PRISM: PEER-REGULARIZED INDUCTION VIA STATE MINING FOR REWARD-FREE BACKDOOR ATTACKS IN REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks can manipulate reinforcement learning (RL) policies under hidden triggers while preserving benign behavior. Existing approaches mainly exploit victim-side information or manipulated training signals, overlooking the behavioral knowledge encoded in independently trained benign policies. We propose PRISM, a multi-policy-assisted backdoor framework that uses clean peer policies to guide selective poisoning. PRISM combines victim-side state salience with peer-based target-action compatibility to identify informative, low-conflict states, and performs reward-preserving target induction with peer regularization. Across visual control, safety control, and autonomous driving tasks, PRISM achieves strong attack success, preserves benign performance, and accelerates early-stage backdoor implantation. Our findings reveal a new attack surface in multi-policy RL: benign policy populations can unintentionally provide useful guidance for backdoor implantation.The anonymous implementation and experimental code are available at:`https://anonymous.4open.science/r/ICLRproject-13C3`

1 INTRODUCTION

Reinforcement learning (RL) has achieved strong performance in sequential decision-making tasks, including robotics, autonomous systems, and complex control Sutton & Barto (2018); Li (2017). However, RL policies remain vulnerable to backdoor attacks, where an agent behaves normally on benign observations but executes attacker-specified actions when a hidden trigger appears Liu et al. (2018). Such attacks are particularly challenging in RL because policy learning depends on sequential interactions, reward feedback, and evolving state distributions.

Recent work has progressively explored more realistic RL backdoor threat models. BadRL studies sparse state-dependent poisoning Cui et al. (2024), SleeperNets develops universal backdoor poisoning Rathbun et al. (2024), and adversarial inception attacks reduce dependence on arbitrary reward manipulation Rathbun et al. (2025). More recent studies further consider component-level and post-training backdoors Vyas et al. (2026), while defenses such as BIRD highlight the continuing difficulty of detecting and removing backdoors from RL policies Chen et al. (2023). Despite these advances, existing attacks mainly derive poisoning opportunities from victim-side information, manipulated trajectories, rewards, or other attacker-controlled signals. The behavioral knowledge encoded in independently trained benign policies remains largely unexplored.

This gap becomes particularly relevant in multi-policy systems. Recent studies show that multi-agent RL (MARL) can be manipulated through online reward/action attacks Liu & Lai (2023), offline reward poisoning Wu et al. (2023), and poisoned data that alter Markov-game equilibria Wu et al. (2024). Existing multi-agent backdoors similarly assume malicious or compromised participants; for example, BACKDOORL uses an adversarial agent to activate hidden behavior in a victim Wang et al. (2021). In contrast, we ask a different question: *can completely benign and independently trained policies themselves provide useful guidance for backdoor implantation?*

To investigate this problem, we propose **PRISM** (**P**eer-**R**egularized **I**nduction via **S**tate **M**ining), a multi-policy-assisted backdoor framework that exploits clean peer policies without compromising them. PRISM combines victim-side state salience with peer-based target-action compatibility to

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

identify informative, low-conflict poisoning states. It then performs reward-preserving target induction while using clean-peer regularization to constrain benign behavior. Experiments across visual control, safety control, and autonomous driving show strong attack effectiveness, competitive benign performance, and efficient early-stage implantation.

Our contributions are summarized as follows:

- We identify a new attack surface in multi-policy RL, where independently trained benign policies can unintentionally guide backdoor implantation.
- We propose PRISM, which combines victim-side salience and clean-peer behavioral compatibility for selective state mining, and performs reward-preserving induction with peer regularization.
- We evaluate PRISM across heterogeneous RL environments and theoretically and empirically characterize targeted induction, benign-performance preservation, and fixed-rollout PPO target preservation.

2 Related Work

2.1 Backdoor Attacks in Reinforcement Learning

RL backdoors implant trigger-activated malicious behavior while preserving benign policy functionality Liu et al. (2018). Early work such as TrojDRL demonstrated the feasibility of training-time backdoors in deep RL Kiourti et al. (2020). Subsequent studies consider increasingly constrained and stealthy attack settings: BadRL performs sparse targeted poisoning Cui et al. (2024), SleeperNets studies universal backdoor poisoning across RL domains Rathbun et al. (2024), adversarial inception attacks operate under restricted reward manipulation Rathbun et al. (2025), and recent work extends attacks to component-level and post-training settings Vyas et al. (2026). Defense methods such as BIRD further demonstrate the difficulty of detecting and removing RL backdoors Chen et al. (2023). These methods predominantly rely on victim-side signals, poisoned experiences, or attacker-controlled mechanisms, whereas PRISM exploits behavioral information from independently trained clean policies.

2.2 Adversarial Attacks in Multi-Agent Reinforcement Learning

MARL introduces additional attack surfaces because manipulating rewards, actions, data, or individual agents can affect joint behavior. Existing work studies offline reward poisoning Wu et al. (2023), online reward and action manipulation Liu & Lai (2023), and data poisoning that changes the apparent equilibrium of Markov games Wu et al. (2024). Backdoor-specific methods typically assume malicious or compromised agents; BACKDOORL, for example, activates victim backdoors through an adversarial participant Wang et al. (2021). PRISM differs in that all peer policies remain independently trained and benign, and their behavioral distributions are used only as guidance for selective poisoning.

2.3 Policy Interaction and Peer Learning

Behavioral knowledge transfer across policies is widely used to improve RL learning and generalization. Policy distillation transfers knowledge from teacher policies Rusu et al. (2016), while population-based training exploits policy populations for more effective optimization Jaderberg et al. (2017). PRISM instead studies the security implication of such behavioral knowledge: agreement among benign policies can reveal states where a target action is compatible with normal behavior.

Related risks have also emerged in broader autonomous-agent systems. AgentPoison exploits poisoned memory or knowledge bases in LLM-based agents Chen et al. (2024), while recent work demonstrates backdoor threats against LLM agents more generally Yang et al. (2024). Although these settings differ from RL, they similarly show that auxiliary behavioral or knowledge sources may introduce previously overlooked security risks.

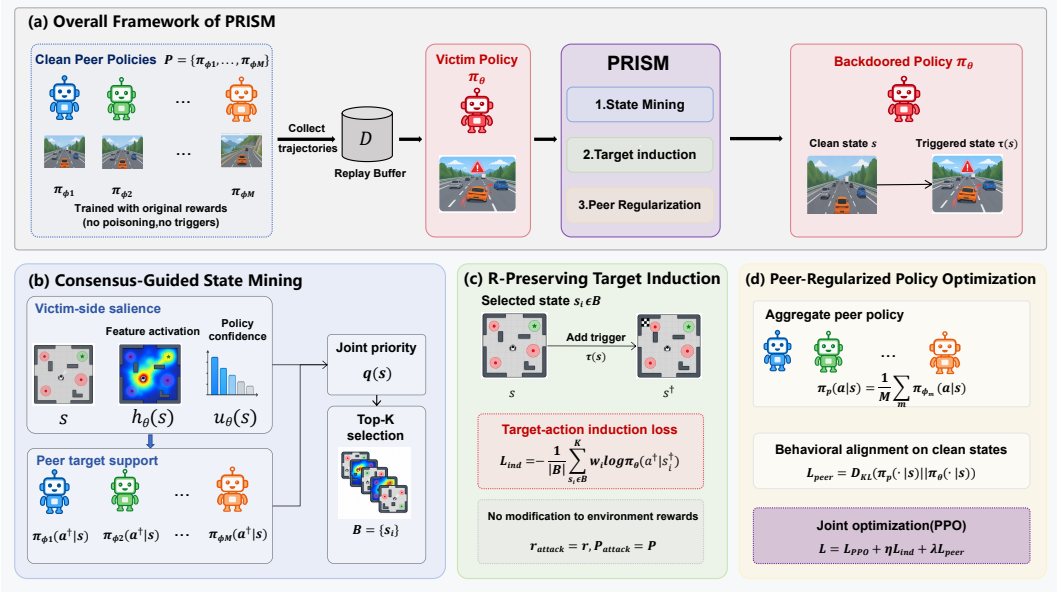


Figure 1: Overview of the PRISM framework. PRISM combines victim-side state salience with behavioral information from independently trained clean peer policies to identify informative poisoning states. The selected states are then used for reward-preserving target induction, while peer regularization helps preserve benign policy behavior.

3 METHOD

PRISM formulates RL backdoor implantation as a selective policy modification problem. Rather than poisoning arbitrary states, PRISM exploits independently trained clean peer policies to identify states in which the target behavior can be introduced with relatively low conflict.

An effective poisoning state should satisfy two complementary properties: (1) it should be influential to the victim policy, and (2) the target action should remain compatible with benign peer behavior. Accordingly, PRISM consists of three components: consensus-guided state mining, reward-preserving target induction, and peer-regularized policy optimization.

3.1 CONSENSUS-GUIDED STATE MINING

Given a victim rollout $\{s_i\}_{i=1}^{N}$, PRISM evaluates candidate poisoning states using both victim-side salience and peer-side behavioral compatibility.

The victim salience score is defined as

$$v_i = \omega_h h_\theta(s_i) + \omega_u u_\theta(s_i), \tag{1}$$

where $h_\theta(s_i)$ denotes the hidden-representation activation of the victim policy π_θ, and $u_\theta(s_i)$ denotes its decision confidence derived from normalized policy entropy. The coefficients ω_h and ω_u balance the two components.

Meanwhile, the M independently trained clean peer policies provide behavioral information about the target action. For a target action $a^\dagger$, the peer consensus score is defined as

$$c_i = \frac{1}{M} \sum_{m=1}^{M} \pi_{\phi_m}(a^\dagger | s_i), \tag{2}$$

where $\pi_{\phi_m}(a^\dagger|s_i)$ denotes the probability assigned to the target action by the m-th clean peer.

PRISM combines victim salience and peer consensus using a multiplicative priority score:

$$q_i = v_i^{\beta} c_i^{1-\beta}, \tag{3}$$

where β controls the trade-off between victim-side salience and peer-side compatibility.

Given poisoning rate ρ, PRISM ranks candidate states according to q_i and selects the highest-priority states:

$$\mathcal{B} = \mathrm{TopK}\left(\{s_i\}_{i=1}^{N}, K\right), \qquad K \sim \mathrm{Binomial}(N, \rho). \tag{4}$$

3.2 Reward-Preserving Target Induction

For each selected state $s_i \in \mathcal{B}$, PRISM generates a triggered observation

$$s_i^{\dagger} = \mathcal{T}(s_i). \tag{5}$$

The target behavior is induced through an auxiliary policy objective:

$$\mathcal{L}_{\mathrm{ind}} = -\frac{1}{|\mathcal{B}|} \sum_{s_i \in \mathcal{B}} q_i \log \pi_\theta \left(a^{\dagger} \mid s_i^{\dagger}\right). \tag{6}$$

Here, q_i denotes the priority score of the selected state and $a^{\dagger}$ is the attacker-specified target action. Consequently, states that are both important to the victim and compatible with benign peer behavior receive stronger induction pressure.

Importantly, PRISM leaves the original environmental reward and transition dynamics unchanged. The triggered observations are used only by the auxiliary induction objective and do not replace the original PPO rollout.

3.3 Peer-Regularized Optimization

To reduce interference with benign behavior, PRISM aggregates the clean peer policies into a reference policy:

$$\pi_p(a|s) = \frac{1}{M} \sum_{m=1}^{M} \pi_{\phi_m}(a|s), \tag{7}$$

where π_p represents the consensus behavior of the clean peers.

For clean observations, PRISM encourages the victim policy to remain close to the peer consensus by minimizing

$$\mathcal{L}_{\mathrm{peer}} = D_{\mathrm{KL}}\left(\pi_p(\cdot|s) \| \pi_\theta(\cdot|s)\right). \tag{8}$$

The final optimization objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{PPO}} + \eta \mathcal{L}_{\mathrm{ind}} + \lambda \mathcal{L}_{\mathrm{peer}}, \tag{9}$$

where η and λ control the strengths of target induction and peer regularization, respectively.

By jointly optimizing these objectives, PRISM increases the probability of the attacker-specified target action under triggered observations while constraining the victim policy to remain close to benign behavioral references on clean states.

3.4 Theoretical Analysis

Theorem 1 (Targeted Policy Induction). For a selected triggered state $s^\dagger = \mathcal{T}(s)$, the induction objective is defined as

$$\ell(\theta) = -w \log \pi_\theta(a^\dagger | s^\dagger), \tag{10}$$

where $w > 0$ denotes the state weight and $a^\dagger$ is the attacker-specified target action. If

$$\ell(\theta) < w \log 2, \tag{11}$$

then

$$\pi_\theta(a^\dagger | s^\dagger) > \max_{a \neq a^\dagger} \pi_\theta(a | s^\dagger). \tag{12}$$

Theorem 2 (Clean-Peer-Anchored Return Preservation). Let π_p denote the aggregated clean-peer policy:

$$\pi_p(a|s) = \frac{1}{M} \sum_{m=1}^{M} \pi_{\phi_m}(a|s). \tag{13}$$

Assume bounded rewards:

$$|r(s,a)| \leq R_{\max}. \tag{14}$$

If

$$\mathbb{E}_{s \sim \mu_b} \left[D_{\mathrm{KL}} \left(\pi_p(\cdot|s) \| \pi_\theta(\cdot|s) \right) \right] \leq \epsilon, \tag{15}$$

and

$$D_{\mathrm{TV}} \left(d_{\pi_\theta}, \mu_b \right) \leq \xi, \tag{16}$$

then

$$|J(\pi_\theta) - J(\pi_{\mathrm{ben}})| \leq \Delta_{\mathrm{peer}} + \frac{4R_{\max}}{(1-\gamma)^2} \left(\sqrt{\frac{\epsilon}{2}} + \xi \right), \tag{17}$$

where

$$\Delta_{\mathrm{peer}} = \left| J(\pi_p) - J(\pi_{\mathrm{ben}}) \right|. \tag{18}$$

Theorem 3 (Fixed-Rollout PPO Target Preservation). For a fixed PPO rollout, PRISM preserves the original temporal-difference residuals and advantage estimates:

$$\delta_t^{\mathrm{PRISM}} = \delta_t^{\mathrm{PPO}}, \tag{19}$$

and

$$\hat{A}_t^{\mathrm{PRISM}} = \hat{A}_t^{\mathrm{PPO}}. \tag{20}$$

The complete proofs of Theorems 1–3 are provided in Appendix A.

4 Experiment

4.1 Experiment Setup

Environments. We evaluate PRISM on three categories of RL tasks: Atari 2600 visual-control environments Bellemare et al. (2013), Safety-Gymnasium control tasks Ji et al. (2023), and autonomous-driving environments from HighwayEnv Leurent (2018). These benchmarks cover both high-dimensional visual observations and low-dimensional state representations, allowing us to evaluate the generality of PRISM across different observation modalities and decision dynamics.

Attack Configuration. We consider an outer-loop poisoning setting in which the attacker can access training trajectories and modify a limited subset of samples before policy optimization. The poisoning budget is controlled by p_{rate}. The objective is to induce the victim policy to execute a predefined target action $a^\dagger$ under triggered observations while preserving its behavior on clean states. Unlike reward-manipulation-based attacks, PRISM does not modify the environmental reward or transition dynamics.

Baselines. We compare PRISM with representative RL backdoor attacks: TrojDRL Kiourti et al. (2020), BadRL Cui et al. (2024), SleeperNets Rathbun et al. (2024), Q-Incept Rathbun et al. (2025), and SpecDRL Qian et al. (2026). Their original attack mechanisms are retained while being implemented under the same PPO training and evaluation framework whenever possible.

Table 1: Overall comparison with existing RL backdoor attacks. ASR denotes Attack Success Rate, and Return denotes the average episodic return obtained by the attacked policy on benign observations.

Environment	TrojDRL		BadRL		SleeperNets		Q-Incept		SpecDRL		Ours	
	ASR	Return	ASR	Return	ASR	Return	ASR	Return	ASR	Return	ASR	Return
Atari 2600 Games (NoFrameskip-v4)												
BeamRider	94.51%	605.36	97.23%	635.68	90.28%	601.48	93.94%	**696.60**	**100%**	641.72	**100%**	664.44
Breakout	99.88%	41.71	97.44%	34.04	**100%**	32.79	98.16%	44.29	**100%**	42.90	**100%**	**46.66**
SpaceInvaders	21.44%	456.85	12.31%	451.95	**100%**	475.30	8.92%	420	**100%**	329.65	**100%**	**503.5**
Qbert	96.62%	1862.75	7.57%	2095.75	**100%**	2510	51.03%	1833.75	**100%**	1893.75	**100%**	**3039.75**
Pong	99.75%	2.24	7.81%	12.10	**100%**	14.76	**100%**	12.42	**100%**	5.90	**100%**	**17.21**
Gopher	99.40%	**1038.60**	99.35%	989.80	**100%**	923.20	13.42%	867.40	**100%**	966.20	**100%**	1021.80
Safety Environments												
CarGoal1	81.71%	11.49	60.72%	11.51	93.40%	11.82	67.77%	11.92	65.49%	11.88	**96.53%**	**11.93**
PointPush1	97.61%	0.67	79.06%	0.39	91.22%	1.15	92.70%	0.92	99.96%	1.16	**100%**	**1.28**
PointGoal1	94.88%	14.71	47.86%	14.22	**100%**	15.76	89.89%	**16.23**	99.98%	13.88	**100%**	16.03
CarPush1	92.22%	**-0.92**	82.33%	-1.93	90.90%	-1.68	77.00%	-1.02	69.63%	-1.03	**97.27%**	-1.30
CarCircle1	84.37%	-1.14	91.17%	7.18	**99.99%**	8.25	81.96%	11.11	93.62%	9.92	99.51%	**11.23**
Highway Environment												
merge	29.80%	7.33	29.82%	7.25	30.21%	7.42	24.57%	7.66	30.15%	6.84	**33.37%**	**7.74**

4.2 Evaluation Metrics

We evaluate PRISM in terms of attack effectiveness, benign task performance, poisoning efficiency, and multi-agent guidance.

Attack Success Rate (ASR). ASR measures the probability that the attacked policy executes the target action under triggered observations:

$$\mathrm{ASR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\pi_\theta(\mathcal{T}(s_i)) = a^\dagger\right), \tag{21}$$

where N is the number of triggered evaluation states, $\mathcal{T}(\cdot)$ denotes the trigger transformation, and $a^\dagger$ is the target action. A higher ASR indicates stronger backdoor effectiveness.

Benign Task Performance. We report the average episodic return of the attacked policy on clean observations. This directly measures how well the original task capability is preserved after backdoor implantation.

Poisoning Efficiency. We evaluate how rapidly the attack establishes the trigger–target association under a limited poisoning budget by tracking the evolution of ASR during training.

Multi-Agent Guidance Effectiveness. We compare PRISM with variants that remove peer information or use only a single peer policy to quantify the contribution of population-level behavioral guidance.

All experiments are repeated with multiple random seeds, and we report the mean and standard deviation.

4.3 Main Results

Attack Effectiveness and Benign Performance. Table 1 compares PRISM with existing RL backdoor attacks in terms of attack effectiveness and benign task performance. Overall, PRISM achieves consistently strong attack performance across different environments while maintaining competitive clean-task returns. These results indicate that PRISM can effectively induce the target behavior without substantially degrading the victim policy's original task capability.

Early-stage Attack Efficiency. Table 2 compares the ASR achieved by different attacks at predefined early training checkpoints under the same poisoning budget. PRISM generally exhibits

Table 2: Attack convergence comparison under different training steps. ASR denotes Attack Success Rate. Higher ASR at earlier training stages indicates faster backdoor implantation.

Environment (Steps)	TrojDRL	BadRL	SleeperNets	Q-Incept	SpecDRL	**Ours**
Atari 2600 Games						
Breakout (100K)	50.03%	53.94%	41.41%	37.63%	41.25%	**84.04%**
BeamRider (100K)	6.08%	8.38%	14.26%	8.19%	7.15%	**22.04%**
SpaceInvaders (100K)	16.01%	22.98%	5.92%	11.36%	12.35%	**92.73%**
Qbert (300K)	3.60%	4.29%	13.21%	7.54%	4.48%	**99.66%**
Gopher (100K)	9.08%	11.52%	10.41%	12.97%	6.37%	**97.46%**
Pong (100K)	13.11%	13.09%	15.97%	21.54%	7.09%	**69.35%**
Safety Environments						
CarGoal1 (500K)	52.63%	43.00%	54.74%	11.97%	18.09%	**63.86%**
CarPush1 (100K)	39.58%	17.50%	51.30%	4.63%	44.49%	**52.22%**
PointGoal1 (100K)	22.27%	16.49%	64.07%	10.14%	56.45%	**66.90%**
PointPush1 (200K)	**71.36%**	53.79%	51.92%	18.13%	25.75%	64.83%
PointButton1 (100K)	18.93%	21.49%	21.81%	37.26%	12.21%	**42.74%**
Highway Environment						
merge (10K)	16.23%	16.69%	16.77%	11.11%	17.34%	**18.48%**

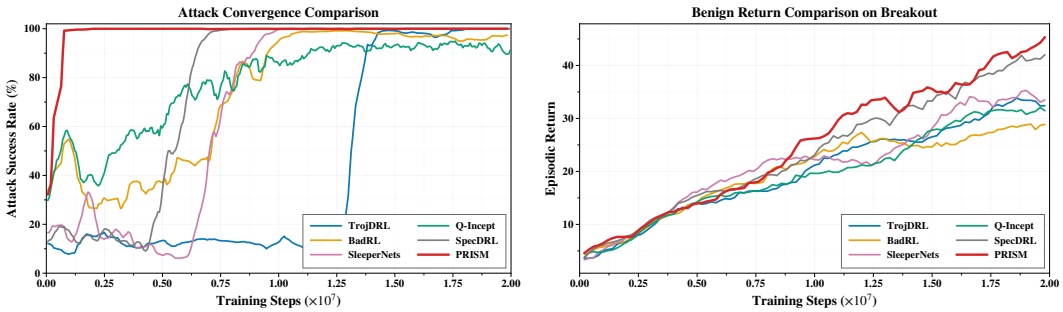


Figure 2: Training dynamics of different backdoor attacks on Breakout. Left: ASR throughout training, where PRISM establishes a high and stable attack success rate substantially earlier than the compared baselines. Right: benign episodic return throughout training, showing that PRISM preserves competitive task performance while rapidly implanting the backdoor.

stronger early-stage attack performance, indicating that clean-peer-guided state mining enables more efficient trigger–target association.

Training Dynamics on Breakout. To further examine the attack dynamics, we compare the ASR and benign episodic return of all methods throughout training on Breakout. Figure 2 shows that PRISM rapidly establishes the trigger–target association and reaches a high, stable ASR much earlier than the compared baselines. At the same time, its benign return continues to improve during training and remains competitive with the strongest baselines. These results show that the accelerated attack convergence of PRISM does not come at the expense of benign task performance.

4.4 Mechanism Analysis

Component Interventions. We investigate the contribution of individual PRISM components through controlled interventions on Gopher. All variants are evaluated under the same initialization, rollout procedure, and optimization configuration, with only the corresponding component removed.

As shown in Fig. 3, removing the induction objective substantially decreases the target-action response, indicating that the auxiliary induction mechanism plays a critical role in establishing the trigger–target association. Removing either victim salience or peer consensus also leads to weaker induction performance than the full PRISM model, suggesting that these two signals provide complementary information for selecting effective trigger states.

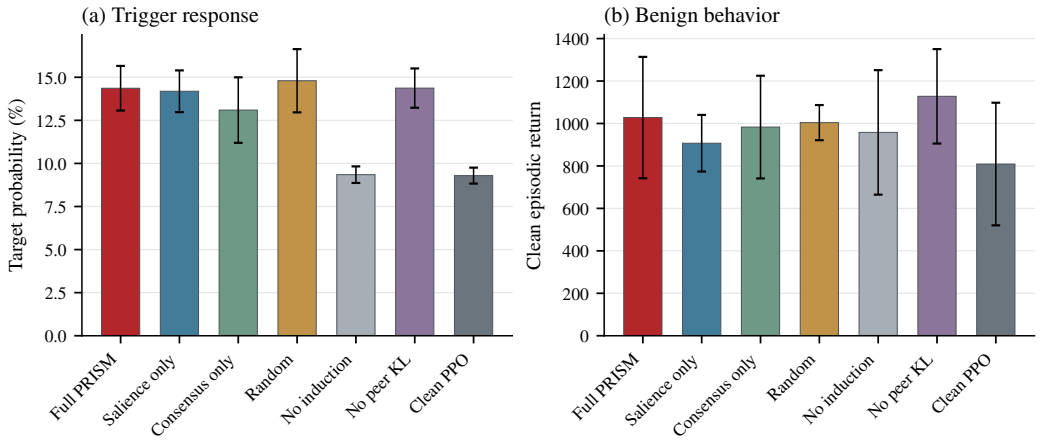


Figure 3: Component interventions on Gopher. Induction response and clean return are reported separately; error bars indicate variation across random seeds.

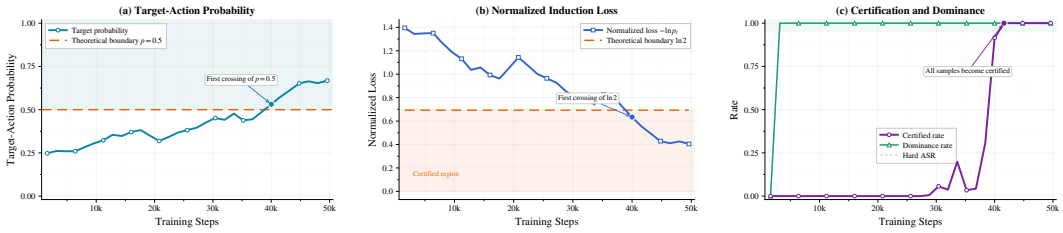


Figure 4: Empirical validation of Theorem 1 on Breakout. Left: target-action probability and the $p = 0.5$ boundary. Middle: normalized induction loss and the $\log 2$ boundary. Right: certification, dominance, and attack-success rates.

The complete PRISM model achieves stronger target response while preserving comparable clean performance. Removing peer consensus causes a moderate change in both induction response and clean return, suggesting that peer regularization mainly improves behavioral consistency by constraining undesired policy deviation. In contrast, victim-side salience provides informative state prioritization, enabling more effective trigger-state selection.

Overall, these controlled interventions demonstrate that the effectiveness of PRISM arises from the joint interaction of selective state mining, reward-preserving induction, and peer-guided regularization, rather than from increasing the amount of poisoned data alone.

4.5 Empirical Validation of Theoretical Properties

We further connect the theoretical properties established in Appendix A with the empirical behavior of PRISM. Specifically, we examine targeted policy induction (Theorem 1), clean-peer-anchored return preservation (Theorem 2), and fixed-rollout PPO target preservation (Theorem 3).

Targeted Policy Induction. Theorem 1 gives a sufficient condition for the attacker-specified action to become dominant under triggered observations.

As shown in Fig. 4, the target-action probability progressively increases while the normalized induction loss decreases, and both cross the corresponding boundaries specified in Theorem 1. The dominance rate rises earlier than the certified rate because the theoretical condition is sufficient rather than necessary for the target action to become dominant. Overall, the results show that PRISM progressively establishes the trigger–target association until the evaluated triggered states enter the theoretically certified region.

Clean-Peer-Anchored Return Preservation. Theorem 2 relates benign-performance preservation to the deviation between the victim policy π_θ and the aggregated clean-peer reference π_p.

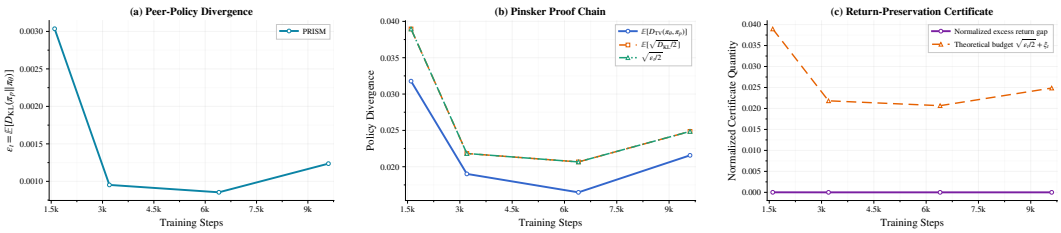


Figure 5: Empirical validation of Theorem 2 on Breakout. Left: victim–peer KL divergence. Middle: empirical TV deviation and KL-derived bounds. Right: normalized return deviation and theoretical budget.

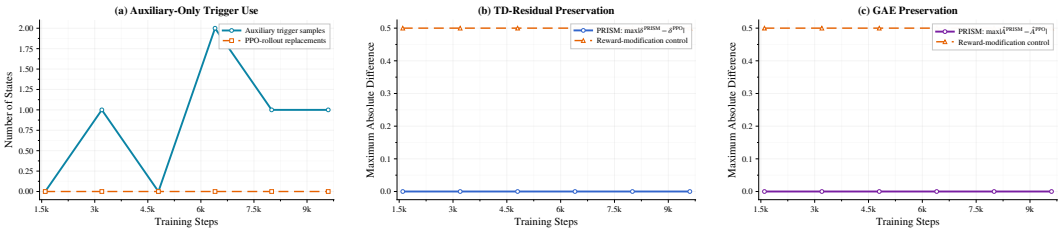


Figure 6: Empirical validation of Theorem 3 on Breakout. Left: auxiliary trigger use without PPO-rollout replacement. Middle: TD-residual preservation. Right: GAE preservation.

As shown in Fig. 5(a), the measured victim–peer divergence ϵ_t remains small, indicating that the victim stays close to the clean-peer behavioral reference on benign states. Fig. 5(b) further shows that the empirical total-variation deviation remains below the corresponding KL-derived bounds, consistent with the relation used in Theorem 2. Finally, Fig. 5(c) shows that the normalized benign-return deviation remains below the theoretical budget determined by ϵ_t, the occupancy mismatch ξ_t, and the intrinsic peer-reference gap Δ_{peer}. Thus, the observed benign-performance deviation remains within the return-preservation region characterized by Theorem 2.

Fixed-Rollout PPO Target Preservation. Theorem 3 states that, for a fixed PPO rollout, PRISM preserves the original temporal-difference residuals and GAE advantages because triggered observations are used only by the auxiliary induction objective.

Fig. 6(a) shows that auxiliary triggered samples are generated during training while the original PPO rollout is never replaced. Accordingly, the PRISM–PPO TD-residual difference remains zero throughout training (Fig. 6(b)), and the same behavior is observed for the GAE advantages (Fig. 6(c)). In contrast, deliberately modifying the rollout reward produces clear nonzero differences in both quantities. These results confirm that PRISM introduces the auxiliary backdoor objective without rewriting the original PPO learning targets of the fixed rollout.

Taken together, the three experiments provide a direct empirical counterpart to the theoretical analysis: PRISM establishes the trigger–target association predicted by Theorem 1, constrains benign-policy deviation as characterized by Theorem 2, and preserves the original PPO targets of a fixed rollout as stated in Theorem 3.

5 Conclusion

We presented PRISM, a multi-policy-assisted backdoor framework that reveals a new attack surface in reinforcement learning, where benign policies can provide unintended guidance for backdoor implantation. By combining victim-side state salience with clean-peer target-action compatibility, PRISM enables selective target induction while preserving benign behavior without reward modification. Experiments across diverse RL domains demonstrate that PRISM achieves effective and efficient backdoor implantation, highlighting the security risks introduced by benign policy populations in multi-policy RL systems.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT) for language polishing and improving the clarity of the manuscript. We also used generative AI tools as an auxiliary discussion partner for exploring the feasibility of research ideas and for assisting code optimization during implementation.

The proposed method, theoretical analysis, experimental design, execution, and interpretation of results were conducted by the authors. We did not use generative AI tools to generate experimental results, derive theoretical claims, or replace the authors' scientific reasoning.

All AI-assisted suggestions were carefully reviewed and verified by the authors. In particular, AI-assisted code modifications were tested and validated in the experimental pipeline, and all manuscript revisions were manually checked for correctness and consistency. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013. doi: 10.1613/JAIR.3912.

Xuan Chen, Wenbo Guo, Guanhong Tao, Xiangyu Zhang, and Dawn Song. BIRD: Generalizable backdoor detection and removal for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1777. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/802e90325f4c8546e13e5763b2ecab88-Abstract-Conference.html.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming llm agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4136. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb113910e9c3f6242541c1652e30dfd6-Abstract-Conference.html.

Jing Cui, Yufei Han, Yuzhe Ma, Jianbin Jiao, and Junge Zhang. BadRL: Sparse targeted backdoor attack against reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11687–11694, 2024. doi: 10.1609/aaai.v38i10.29052.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017. URL https://arxiv.org/abs/1711.09846.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety-Gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. TrojDRL: Evaluation of backdoor attacks on deep reinforcement learning. In *Proceedings of the 57th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2020. doi: 10.1109/DAC18072.2020.9218663.

Edouard Leurent. An environment for autonomous driving decision-making. GitHub repository, 2018. URL https://github.com/Farama-Foundation/HighwayEnv.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017. URL https://arxiv.org/abs/1701.07274.

Guanlin Liu and Lifeng Lai. Efficient adversarial attacks on online multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1062. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4cddc8fc57039f8fe44e23aba1e4df40-Abstract-Conference.html.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

Yaguan Qian, Taining Zhang, Qiqi Bao, Yanru Guo, Lufang Zhang, Zhaoquan Gu, Shouling Ji, Bin Wang, and Zhen Lei. Toward subspace-perturbed trajectory-aware backdoor attacks in deep reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Ethan Rathbun, Christopher Amato, and Alina Oprea. SleeperNets: Universal backdoor poisoning attacks against reinforcement learning agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111994–112024, 2024. doi: 10.52202/079017-3556.

Ethan Rathbun, Alina Oprea, and Christopher Amato. Adversarial inception backdoor attacks against reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51273–51296. PMLR, 2025. URL https://proceedings.mlr.press/v267/rathbun25a.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition, 2018.

Sanyam Vyas, Alberto Caron, Chris Hicks, Pete Burnap, and Vasilios Mavroudis. Beyond training-time poisoning: Component-level and post-training backdoors in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26072–26080, 2026. doi: 10.1609/aaai.v40i31.39809. URL https://ojs.aaai.org/index.php/AAAI/article/view/39809.

Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. BACKDOORL: Backdoor attack against competitive reinforcement learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3699–3705, 2021. doi: 10.24963/ijcai.2021/509.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Reward poisoning attacks on offline multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10426–10434, 2023. doi: 10.1609/aaai.v37i9.26240. URL https://ojs.aaai.org/index.php/AAAI/article/view/26240.

Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie. Data poisoning to fake a nash equilibria for markov games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15979–15987, 2024. doi: 10.1609/aaai.v38i14.29529. URL https://ojs.aaai.org/index.php/AAAI/article/view/29529.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. Watch out for your agents! investigating backdoor threats to llm-based agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3201. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b6e9d6f4f3428cd5f3f9e9bbae2cab10-Abstract-Conference.html.

A Theoretical Analysis

We provide the complete theoretical analysis of three key properties of PRISM: (1) targeted policy induction under triggered observations, (2) benign performance preservation through clean-peer regularization, and (3) preservation of the original PPO learning targets under a fixed rollout.

A.1 TARGETED POLICY INDUCTION

Theorem 1 (Targeted Policy Induction).

For a selected triggered state $s^\dagger = \mathcal{T}(s)$, the induction objective is defined as

$$\ell(\theta) = -w \log \pi_\theta(a^\dagger | s^\dagger), \tag{22}$$

where $w > 0$ denotes the state weight and $a^\dagger$ is the attacker-specified target action.

If

$$\ell(\theta) < w \log 2, \tag{23}$$

then

$$\pi_\theta(a^\dagger | s^\dagger) > \max_{a \neq a^\dagger} \pi_\theta(a | s^\dagger). \tag{24}$$

Proof.

Let

$$p^\dagger = \pi_\theta(a^\dagger | s^\dagger). \tag{25}$$

According to the induction loss,

$$\ell(\theta) = -w \log p^\dagger. \tag{26}$$

Since $w > 0$, the condition $\ell(\theta) < w \log 2$ implies

$$-\log p^\dagger < \log 2, \tag{27}$$

which is equivalent to

$$p^\dagger > \frac{1}{2}. \tag{28}$$

Because the policy output is a valid probability distribution,

$$\sum_{a \neq a^\dagger} \pi_\theta(a | s^\dagger) = 1 - p^\dagger < \frac{1}{2}. \tag{29}$$

Therefore, every non-target action satisfies

$$\pi_\theta(a | s^\dagger) < \frac{1}{2}, \qquad a \neq a^\dagger. \tag{30}$$

Hence,

$$\pi_\theta(a^\dagger | s^\dagger) > \max_{a \neq a^\dagger} \pi_\theta(a | s^\dagger). \tag{31}$$

Thus, minimizing the induction loss sufficiently establishes target-action dominance under the trigger condition. □

A.2 CLEAN-PEER-ANCHORED RETURN PRESERVATION

Theorem 2 (Clean-Peer-Anchored Return Preservation).

Let π_p denote the aggregated clean-peer policy:

$$\pi_p(a|s) = \frac{1}{M} \sum_{m=1}^{M} \pi_{\phi_m}(a|s). \tag{32}$$

Assume bounded rewards:

$$|r(s,a)| \leq R_{\max}. \tag{33}$$

If the KL divergence between the victim policy and the clean-peer reference satisfies

$$\mathbb{E}_{s\sim\mu_b} \left[D_{\mathrm{KL}} \left(\pi_p(\cdot|s) \| \pi_\theta(\cdot|s) \right) \right] \leq \epsilon, \tag{34}$$

and the occupancy-distribution shift satisfies

$$D_{\mathrm{TV}}(d_{\pi_\theta}, \mu_b) \leq \xi, \tag{35}$$

then the return degradation is bounded by

$$|J(\pi_\theta) - J(\pi_{\mathrm{ben}})| \leq \Delta_{\mathrm{peer}} + \frac{4R_{\max}}{(1-\gamma)^2} \left(\sqrt{\frac{\epsilon}{2}} + \xi \right), \tag{36}$$

where

$$\Delta_{\mathrm{peer}} = |J(\pi_p) - J(\pi_{\mathrm{ben}})| \tag{37}$$

denotes the intrinsic performance gap between the clean-peer consensus and the benign reference policy.

Proof.

Using the Performance Difference Lemma,

$$J(\pi_\theta) - J(\pi_p) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{\pi_\theta}} \left[\mathbb{E}_{a\sim\pi_\theta} A_{\pi_p}(s,a) \right]. \tag{38}$$

Since

$$\mathbb{E}_{a\sim\pi_p} A_{\pi_p}(s,a) = 0, \tag{39}$$

the performance difference depends on the policy deviation:

$$\left| \mathbb{E}_{a\sim\pi_\theta} A_{\pi_p}(s,a) \right| \leq \frac{4R_{\max}}{1-\gamma} D_{\mathrm{TV}} \left(\pi_\theta(\cdot|s), \pi_p(\cdot|s) \right). \tag{40}$$

By Pinsker's inequality,

$$D_{\mathrm{TV}}(p,q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p\|q)}. \tag{41}$$

Therefore,

$$\mathbb{E}_{s\sim\mu_b} D_{\mathrm{TV}}(\pi_\theta, \pi_p) \le \sqrt{\frac{\epsilon}{2}}. \tag{42}$$

The mismatch between d_{π_θ} and μ_b introduces an additional error bounded by ξ. Combining the two terms gives

$$|J(\pi_\theta) - J(\pi_p)| \le \frac{4R_{\max}}{(1-\gamma)^2}\left(\sqrt{\frac{\epsilon}{2}} + \xi\right). \tag{43}$$

Finally, by the triangle inequality,

$$|J(\pi_\theta) - J(\pi_{\mathrm{ben}})| \le |J(\pi_\theta) - J(\pi_p)| + |J(\pi_p) - J(\pi_{\mathrm{ben}})|. \tag{44}$$

Substituting Δ_{peer} completes the proof. □

A.3 Fixed-Rollout PPO Target Preservation

Theorem 3 (Fixed-Rollout PPO Target Preservation).

For a fixed PPO rollout, PRISM preserves the original temporal-difference residuals and advantage estimates:

$$\delta_t^{\mathrm{PRISM}} = \delta_t^{\mathrm{PPO}}, \tag{45}$$

and

$$\hat{A}_t^{\mathrm{PRISM}} = \hat{A}_t^{\mathrm{PPO}}. \tag{46}$$

Proof.

The PPO temporal-difference residual is

$$\delta_t = r_t + \gamma(1 - d_{t+1})V_{t+1} - V_t. \tag{47}$$

For a fixed rollout, PRISM does not modify the environmental rewards r_t, transition dynamics, termination signals d_t, or stored value estimates V_t.

The triggered observations

$$s^\dagger = \mathcal{T}(s) \tag{48}$$

are used only by the auxiliary induction objective and do not replace the original PPO rollout. Therefore,

$$\delta_t^{\mathrm{PRISM}} = \delta_t^{\mathrm{PPO}}. \tag{49}$$

Generalized Advantage Estimation (GAE) Schulman et al. (2016) is computed recursively as

$$\hat{A}_t = \delta_t + \gamma\lambda(1 - d_{t+1})\hat{A}_{t+1}. \tag{50}$$

Because the TD residuals remain unchanged for the fixed rollout, the resulting advantage estimates are also unchanged:

$$\hat{A}_t^{\mathrm{PRISM}} = \hat{A}_t^{\mathrm{PPO}}. \tag{51}$$

Thus, PRISM preserves the original PPO learning targets of the fixed rollout while introducing additional policy-space optimization for backdoor induction. □

B Additional Experimental Results

Controlled-update protocol. We evaluate PRISM under controlled policy updates using a common initial PPO checkpoint at 2M environment steps, fixed clean rollouts, and frozen peer policies. Gopher is used for parameter and mechanism analyses, with the same component interventions also tested on Pong. Three peers are obtained by independent clean PPO continuations from the common checkpoint using seeds 1, 2, and 3, each trained for 16K environment steps with the original reward.

Optimization and evaluation settings. All experiments use Adam with a learning rate of 2.5×10^{-4} and $\epsilon = 10^{-5}$. The discount factor and GAE parameter are set to 0.99 and 0.95, respectively. PPO uses a clipping coefficient of 0.1, an entropy coefficient of 0.01, and a value-loss coefficient of 0.5, with gradient norms clipped at 0.5. For each seed, we collect 25,600 rollout positions and 2,048 evaluation observations. Optimization consists of one shuffled pass with a batch size of 256, resulting in 100 updates. Results are reported over five paired seeds, with four complete return-evaluation games per seed. Experiments are run on an NVIDIA RTX A6000 (48 GiB) using Python 3.10.0, PyTorch 2.9.1, and Gymnasium 0.29.1.

B.1 Effect of the Number of Clean Peer Policies

We investigate how the number of clean peer policies affects the attack effectiveness of PRISM on SafetyPointGoal and Breakout under a fixed 100K training budget. We vary the number of peers as $M \in \{1, 2, 3, 5, 8, 10, 15\}$ while keeping the remaining attack settings unchanged.

The results show that the relationship between the number of clean peers and attack success is clearly non-monotonic. On SafetyPointGoal, the ASR generally increases as more peer policies are introduced, rising from 32.62% with one peer to 38.38% with five peers, and further to 49.42% and 66.90% with eight and ten peers, respectively. However, increasing the peer pool to fifteen reduces the ASR to 51.23%. A similar non-monotonic pattern is observed on Breakout. The ASR varies substantially across different peer counts, from 84.04% with one peer to 55.59% with two peers, before recovering to 79.19% and 81.28% with three and five peers. The highest ASR of 93.98% is again achieved with ten peers, whereas using fifteen peers causes a sharp decrease to 43.06%.

These results indicate that simply increasing the number of clean peer policies does not continuously strengthen the attack. A moderate peer pool can provide more diverse and complementary policy references, improving the reliability of peer-based state evaluation and facilitating the identification of states that are suitable for backdoor induction. In contrast, an excessively large peer set may introduce heterogeneous or conflicting policy preferences, causing the aggregated peer signal to become less discriminative and thereby weakening the effectiveness of state selection and policy induction. Notably, both environments achieve their highest ASR with ten peers, suggesting that PRISM benefits from sufficient peer diversity but is also sensitive to over-aggregation. Therefore, the number of peer policies should be treated as a controllable design parameter rather than assuming that more peers necessarily lead to stronger attacks.

B.2 Ablation Study

Component Ablation. We investigate the contribution of the key components in PRISM's consensus-guided state-mining strategy. As shown in Fig. 7, removing either victim-side salience or clean-peer consensus leads to a pronounced degradation in attack effectiveness, indicating that the two signals provide complementary information rather than serving as interchangeable heuristics. Victim salience helps identify states that are influential to the current policy, while peer consensus favors states in which the target action is more compatible with benign behavioral patterns. Replacing the joint priority criterion with random state selection also results in weaker performance under the same poisoning budget, confirming that the effectiveness of PRISM is not merely due to the presence of poisoned samples. Overall, these results validate the design of consensus-guided state mining and demonstrate the importance of jointly considering victim-side importance and clean-peer behavioral compatibility.

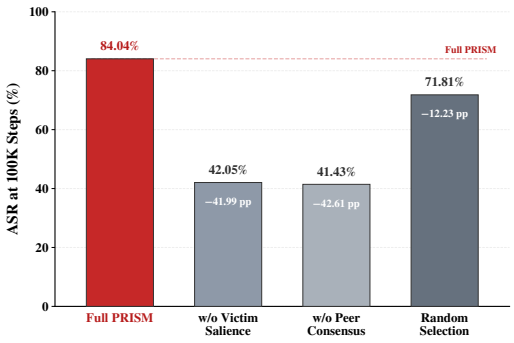


Figure 7: Component ablation of PRISM on Breakout at 100K training steps. Removing victim-side salience or clean-peer consensus substantially weakens attack effectiveness, while random state selection also underperforms the full priority-guided strategy.