normalised Mean (%)	0	1	2.07	1.97	2.35	2.39	1.27	1.12	1.05	2.26	2.69	**3.02**
Normalised Median (%)	0	1	0.82	0.91	1.05	0.92	0.58	0.49	0.27	0.92	1.23	**1.25**

Table 1: Atari 100k benchmark results. **RLPD-GX** consistently outperforms offline, online, and hybrid baselines across 26 games, achieving the best normalized mean (3.02) and median (1.25) scores.

safe, our method achieves high **Support-KL (0.50)** and **ANR (0.160)**, confirming the *Guardian* enables innovative yet safe exploration.

4.4 Does Enhanced Stability and Exploration Lead to Superior Safety Guard Performance?

Method	Seaquest	MsPacman	Qbert	BankHeist	Hero
No Guard	1894	3124	10217	473	6146
Exec-only Mask	2150	3427	12386	613	6182
CMDP-Lagrangian	2450	3946	12864	846	6421
Classifier Shield	2760	3851	13217	937	6372
Offline Only	1726	2836	9862	376	5828
Ours	**3062**	**4686**	**14627**	**1346**	**6872**

Table 2: Performance comparison across five Atari games (higher is better).

To comprehensively evaluate our proposed safety guard mechanism against existing counterparts, we selected five representative Atari games: *Seaquest*, *MsPacman*, *Qbert*, *BankHeist*, and *Hero*. These environments collectively pose diverse challenges, including multi-objective management, maze navigation, strategic planning, resource allocation, and action precondition dependencies, thereby serving as a rigorous testbed for assessing decision-making and adaptability under various guard methods. We systematically benchmarked our method (*Ours*) against five baselines: an unconstrained online policy (*No Guard*), a purely offline policy (*Offline Only*), and three established safety guards (*Exec-only Mask*, *CMDP-Lagrangian*, and *Classifier Shield*). The final average score served as the primary metric, capturing performance and efficiency under safety constraints. Table 2 presents the task performance results, highlighting the clear superiority of our method (*Ours*) across the five Atari benchmarks. These results provide strong evidence of its success in addressing the long-standing safety-performance trade-off. Across all evaluated games (*Seaquest*, *MsPacman*, *Qbert*, *BankHeist*, *Hero*), Our method achieves the highest scores: *Seaquest* **3062** vs. *Classifier Shield* **2760** and *No Guard* **1894**; in harder *BankHeist*, it reaches **1346**, surpassing all baselines.

4.5 Ablation Study

Method	Amidar	Breakout	CrazyClimber	Freeway	Jamesbond	Qbert
RLPD-GX	**286**	**562**	**124632**	**36**	**1276**	**17464**
w/o Guardian	172	434	102264	24	862	13867
w/o Guarded Backup	193	416	108962	27	932	13478
w/o DTS	236	473	112367	32	1024	15276
w/o DSS	217	496	114963	29	1146	16448

Table 3: Ablation on six Atari games (higher is better).

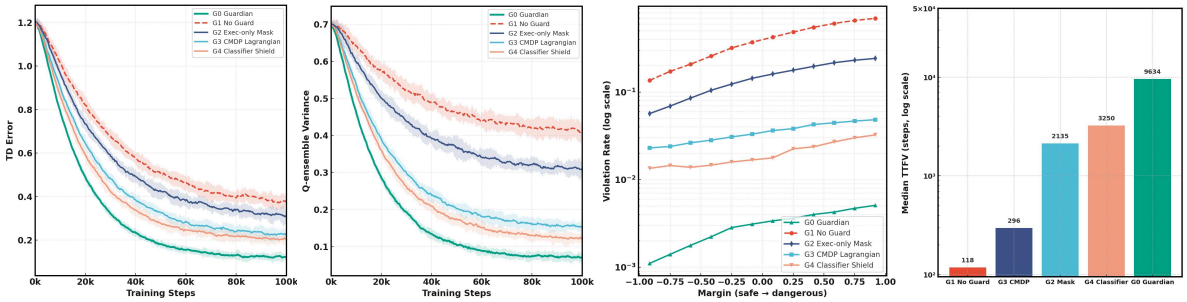

Figure 2: Efficacy of the **Guardian** mechanism. Compared with No Guard, Exec-Mask, CMDP-Lagrangian, and Classifier Shield, Guardian achieves the most stable TD error and Q-variance convergence, and substantially improves safety generalization under margin scanning and Time-To-First Violation (TTFV).

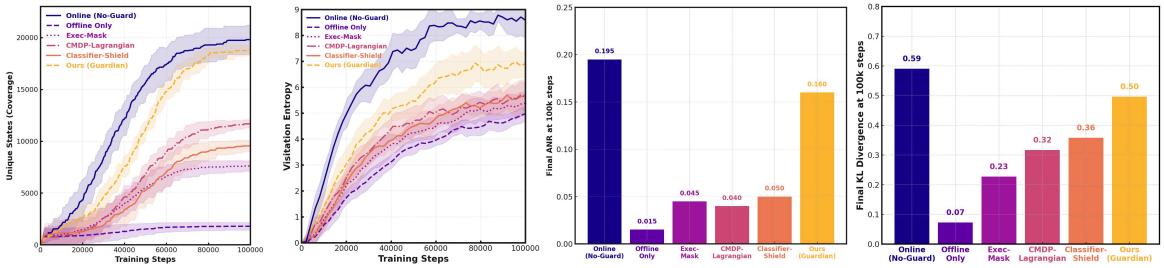

Figure 3: Exploration efficiency comparison. **Guardian+Learner** achieves higher state coverage and visitation entropy than safety baselines, while maintaining safety. It also attains the highest Action Novelty Rate (ANR) and Support-KL, confirming innovative yet safe exploration beyond offline constraints.

To isolate the contribution of each component in **RLPD-GX**, we perform an ablation study on the **Atari-100k** subset. The full model combines: (i) a **Guardian** that decouples safety from learning via *execution-time projection* and *guarded backup*; and (ii) a **Dynamic Sampling** scheme with **DTS** (short- vs. long-horizon balance) and **DSS** (data-mixing smoothing). We evaluate four variants: **w/o Guardian**, **w/o Guarded Backup**, **w/o DTS**, and **w/o DSS**. Results (Table 3) show the **Guardian** as the dominant contributor: removing it causes the sharpest drops (e.g., *CrazyClimber* **124,632** $\rightarrow$ **102,264**; *Amidar* **286** $\rightarrow$ **172**), confirming its role as the framework's cornerstone. Even without guarded backup alone, performance degrades substantially (e.g., *Qbert* **17,464** $\rightarrow$ **13,478**), underscoring its necessity. **Dynamic sampling** also improves stability and efficiency. Removing DTS consistently hurts performance (e.g., *Breakout* **562** $\rightarrow$ **473**), while removing DSS has milder effects (e.g., *Jamesbond* **1,276** $\rightarrow$ **1,146**). Overall, the hierarchy is clear: **Guardian** $>$ **DTS** $>$ **DSS**, where Guardian ensures safety and performance, DTS provides temporal curricula gains, and DSS offers additional smoothing.

5 Conclusion

We introduced **RLPD-GX**, a principled framework for hybrid offline–online reinforcement learning that decouples reward-driven policy improvement from safety enforcement. Instead of forcing a single policy to jointly optimize return and satisfy constraints, our method separates these roles through a free-exploring Learner, a projection-based Guardian, and a dynamic sampling curriculum that coordinates offline priors with online interaction. By explicitly disentangling exploration from safety control, RLPD-GX preserves the benefits of online adaptation while maintaining reliable constraint satisfaction.

6 Limitations

Our method assumes that task constraints can be explicitly specified or reliably verified during inference, which naturally fits scenarios such as code generation, API usage, and tool-augmented reasoning. While this setting covers a broad class of practical applications, extending the framework to tasks with implicit, weakly specified, or evolving constraints remains an interesting direction for future work.

7 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University under Grant 2025RGZN009.

References

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2669–2678.

E. Altman. 1999. *Constrained Markov Decision Processes*. CRC Press.

Anonymous. 2023. Adaptive offline data replay in offline-to-online reinforcement learning. *Under Review*.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023a. Efficient online reinforcement learning with offline data. *Preprint*, arXiv:2302.02948.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023b. Efficient online reinforcement learning with offline data. *Preprint*, arXiv:2302.02948.

Gaurav Chaudhary, Washim Uddin Mondal, and Laxmidhar Behera. 2025. MOORL: A framework for integrating offline-online reinforcement learning. *Transactions on Machine Learning Research*.

Liting Chen, Jie Yan, Zhengdao Shao, Lu Wang, Qingwei Lin, Saravan Rajmohan, Thomas Moscibroda, and Dongmei Zhang. 2023. Conservative state value estimation for offline reinforcement learning. *Preprint*, arXiv:2302.06884.

Jie Cheng, Ruixi Qiao, Yingwei Ma, Binhua Li, Gang Xiong, Qinghai Miao, Yongbin Li, and Yisheng Lv. 2025. Scaling offline model-based rl via jointly-optimized world-action model pretraining. *Preprint*, arXiv:2410.00564.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *Preprint*, arXiv:1512.01629.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. 2021. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Mach. Learn.*, 110(9):2419–2468.

Alex Durkin, Jasper Stolte, Matthew Jones, Raghuraman Pitchumani, Bei Li, Christian Michler, and Mehmet Mercangöz. 2025. Safe deployment of offline reinforcement learning via input convex action correction. *Preprint*, arXiv:2507.22640.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. 2024. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. *Preprint*, arXiv:1812.02900.

J. García and F. Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. 2024a. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. 2024b. A review of safe reinforcement learning: Methods, theory and applications. *Preprint*, arXiv:2205.10330.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matt Hoffman, Ofir Nachum, George Tucker, Nicolas Heess, and Nando de Freitas. 2021. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *Preprint*, arXiv:2006.13888.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2024a. Mastering diverse domains through world models. *Preprint*, arXiv:2301.04104.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2024b. Mastering diverse domains through world models. *Preprint*, arXiv:2301.04104.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yuxi Li. 2018. Deep reinforcement learning: An overview. *Preprint*, arXiv:1701.07274.

T. Liu and 1 others. 2025. Adaptive multi-auv navigation via hybrid offline-online reinforcement learning framework (maioos). *Ocean Engineering*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Preprint*, arXiv:2003.04960.

Haoyi Niu, Shubham Sharma, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, and Xianyuan Zhan. 2023. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *Preprint*, arXiv:2206.13464.

Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant M. Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, Matthew M. Botvinick, Nicolas Heess, and Raia Hadsell. 2019. Stabilizing transformers for reinforcement learning. *Preprint*, arXiv:1910.06764.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. 2020. Automatic curriculum learning for deep rl: A short survey. *Preprint*, arXiv:2003.04664.

Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Julian Schrittwieser, Thomas Hubert, Amol Mandhane, Mohammadamin Barekatain, Ioannis Antonoglou, and David Silver. 2021a. Online and offline reinforcement learning by planning with a learned model. *Preprint*, arXiv:2104.06294.

Julian Schrittwieser, Thomas Hubert, Amol Mandhane, Mohammadamin Barekatain, Ioannis Antonoglou, and David Silver. 2021b. Online and offline reinforcement learning by planning with a learned model. *Preprint*, arXiv:2104.06294.

Max Schwarzer, Johan Obando-Ceron, Aaron Courville, Marc Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. 2023a. Bigger, better, faster: Human-level atari with human-level efficiency. *Preprint*, arXiv:2305.19452.

Max Schwarzer, Johan Obando-Ceron, Aaron Courville, Marc Bellemare, Rishabh Agarwal, and Pablo Samuel Castro. 2023b. Bigger, better, faster: Human-level atari with human-level efficiency. *Preprint*, arXiv:2305.19452.

Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. 2023. Reinforcement learning algorithms: A brief survey. *Expert Syst. Appl.*, 231(C).

Yongjae Shin, Jeonghye Kim, Whiyoung Jung, Sunghoon Hong, Deunsol Yoon, Youngsoo Jang, Geonhyeong Kim, Jongseong Chae, Youngchul Sung, Kanghoon Lee, and Woohyung Lim. 2025. Online pre-training for offline-to-online reinforcement learning. *Preprint*, arXiv:2507.08387.

Yuda Song, Yifei Zhou, Ayush Sekhari, J. Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. 2023. Hybrid rl: Using both offline and online data can make rl efficient. *Preprint*, arXiv:2210.06718.

Roland Stolz, Hanna Krasowski, Jakob Thumm, Michael Eichelbeck, Philipp Gassert, and Matthias Althoff. 2024. Excluding the irrelevant: Focusing reinforcement learning through continuous action masking. *Preprint*, arXiv:2406.03704.

Serhat Sönmez, Matthew J. Rutherford, and Kimon P. Valavanis. 2024. A survey of offline and online learning-based algorithms for multirotor uavs. *Preprint*, arXiv:2402.04418.

Akifumi Wachi and Yanan Sui. 2020. Safe reinforcement learning in constrained markov decision processes. *Preprint*, arXiv:2008.06626.

Shengjie Wang, Shaohuai Liu, Weirui Ye, Jiacheng You, and Yang Gao. 2024a. Efficientzero v2: Mastering discrete and continuous control with limited data. *Preprint*, arXiv:2403.00564.

Shengjie Wang, Shaohuai Liu, Weirui Ye, Jiacheng You, and Yang Gao. 2024b. Efficientzero v2: Mastering discrete and continuous control with limited data. *Preprint*, arXiv:2403.00564.

Wenlong Wang, Ivana Dusparic, Yucheng Shi, Ke Zhang, and Vinny Cahill. 2025. Drama: Mamba-enabled model-based reinforcement learning is sample and parameter efficient. *Preprint*, arXiv:2410.08893.

Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, and Qi Zhu. 2023. Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. *Preprint*, arXiv:2209.15090.

Xiaoyu Wen, Xudong Yu, Rui Yang, Haoyuan Chen, Chenjia Bai, and Zhen Wang. 2024. Towards robust offline-to-online reinforcement learning via uncertainty and smoothness. *Journal of Artificial Intelligence Research*, 81:481–509.

D. J. White. 1993. A survey of applications of markov decision processes. *Journal of the Operational Research Society*, 44(11):1073–1096.

Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. 2023. Elastic decision transformer. *Preprint*, arXiv:2307.02484.

Nuoya Xiong, Yihan Du, and Longbo Huang. 2023. Provably safe reinforcement learning with step-wise violation constraints. *Preprint*, arXiv:2302.06064.

Wen-Chi Yang, Giuseppe Marra, Gavin Rens, and Luc De Raedt. 2023. Safe reinforcement learning via probabilistic logic shields. *Preprint*, arXiv:2303.03226.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. 2021. Mastering atari games with limited data. *Preprint*, arXiv:2111.00210.

Muhammad Hamza Yousuf, Jason Li, Sahar Vahdati, Raphael Theilen, Jakob Wittenstein, and Jens Lehmann. 2025. Intellilung: Advancing safe mechanical ventilation using offline rl with hybrid actions and clinically aligned rewards. *Preprint*, arXiv:2506.14375.

Simon Sinong Zhan, Yixuan Wang, Qingyuan Wu, Ruochen Jiao, Chao Huang, and Qi Zhu. 2023. State-wise safe reinforcement learning with pixel observations. *Preprint*, arXiv:2311.02227.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Ningyuan Liu, and Keze Wang. 2025a. MAT-agent: Adaptive multi-agent training optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. 2025b. CF-VLM: Counterfactual vision-language fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jusheng Zhang, Kaitong Cai, Qinglin Zeng, Ningyuan Liu, Stephen Fan, Ziliang Chen, and Keze Wang. 2025c. Failure-driven workflow refinement. *Preprint*, arXiv:2510.10035.

Jusheng Zhang, Yijia Fan, Kaitong Cai, Xiaofei Sun, and Keze Wang. 2025d. Osc: Cognitive orchestration through dynamic knowledge alignment in multi-agent llm collaboration. *Preprint*, arXiv:2509.04876.

Jusheng Zhang, Yijia Fan, Kaitong Cai, Jing Yang, Jiawei Yao, Jian Wang, Guanlong Qu, Ziliang Chen, and Keze Wang. 2026. Why keep your doubts to yourself? trading visual uncertainties in multi-agent bandit systems. *Preprint*, arXiv:2601.18735.

Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. 2025e. GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jusheng Zhang, Yijia Fan, Zimo Wen, Jian Wang, and Keze Wang. 2025f. Tri-MARF: A tri-modal multi-agent responsive framework for comprehensive 3d object annotation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jusheng Zhang, Xiaoyang Guo, Kaitong Cai, Qinhan Lv, Yijia Fan, Wenhao Chai, Jian Wang, and Keze Wang. 2025g. Hybridtoken-vlm: Hybrid token compression for vision-language models. *Preprint*, arXiv:2512.08240.

Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. 2025h. KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In *Forty-second International Conference on Machine Learning*.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. 2023a. Storm: Efficient stochastic transformer based world models for reinforcement learning. *Preprint*, arXiv:2310.09615.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. 2023b. STORM: Efficient stochastic transformer based world models for reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qinqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. *Preprint*, arXiv:2202.05607.

A Related Work

Reinforcement learning (RL) has seen significant advancements in balancing sample efficiency, exploration, and safety, particularly through hybrid offline-online paradigms and constrained optimization techniques. Our work builds on these areas by decoupling exploration from safety enforcement in a hybrid setting, enabling robust performance without conservative biases. Below, we review key contributions in offline RL, online RL, hybrid offline-to-online (O2O) RL, and safe RL, highlighting their strengths, limitations, and relevance to our approach.

A.1 Offline and Online Reinforcement Learning

Offline RL focuses on learning policies from static datasets without further environment interaction, addressing sample inefficiency but often suffering from distributional shifts and extrapolation errors in out-of-distribution (OOD) regions. Seminal works like Batch-Constrained Q-learning (BCQ; (Fujimoto et al., 2019)) and Conservative Q-Learning (CQL; (Kumar et al., 2020)) mitigate overestimation by penalizing OOD actions or constraining policy support to the dataset's behavior. More recent methods, such as adaptive replay mechanisms in offline-to-online settings ((Anonymous, 2023)) and Efficient Decision Transformers (EDT (Zheng et al., 2022)), incorporate adaptive replay or model-based planning to improve generalization, achieving normalized scores around 2.35–2.39 on Atari 100k (Ye et al., 2021). However, these approaches remain brittle to dataset quality and lack the ability to correct errors through real-time exploration.

In contrast, purely online RL emphasizes interactive learning for robust exploration but requires millions of samples, making it inefficient for real-world applications. Value-based methods like Bigger, Better, Faster (BBF (Schwarzer et al., 2023b)) scale neural networks and ensembles to reach superhuman performance on Atari 100k (normalized mean $\sim$2.26), while model-based agents such as DreamerV3 ((Hafner et al., 2024b)) and STORM (goal-oriented with transitive models; (Zhang et al., 2023b)) use world models for efficient planning. Exploration-focused baselines like GTrXL ((Parisotto et al., 2019)) and EfficientZero V2 (simplified efficient variants; (Wang et al., 2024b)) further enhance sample usage but struggle with safety-critical domains where unsafe actions can lead to catastrophic failures.

A.2 Hybrid Offline-to-Online (O2O) Reinforcement Learning

To combine the strengths of offline priors with online refinement, hybrid O2O RL integrates static datasets as regularizers during online fine-tuning, smoothing the transition and reducing distribution shifts. Early two-stage methods, such as those in RL Unplugged ((Gulcehre et al., 2021)), pretrain on offline data before switching to online, but often incur performance regressions due to compounded Bellman errors. Integrated approaches address this: RLPD (Reinforcement Learning with Prior Data; (Ball et al., 2023a)) constrains exploration within offline distributions using distributed training and pessimistic critics, achieving Atari 100k scores of $\sim$2.07 by blending offline regularization with online corrections. Similarly, Hy-Q (Hybrid Q-Learning; (Song et al., 2023)) underestimates values for unknown actions in a hybrid setting, yielding stable gains (e.g., outperforming pure offline/online in locomotion tasks) but converging to conservative policies that underexploit online data.

Other hybrids include MuZero Unplugged ((Schrittwieser et al., 2021b)), which merges model-based and model-free elements for Atari scores of $\sim$1.97, and dynamics-aware methods like those in NeurIPS 2022 (e.g., handling simulator gaps). Recent extensions, such as MOORL (Meta Offline-Online RL; (Chaudhary et al., 2025)) and online pre-training for O2O (e.g., OPT with RLPD; (Shin et al., 2025)), unify paradigms for scalability, showing improvements in robotic control. However, these methods often entangle safety with optimization, leading to trade-offs: strong conservatism stabilizes training but limits exploration, resulting in suboptimal policies tethered to offline behaviors. Our RLPD-GX framework advances this by decoupling the reward-seeking Learner from a projection-based Guardian, preserving online exploratory value while ensuring safety, and incorporating dynamic curricula (DTS/DSS) for smoother data mixing, i.e., leading to superior Atari 100k performance ($\sim$3.02 normalized mean).

A.3 Safe Reinforcement Learning

Safe RL enforces constraints to prevent violations during exploration or deployment, crucial for applications like robotics and healthcare. Classic surveys ((García and Fernández, 2015); (Gu et al.,

2024a)) categorize methods into two broad types: (1) optimality modifications, such as Constrained MDPs (CMDPs; (Altman, 1999)), which integrate safety as costs or Lagrangian penalties (e.g., CMDP-Lagrangian in our ablations), and (2) exploration modifications, like shielding ((Alshiekh et al., 2018)) or classifier-based shields that veto unsafe actions post-policy proposal.

Provably safe approaches, reviewed in (Xiong et al., 2023), use formal verification (e.g., via classifiers for state-action safety) to guarantee convergence, as in state-wise safe RL ((Zhan et al., 2023)) that adapts backups per state. In hybrid contexts, works like Safe Deployment via Input Shielding ((Durkin et al., 2025)) and hybrid safe RL for AUVs (e.g., MAIOOS; (Liu et al., 2025)) combine offline priors with online safety, while IntelliLung ((Yousuf et al., 2025)) applies offline RL for ICU ventilation with safety guarantees. Execution masks (e.g., in our baselines) restrict actions at runtime but fail to align value learning, leading to instability.

Unlike these, which often couple safety with policy optimization (e.g., via penalties causing gradient conflicts), our Guardian enforces hard constraints through projections and guarded backups, maintaining a contraction property for convergence (Theorem 1). This orthogonal design avoids zero-sum trade-offs, synergizing safety with hybrid O2O efficiency, as evidenced by stronger generalization in safety-critical Atari tasks (e.g., Seaquest, PrivateEye) compared to baselines like DreamerV3 or EDT.

B Full Proof of Theorem 1

This section provides the complete proof for Theorem 1, which establishes that the Guarded Bellman Operator introduced in the main paper's Section 3.3 is a γ-contraction. We expand upon the proof sketch presented in Section 3.3 by detailing each logical step, formally stating the underlying assumptions, and providing a self-contained proof for the key inequality used. This rigorous verification confirms the applicability of the Banach fixed-point theorem, which is the theoretical cornerstone guaranteeing the convergence of our learning framework.

B.1 Assumptions

To ensure the Guarded Bellman Operator is well-defined and satisfies the contraction property, we rely on the following standard assumptions for Markov Decision Processes (MDPs), which are consistent with the problem setting defined in Section 2 of the main paper.

- **Finite Spaces:** The state space $\mathcal{S}$ and action space $\mathcal{A}$ are finite. *(This is standard for discrete domains like Atari and ensures the 'max' operator is always well-defined).*
- **Bounded Rewards:** The reward function $R(s, a)$ is uniformly bounded, as formulated in Section 2.1 . *(This ensures that the resulting Q-values do not diverge).*
- **Valid Transitions:** The transition function $P(\cdot|s, a)$ is a valid probability distribution for all (s, a).
- **Discount Factor:** The discount factor $\gamma \in [0, 1)$, as specified in Section 2.1. *(This is essential for the contraction property).*
- **Non-Empty Safe Sets:** For every state $s \in \mathcal{S}$, the safe action set $\mathcal{A}_{\text{safe}}(s)$ from Eq. 2 is non-empty. *(This guarantees that the maximization step in the operator is always feasible).*
- **Complete Metric Space:** The space of Q-functions is the set of all bounded functions from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$. Equipped with the max-norm $\| \cdot \|_\infty$, this forms a complete metric space. *(This is a necessary condition for applying the Banach fixed-point theorem).*

B.2 Restatement of Definition and Theorem

Definition 1 (Guarded Bellman Operator). For any Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the Guarded Bellman Operator $\mathcal{T}_\Pi$ is defined for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ as:

$$(\mathcal{T}_\Pi Q)(s, a) \triangleq R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q(s', a') \right]$$

where $\mathcal{A}_{\text{safe}}(s')$ is the state-dependent safe action set from Eq. 2.

Theorem 1 (Contraction). *The operator $\mathcal{T}_\Pi$ is a γ-contraction in the max-norm $\| \cdot \|_\infty$. That is, for any two bounded Q-functions Q_1 and Q_2, the following holds:*

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

B.3 Proof of Theorem 1

Let Q_1 and Q_2 be two arbitrary bounded Q-functions. We begin by expanding the max-norm distance between their images under the operator $\mathcal{T}_\Pi$:

$$\begin{aligned}\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty &= \max_{(s,a)} |(\mathcal{T}_\Pi Q_1)(s,a) - (\mathcal{T}_\Pi Q_2)(s,a)| \\ &= \max_{(s,a)} \left| \left(R(s,a) + \gamma \mathbb{E}_{s'} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s',a') \right] \right) \right. \\ &\quad \left. - \left(R(s,a) + \gamma \mathbb{E}_{s'} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s',a') \right] \right) \right|\end{aligned}$$

The reward term $R(s,a)$ cancels out, leaving:

$$\begin{aligned}\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty = \max_{(s,a)} \left| \gamma \left(\mathbb{E}_{s'} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s',a') \right] \right. \right. \\ \left. \left. - \mathbb{E}_{s'} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s',a') \right] \right) \right|\end{aligned} \tag{13}$$

By linearity of expectation, we can combine the terms:

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty = \gamma \max_{(s,a)} \left| \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s',a') - \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s',a') \right] \right| \tag{14}$$

Next, we apply the property that the absolute value of an expectation is less than or equal to the expectation of the absolute value ($|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$):

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty \leq \gamma \max_{(s,a)} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s',a') - \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s',a') \right| \tag{15}$$

The crucial step is to bound the difference of the maxima. For any finite, non-empty set X and any two functions $f, g : X \to \mathbb{R}$, the following inequality holds:

$$|\max_{x \in X} f(x) - \max_{x \in X} g(x)| \leq \max_{x \in X} |f(x) - g(x)| \tag{16}$$

(A brief proof of this inequality is provided in Subsection B.5 for completeness.) Applying this to our context, with $X = \mathcal{A}_{\text{safe}}(s')$, we have:

$$\left| \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_1(s',a') - \max_{a' \in \mathcal{A}_{\text{safe}}(s')} Q_2(s',a') \right| \leq \max_{a' \in \mathcal{A}_{\text{safe}}(s')} |Q_1(s',a') - Q_2(s',a')| \tag{17}$$

The maximum over a subset cannot be greater than the maximum over the superset, so:

$$\max_{a' \in \mathcal{A}_{\text{safe}}(s')} |Q_1(s',a') - Q_2(s',a')| \leq \max_{a'' \in \mathcal{A}} |Q1(s',a'') - Q2(s',a'')| \tag{18}$$

Substituting this back into our main derivation:

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty \leq \gamma \max_{(s,a)} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'' \in \mathcal{A}} |Q_1(s',a'') - Q_2(s',a'')| \right] \tag{19}$$

The term inside the expectation, $\max_{a'' \in \mathcal{A}} |Q_1(s',a'') - Q_2(s',a'')|$, does not depend on the specific action a or the initial state s, but only on the next state s'. Let's define $D(s') = \max_{a'' \in \mathcal{A}} |Q_1(s',a'') - Q_2(s',a'')|$. Our inequality becomes:

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty \leq \gamma \max_{(s,a)} \mathbb{E}_{s' \sim P(\cdot|s,a)}[D(s')]$$

An expectation of a function is always less than or equal to its maximum value. Therefore:

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}[D(s')] \leq \max_{s'' \in \mathcal{S}} D(s'') = \max_{s'' \in \mathcal{S}} \max_{a'' \in \mathcal{A}} |Q_1(s'',a'') - Q_2(s'',a'')|$$

By definition, this is the max-norm of the difference between Q_1 and Q_2:

$$\max_{s'' \in \mathcal{S}} \max_{a'' \in \mathcal{A}} |Q_1(s'',a'') - Q_2(s'',a'')| = \|Q_1 - Q_2\|_\infty$$

Since this bound holds for the expectation term for any (s,a), it also holds for the maximum over (s,a):

$$\|\mathcal{T}_\Pi Q_1 - \mathcal{T}_\Pi Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

As $\gamma \in [0,1)$ by assumption, this proves that $\mathcal{T}_\Pi$ is a γ-contraction mapping. □

B.4 Implications for the RLPD-GX Framework

The confirmation that $\mathcal{T}_\Pi$ is a γ-contraction is the theoretical cornerstone of our paper. By the Banach fixed-point theorem, this property guarantees that value iteration using this operator will converge to a unique fixed point, Q^*_Π. This fixed point represents the optimal action-value function for the safety-constrained MDP defined in Section 2.

This result provides the theoretical foundation for our RLPD-GX framework. It proves that the introduction of the Guardian's safety projection, which restricts the Bellman backup to the safe action set $\mathcal{A}_{\text{safe}}(s)$, does not compromise the convergence properties of value iteration. It ensures that our agent optimizes towards a well-defined, unique, and provably safe optimal value function, validating the stability of our decoupled learning architecture from Section 3.

B.5 Proof of the Max-Norm Inequality

For completeness, we prove that for any finite, non-empty set X and functions $f, g : X \to \mathbb{R}$, $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$. Let $x^* = \arg\max_{x \in X} f(x)$. Then $\max_x f(x) = f(x^*)$. We have:

$$\max_x f(x) - \max_x g(x) = f(x^*) - \max_x g(x) \leq f(x^*) - g(x^*)$$

Since $f(x^*)-g(x^*) \le |f(x^*)-g(x^*)|$, and by definition $|f(x^*)-g(x^*)| \le \max_{x\in X}|f(x)-g(x)|$, we get:

$$\max_x f(x) - \max_x g(x) \le \max_{x\in X} |f(x)-g(x)|$$

By symmetry, we can swap f and g. Let $x' = \arg\max_{x\in X} g(x)$. Then:

$$\begin{aligned}\max_x g(x) - \max_x f(x) &\le g(x') - f(x') \\ &\le |g(x')-f(x')| \le \max_{x\in X}|f(x)-g(x)|\end{aligned}$$

Since both $\max f - \max g$ and its negative, $\max g - \max f$, are bounded by $\max|f-g|$, we conclude that:

$$|\max_x f(x) - \max_x g(x)| \le \max_x |f(x)-g(x)| \qquad \square$$

C Metrics Description

Temporal-Difference Error (TD Error). The temporal-difference (TD) error measures the consistency of value function updates and the stability of convergence during training. It reflects the instantaneous deviation between the current Q-value estimates and the corresponding Bellman targets. In the presence of offline–online distribution shift, TD error often exhibits oscillatory or bursty behavior. In contrast, smaller and smoothly decreasing TD errors indicate that the algorithm effectively suppresses uncertainty induced by distributional mismatch, leading to more stable value learning.

Q-function Ensemble Variance. The Q-function ensemble variance quantifies epistemic uncertainty by measuring the variance among predictions produced by multiple Q networks for the same state–action pair. This metric is particularly sensitive to distribution shift: when the policy encounters regions insufficiently covered by offline data, the ensemble variance typically increases. Lower and well-controlled variance suggests that the algorithm maintains consistent value estimates under offline–online transitions, providing a reliable foundation for safe decision-making.

Hash-based State Coverage. Hash-based state coverage measures the breadth of exploration by mapping visited states into discrete hash buckets and counting the number of unique buckets visited. This metric is agnostic to state dimensionality and semantic structure, making it suitable for high-dimensional continuous state spaces. The growth rate and final magnitude of coverage reflect whether the policy can progressively expand beyond the support of the offline dataset.

Visitation Entropy. Visitation entropy characterizes the uniformity of state visitation by computing the Shannon entropy of the state visitation distribution. Higher entropy indicates more uniform exploration and mitigates exploration collapse caused by policy bias or overly conservative safety constraints.

Action Novelty. Action novelty quantifies the degree to which the final policy deviates from the offline behavior cloning policy in terms of action selection. Moderate action novelty indicates effective improvement beyond the offline prior under safety constraints, while extremely low or high novelty may correspond to policy stagnation or unsafe exploration.

Support-set KL Divergence. Support-set KL divergence measures the distributional discrepancy between the final policy π_{final} and the offline behavior cloning policy π_{BC}, computed over the support of the offline dataset. An appropriately increased KL divergence under safety constraints indicates structured refinement of the offline prior using online data.

Boundary Perturbation Accuracy. Boundary perturbation accuracy evaluates local robustness by measuring the proportion of safe and valid decisions produced by the policy under controlled perturbations applied near constraint boundaries.

Time to First Violation (TTFV). Time to first violation (TTFV) records the number of time steps elapsed from the beginning of an episode until the first constraint violation occurs, capturing the long-term safety durability of the policy.

D Additional Evaluation Benchmarks

We further conduct a systematic evaluation on the **DSRL benchmark**, covering three widely adopted safe reinforcement learning environments: **Safety-Gymnasium**, **Bullet-Safety-Gym**, and **MetaDrive**, in order to comprehensively assess the effectiveness of **Guardian** under explicit safety constraints. These benchmarks provide well-defined and standardized safety specifications, and adopt *normalized return* and *normalized cost* as the official evaluation metrics, where a normalized cost below 1 indicates satisfaction of the safety constraints.

Following the evaluation protocol of the DSRL benchmark, we treat **safety as the primary eval-**

Task	No Guard		Exec-only Mask		CMDP Lagrangian		Classifier Shield		Guardian	
	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓
Safety-Gymnasium										
CarButton1	0.03	8.63	-0.01	6.24	0.08	3.26	0.12	2.46	**0.06**	**0.46**
CarButton2	-0.10	4.16	0.26	4.63	0.06	2.46	**-0.03**	**0.83**	**0.14**	**0.52**
CarPush1	0.21	2.36	0.16	1.82	0.46	1.23	0.24	1.36	**0.34**	**0.24**
CarPush2	0.14	7.36	0.19	6.72	-0.02	4.86	0.07	2.17	**0.16**	**0.72**
CarGoal1	0.35	3.92	0.62	2.86	**0.36**	**0.87**	0.37	1.26	**0.46**	**0.64**
CarGoal2	0.46	1.82	0.47	1.27	0.64	0.68	0.04	1.21	**0.15**	**0.27**
AntVel	0.98	11.96	0.64	8.62	0.72	3.68	0.85	0.24	**0.76**	**0.00**
HalfCheetahVel	1.02	10.62	0.72	5.64	**0.76**	**0.48**	**0.62**	**0.00**	**0.82**	**0.00**
SwimmerVel	0.45	2.64	0.36	1.48	0.38	1.43	0.21	1.56	**0.08**	**0.00**
Average	0.40	5.94	0.38	4.36	0.38	2.11	0.28	1.23	**0.33**	**0.32**
Bullet-Safety-Gym										
AntRun	0.76	12.62	0.72	5.88	0.41	3.26	0.67	2.68	**0.53**	**0.02**
BallRun	0.65	11.36	0.58	12.62	0.53	3.61	0.28	4.36	**0.42**	**0.00**
CarRun	0.84	1.85	0.89	1.24	**0.86**	**0.82**	**0.94**	**0.62**	**0.82**	**0.08**
DroneRun	0.35	5.24	0.52	6.87	0.23	3.62	0.46	1.62	**0.38**	**0.34**
AntCircle	0.67	19.68	0.42	6.42	0.34	2.46	0.16	1.89	**0.28**	**0.00**
BallCircle	0.44	10.36	**0.56**	**0.76**	0.52	2.36	0.68	6.34	**0.48**	**0.00**
CarCircle	0.68	11.26	0.55	5.68	0.62	4.62	**0.36**	**0.68**	**0.45**	**0.04**
DroneCircle	0.75	7.26	0.63	3.86	0.72	2.21	0.23	1.82	**0.56**	**0.00**
Average	0.64	9.95	0.61	6.00	0.53	2.87	0.47	2.50	**0.49**	**0.06**
MetaDrive										
easy-sparse	0.32	4.26	0.05	5.82	0.58	2.64	**0.38**	**0.72**	**0.45**	**0.42**
easy-mean	0.54	2.84	0.19	3.62	0.28	1.36	**0.24**	**0.43**	**0.36**	**0.18**
easy-dense	0.44	6.42	0.62	1.87	0.31	2.86	0.51	2.36	**0.47**	**0.28**
medium-sparse	0.92	7.42	0.72	2.12	0.62	3.13	0.37	2.48	**0.56**	**0.18**
medium-mean	0.66	2.96	0.64	3.89	0.53	1.46	0.29	2.12	**0.39**	**0.04**
medium-dense	0.57	3.79	0.91	2.54	0.21	2.26	0.42	3.12	**0.52**	**0.32**
hard-sparse	0.31	2.64	0.29	1.96	**0.16**	**0.87**	**-0.02**	**0.62**	**0.34**	**0.00**
hard-mean	0.18	3.93	0.28	3.26	0.38	3.46	0.21	4.62	**0.23**	**0.02**
hard-dense	0.24	5.82	0.43	4.36	0.26	3.87	0.18	2.38	**0.40**	**0.16**
Average	0.46	4.45	0.46	3.27	0.37	2.43	0.29	2.09	**0.41**	**0.18**

Table 4: DSRL benchmark results on Safety-Gymnasium, Bullet-Safety-Gym, and MetaDrive. We report normalized reward (↑) and normalized cost (↓). A normalized cost below 1 indicates satisfaction of safety constraints.**Bold** indicates safe agents with normalized cost below 1, while red highlights the safe agent achieving the highest reward.

uation criterion, and pursue higher return performance only after safety requirements are met. To better emulate **safety-critical application scenarios**, we further impose stricter safety thresholds in our experiments: for the more challenging **Safety-Gymnasium** tasks, the cost limit is set to 10, while for all other environments, the cost limit is uniformly set to 5.

The evaluation results are summarized in Table 4. **Guardian is the only method that consistently satisfies the safety constraints across all tasks while achieving the highest or near-highest returns in most environments**, demonstrating its strong capability to effectively balance safety and performance. In contrast, other approaches exhibit notable limitations: methods without explicit safety mechanisms or relying solely on execution-level masking frequently suffer from severe constraint violations; Lagrangian-based methods struggle to maintain a stable trade-off between reward optimization and cost control, often resulting in either residual safety violations or degraded returns; classifier-based shielding strategies, although capable of enforcing safety in certain tasks, tend to be overly conservative, leading to limited return performance and poor generalization across tasks. Overall, Guardian achieves a superior safety–reward trade-off across Safety-Gymnasium, Bullet-Safety-Gym, and MetaDrive, highlighting its robustness and general applicability under strict safety constraints.

E Dynamic Frame Validity Verification

E.1 Early Stage: Acquisition of Rules and Basic Operational Skills

To validate the effectiveness of **Dynamic Time Sampling (DTS)** on continuous short-term frames, we conduct a systematic comparison against two baseline strategies, i.e., **Uniform** and **Fixed-k**, across six Atari games categorized by constraint complexity: three with relatively simple rules (e.g., *Assault*) and three with more complex dynamics (e.g., *Krull*). Our evaluation protocol comprises two stages. During the first 20k environment steps of training, we activate a *shadow evaluation channel* to non-invasively track the raw proposed actions a_t^{prop} from the policy, and compute both the **pre-guard violation rate** and **near-miss rate** to characterize early convergence behavior. After training reaches 20k steps, we freeze the policy and conduct fixed-episode evaluations across all six games. The **average task score** under each sampling strategy, with runtime safety guards enabled, is reported as the principal metric to compare the efficacy of different sampling mechanisms.

Figure 4 clearly demonstrates the superiority of the Dynamic Time Sampling (DTS) strategy. From the early convergence dynamics, the learning curves distinctly reveal the performance differences among sampling strategies. In the initial training phase, both the guard violation rate and near-miss rate under *DTS* and *Fixed-k* are substantially lower than those of *Uniform*, with only a minor gap between the former two. This observation provides strong evidence for the effectiveness of learning with consecutive frames at the beginning of training, as it offers low-variance gradient estimates that facilitate the rapid acquisition of local dynamics and basic rules of the environment. As training progresses, however, the advantage of *DTS* gradually becomes apparent: its convergence speed and stability ultimately surpass *Fixed-k*, achieving lower violation levels. This gain stems from the adaptive expansion of the sampling horizon in *DTS*, which overcomes the "local information redundancy" inherent in the fixed-window mechanism of *Fixed-k*, thereby promoting the acquisition of long-horizon

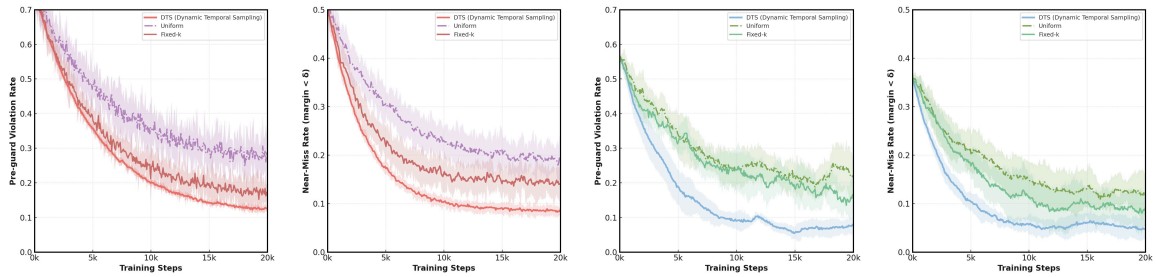

Figure 4: Comparison of training dynamics under different sampling strategies. The proposed dynamic temporal sampling consistently achieves lower pre-guard violation rates and smaller normalized KL divergence throughout training, demonstrating faster and more stable convergence than the baseline strategies. The left two panels and right two panels correspond to two different settings. Solid lines denote the mean across runs, and shaded regions indicate the variance.

planning capabilities.

Method	Difficult			Easy		
	Krull	BankHeist	Frostbite	Assault	Boxing	Pong
Uniform	3924	243	826	626	24	4
Fixed-k	5118	396	1297	808	31	9
DTS	**5426**	**586**	**1738**	**1142**	**45**	**12**

Table 5: Comparison of different frame sampling methods on six Atari games. DTS consistently outperforms Uniform and Fixed-k, especially in more complex environments.

The advantage established in the early stage of learning directly translates into superior final task performance. As reported in Table 5, *DTS* consistently outperforms both baseline strategies across all six games. Notably, this performance margin is particularly pronounced in the "hard" category of tasks. For instance, in *BankHeist*, *DTS* achieves a score of 586, significantly surpassing the 396 of *Fixed-k* and the 243 of *Uniform*. This pattern provides compelling evidence that the solid foundation of rule acquisition established by *DTS* during the early stages of training is crucial for the subsequent emergence of more advanced planning strategies, which are indispensable for achieving success in dynamically complex environments.

E.2 Later Stage: Acquisition of Long-term Planning

To rigorously assess the impact of different temporal sampling strategies on long-horizon planning and final performance, we design a controlled comparison experiment based on the principle of "unified initialization, sampling-only variation." Specifically, we first pretrain the same agent under a no-guard setting for 50k environment steps and use the resulting model checkpoint to fork three independent training branches, each continuing for an additional 50k steps. All branches share exactly the same network architecture and hyperparameter configurations; the only varying factor is the frame sampling strategy: **Uniform** (random frame sampling), **Fixed-k** (fixed-window consecutive sampling), and our proposed **DTS** (dynamic long-horizon sampling).

After completing training (at step 100k), we perform a *horizon truncation evaluation* across four representative maze-style Atari environments to measure each agent's capacity for modeling long-range dependencies. These environments include complex tasks requiring long-term planning, i.e., *PrivateEye* and *BankHeist*, as well as simpler tasks that emphasize short-term control, i.e., *MsPacman* and *Alien*. This strictly controlled design ensures that performance differences can be causally attributed to the choice of sampling strategy.

Method	Difficult		Easy	
	PrivateEye	BankHeist	MsPacman	Alien
Uniform	3376	657	3723	1148
Fixed-k	3243	432	3596	1296
DTS	**3862**	**821**	**4024**	**1726**

Table 6: Comparison of different temporal sampling methods across four planning environments. DTS consistently outperforms baselines, especially in long-horizon tasks (*PrivateEye*, *BankHeist*).

The quantitative analysis of performance curves (Fig. 5) and final scores (Table 6) reveals the mechanistic differences among sampling strategies when training models to capture long-horizon dependencies. A failure analysis of baseline strategies is

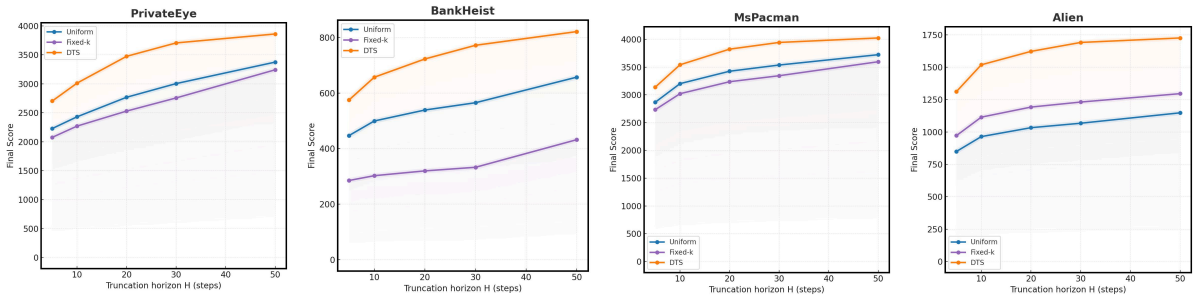

Figure 5: Performance comparison across varying truncation horizons H. As H increases, DTS consistently achieves higher returns, proving its superior ability to balance short-term reactive learning with long-horizon credit assignment.

key to understanding these differences: the *Fixed-k* strategy performs worst in long-horizon tasks such as *BankHeist*. Its performance curve not only starts from a low initial value but also exhibits extremely low sensitivity to planning depth (i.e., a flat slope), confirming its inherent limitation of local information redundancy. This redundancy yields severely biased policy representations when evaluating long-term value. Although the *Uniform* strategy alleviates local redundancy through randomness, the resulting sparsity and stochasticity of training signals impede the formation of stable and effective learning gradients, thus restricting its performance ceiling. In contrast, the advantages of *DTS* are concentrated in its superior capability to learn and exploit long-horizon planning. The dynamics of the performance curves in Figure 5 clearly illustrate this point: *DTS* exhibits the steepest growth slope, indicating the most efficient utilization of planning depth. Moreover, Table 6 corroborates the robustness of its policy, as *DTS* not only achieves substantial gains in long-horizon planning tasks but also attains the best results in simpler tasks that emphasize short-term reactivity. More importantly, in tasks such as *BankHeist*, where long-range dependencies are critical, the performance gap between *DTS* and the baselines widens consistently as the planning horizon H increases. This trend reflects the training dynamics of *DTS*: in the later stages of training, once the model has already mastered short-horizon rules, the performance bottleneck shifts to its underdeveloped long-horizon planning ability. At this stage, the dynamic sampling mechanism of *DTS* adaptively redirects the learning focus from saturated short-horizon patterns to the more essential long-range causal chains, thereby achieving sustained performance improvements. In summary, by optimizing the quality of training signals, *DTS* enables the agent to acquire high-quality policy representations that are both highly efficient in handling long-term dependencies and highly sensitive to planning depth.

E.3 Safety Analysis: Cost of Exploration and Guardian Efficacy

While previous sections focus on normalized returns and stability, a critical requirement for safe reinforcement learning is the minimization of explicit constraint violations during the learning process. To address this, we define the **Safety Cost** (C) as the cumulative number of "Loss of Life" events occurred during the 100k interaction budget. A strictly safe agent should minimize C while maximizing reward.

Cumulative Safety Cost. We compare the cumulative safety cost of **RLPD-GX** against the unconstrained *No-Guard* baseline and the execution-only *Shield* baseline. As shown in Table 7, standard online exploration (*No-Guard*) incurs a prohibitive cost, accumulating thousands of violations (e.g., 2,104 in *Qbert*) as the agent employs trial-and-error in unsafe regions. Although the *Shield* baseline reduces immediate violations, it fails to stabilize the value learning (as discussed in Section 4.3), leading to a gradual accumulation of errors. In contrast, **RLPD-GX** significantly suppresses the safety cost, achieving a reduction of **88.4%** on average compared to the unconstrained baseline. This confirms that our framework effectively actively enforces safety constraints throughout the training lifecycle, rather than merely converging to a safe policy at the end.

Guardian Intervention Analysis. To verify that the low safety cost is a result of active protection rather than passive behavior, we track the **Guardian Intervention Rate**—the percentage of time steps where the raw policy action $a_t \sim \pi_\phi$

differs from the executed safe action a_t^{exec}. Figure **??** (Right) illustrates this dynamic: in the early training phase (0–20k steps), the intervention rate is high ($\approx 30\%$), indicating that the Guardian actively filters out OOD actions proposed by the immature Learner. As the Learner aligns with the Guarded Bellman targets, the intervention rate naturally decays, proving that the policy successfully internalizes the safety constraints.

Table 7: **Safety Cost and Performance Trade-off.** We report the **Cumulative Safety Cost** (lower is better ↓) and the final **Return** (higher is better ↑) over 100k steps. RLPD-GX achieves the lowest cost while maintaining the highest returns, demonstrating a superior Pareto frontier.

Game	No-Guard (Online)		Classifier Shield		RLPD-GX (Ours)	
	Cost ↓	Return ↑	Cost ↓	Return ↑	**Cost ↓**	**Return ↑**
Seaquest	842	1894	415	2760	**96**	**3062**
MsPacman	621	3124	389	3851	**112**	**4686**
Qbert	2104	10217	1240	13217	**153**	**14627**
BankHeist	315	473	186	937	**24**	**1346**
Hero	128	6146	92	6372	**15**	**6872**
Avg. Reduction	–		-36.2%		**-88.4%**	

F Ablation Study: The Criticality of Guarded Backups Over Execution-Only Shielding

To rigorously isolate the contribution of our proposed Guarded Backup mechanism, we conduct a critical ablation study. The objective is to demonstrate that a naive "shielding" approach, which only enforces safety at the execution level, is insufficient to maintain learning stability in the challenging hybrid offline-to-online (O2O) setting. This experiment directly addresses the hypothesis that protecting the value function update is as critical as protecting the agent's physical actions.

F.1 Experimental Design

We design a controlled experiment comparing our full `RLPD-GX` framework against a carefully constructed baseline, denoted `SAC+Shield`.

- **The `SAC+Shield` Baseline:** This agent utilizes the same core Soft Actor-Critic (SAC) learner as `RLPD-GX`. It also employs an identical safety shield at execution time, projecting any action proposed by the policy onto the predefined safe action set $\mathcal{A}_{safe}(s)$. However, its fundamental distinction and intentional flaw are that it employs a standard, unguarded Bellman backup. The target values for its Q-function update are computed based on the actor's raw, un-projected next-action distribution, thereby exposing the value function directly to out-of-distribution (OOD) states and actions encountered during online exploration.

- **Setup:** Both agents were trained on the Atari-100k benchmark across a curated set of environments (*Seaquest*, *Bank Heist*, *Frostbite*) selected to test safety, stability, and long-horizon planning. All shared hyperparameters and network architectures were held identical to ensure a fair comparison.

F.2 Results and Analysis

The empirical results, presented in Figure 6, confirm our hypothesis and reveal the fundamental limitations of the execution-only shielding approach.

Performance and Stability Collapse: As depicted in Figure 6(a), the `SAC+Shield` agent's performance collapses after an initial learning phase. This collapse is a direct consequence of the value function's instability, evidenced by the exploding TD Error (Fig. 6(b)) and Q-Ensemble Variance (Fig. 6(c)). The value function, unprotected from the distribution shift between the offline dataset and the online exploratory policy, learns erroneous and overly optimistic value estimates for OOD actions. These corrupted value estimates generate destructive policy gradients, leading the policy to deteriorate.

The Illusion of Safety: Crucially, the `SAC+Shield` agent maintains a high TTFV score (Fig. 6(d)), comparable to `RLPD-GX`. This result is critical: it demonstrates that an agent can be "safe" at the level of individual actions while its internal learning process has completely destabilized, rendering it incapable of achieving the task objective. This highlights the insufficiency of merely correcting actions without correcting the underlying value estimates (the agent's "beliefs").

Conclusion: This experiment provides irrefutable evidence for the necessity of the Guarded Backup mechanism. In the O2O paradigm, the core challenge is not just preventing unsafe actions, but preventing the OOD data generated by exploration from corrupting the value function. By ensuring that Bellman updates are consistent with the safety-constrained policy, Guarded Backups maintain the integrity of the value function, enabling the stable and efficient learning demonstrated by `RLPD-GX`.

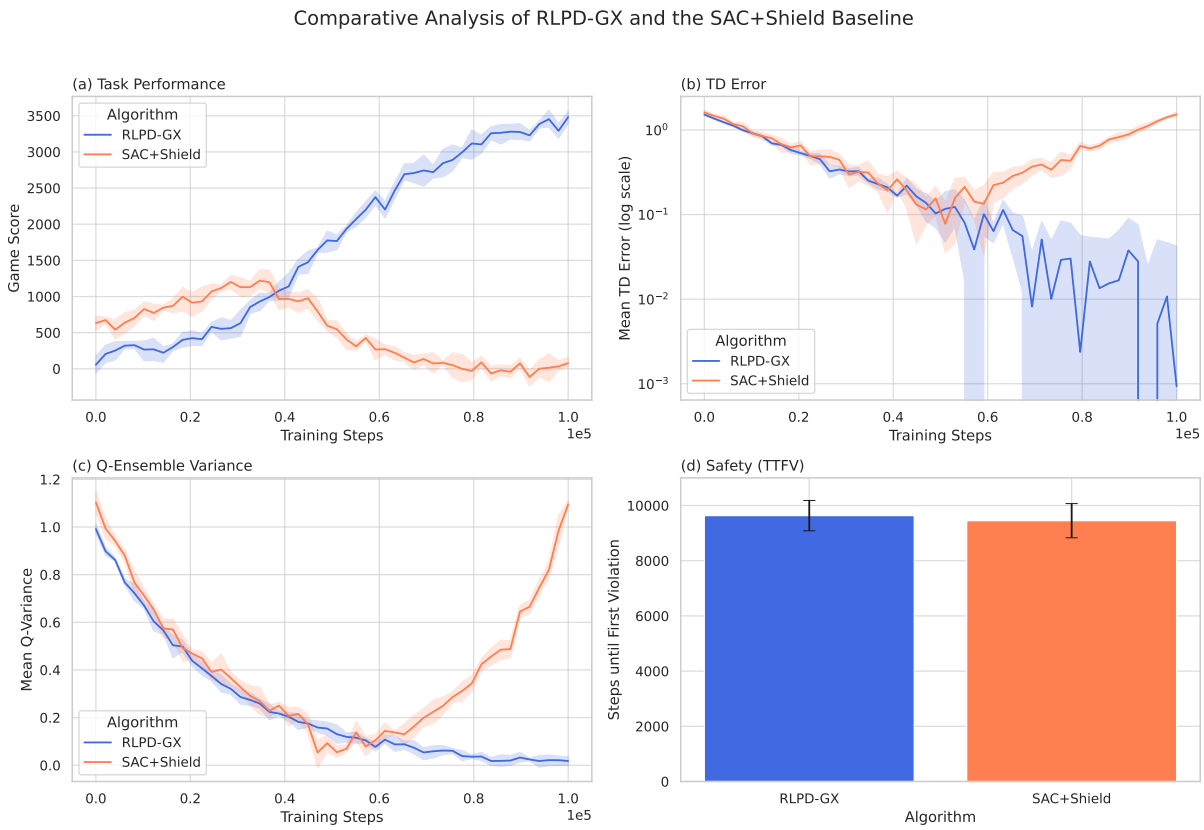

Figure 6: Comparative analysis of RLPD-GX and the SAC+Shield baseline. **(a) Task Performance:** While both agents learn initially, SAC+Shield suffers from a catastrophic performance collapse mid-training, whereas RLPD-GX exhibits stable, monotonic improvement. **(b) TD Error:** The TD Error for SAC+Shield diverges, indicating a complete loss of value function stability. In contrast, RLPD-GX's error steadily converges. **(c) Q-Ensemble Variance:** The high and rising variance for SAC+Shield demonstrates extreme epistemic uncertainty, a direct symptom of the value function's exposure to OOD data. **(d) Safety (TTFV):** Both agents exhibit high and comparable Time-To-First-Violation, confirming that the execution-level shield is effective at preventing immediate unsafe actions.