Counterfactual Shapley Credit Assignment

Mingxuan Li, Kai-Zhan Lee, Elias Bareinboim

Keywords: temporal credit assignment, counterfactual Shapley values, causal inference

Summary

The Credit Assignment Problem (CAP) is fundamental to developing efficient and explainable Reinforcement Learning (RL) agents. Existing frameworks, whether relying on temporal contiguity or hindsight-conditioned reward reweighting, frequently fail to attribute properly between an agent's policy (skill) and environmental stochasticity (luck). A principled approach to CAP must isolate the true causal drivers of observed outcomes from spurious correlations and environmental randomness. We introduce Counterfactual Shapley Credit Assignment, a novel framework grounded in causal theory that attributes credit and blame via the Counterfactual Shapley Value (ϕ-value). By redistributing environmental rewards, ϕ-values enhance temporal credit assignment across three critical dimensions: sparse causality, high stochasticity, and delayed rewards, all while preserving the optimal policy. We derive a consistent estimator that computes ϕ-values efficiently, enabling a new class of policy gradient methods, ϕ-PPO, combined with Prioritized Trajectory Replay (PTR). Empirical results demonstrate that ϕ-values align precisely to the ground truth causes of task rewards with superior sample efficiency in challenging environments where prior state-of-the-art methods fail to converge.

Contribution(s)

1. We apply the Counterfactual Shapley Value (ϕ-value) framework Lee et al. (2025) to RL temporal credit assignment, showing that ϕ-values yield a reward redistribution that preserves the optimal policy and concentrates credit on causal actions. We show that the proposed framework satisfies the four causal credit assignment criteria: (1) Causal Admissibility (2) Causal Power (3) Causal Normality and (4) Causal Effect Scaling. The four criteria together guarantee that the proposed framework allocates credits faithful to how much causal contribution each action has to the outcome compared to a baseline policy, effectively rectifying the long-standing challenges in RL temporal credit assignment: sparsity, high stochasticity and delayed rewards.
   Context: Prior credit assignment methods, including HCA (Harutyunyan et al., 2019a), CCA (Mesnard et al., 2021), CoCoA (Meulemans et al., 2023), RUDDER (Arjona-Medina et al., 2019), and Synthetic Returns (Raposo et al., 2021), are not grounded in counterfactuals in the Structural Causal Model (SCM) framework (Pearl, 2009; Bareinboim et al., 2022).
2. We derive a consistent ϕ-value function estimator converging to the true causal contributions in probability given sufficient samples with amortized $O(1)$ time per action and a tunable bias–variance tradeoff via λ-bootstrapping.
   Context: Previous work on estimating Shapley values (Shapley, 1953) and ϕ-values (Lee et al., 2025) don't consider bootstrapping with value function approximators. Thus, they usually suffer from either exponential sample complexity or biased estimation.
3. We introduce ϕ-PPO with Prioritized Trajectory Replay (PTR) and demonstrate superior sample efficiency on the SkillLuck and Combinatorial Lock benchmarks, where prior methods either take much longer to learn or fail to converge; through ablation, we show that both ϕ-values and PTR are necessary.
   Context: Compared against PPO and REINFORCE (Schulman et al., 2017; Williams, 1992), HCA (Harutyunyan et al., 2019b), CCA (Mesnard et al., 2021), CoCoA (Meulemans et al., 2023), QCA (Mesnard et al., 2023), RUDDER (Arjona-Medina et al., 2019), Synthetic Returns (Raposo et al., 2021), and H-DICE (Velu et al., 2024) in three classical tabular environments characterizing the challenges in temporal credit assignment.

Counterfactual Shapley Credit Assignment

Mingxuan Li[1,†], **Kai-Zhan Lee**[1,†], **Elias Bareinboim**[1]
ml@cs.columbia.edu, kl2792@columbia.edu, eb@cs.columbia.edu

[1]**Department of Computer Science, Columbia University, New York, NY, USA**

[†] Equal contribution

Abstract

The Credit Assignment Problem (CAP) is fundamental to developing efficient and explainable Reinforcement Learning (RL) agents. Existing frameworks, whether relying on temporal contiguity or hindsight-conditioned reweighting of rewards, frequently fail to distribute credit properly between an agent's policy (skill) and environmental stochasticity (luck). A principled approach to the CAP must isolate the causes of observed rewards from spurious correlated features and environmental randomness. We introduce Counterfactual Shapley Credit Assignment, a novel causal credit assignment framework based on the counterfactual Shapley value (ϕ-value). By redistributing rewards, ϕ-values enhance credit assignment across three critical dimensions: high stochasticity, sparse causality, and delayed rewards, all while theoretically preserving the original optimal policy. We derive a consistent estimator that computes each ϕ-value in amortized constant time complexity, enabling a new class of policy gradient methods, ϕ-PPO. Empirical results demonstrate that ϕ-values align precisely with the ground truth causes of task rewards. Furthermore, we show its superior sample efficiency in challenging environments where prior state-of-the-art methods fail to converge.

1 Introduction

Reinforcement learning (RL) agents have achieved remarkable success in domains such as healthcare (Yu et al., 2020), autonomous driving (Shao et al., 2019), and code generation (Anthropic, 2024; Yang et al., 2025), driven by advances in deep RL (Mnih et al., 2015; Silver et al., 2017; OpenAI et al., 2019; Schrittwieser et al., 2019; Guo et al., 2025). As RL scales to more challenging domains with longer horizons, sparse and delayed rewards, and greater stochasticity, the long-standing *Credit Assignment Problem* (CAP) grows increasingly relevant: learning depends on successfully "distributing credit for success of a complex strategy among the many decisions that were involved" (Minsky, 1961). We argue that this credit distribution must be causal: the credit each action receives should reflect its causal effect on the total discounted reward.

However, locating the true causes of the rewards is challenging (Pignatelli et al., 2024). In Figure 1, we construct a simple MDP to reflect this intricate nature of CAP. Under a uniform random policy, the agent's first action determines which branch it follows. On the top branch, the last action determines the reward: one choice yields reward 1 with certainty ("Skill," red) while the other yields 0 (gray). On the bottom branch, the reward is 0 or 1 with equal probability regardless of the last action ("Luck," blue). Thus, a good credit assignment method should assign bigger positive credit to both the first and last actions on Skill trajectory than those on Luck, since changing either one would have altered the return distribution. On the Luck trajectory, the last action has no causal effect on the return and should receive zero credit. Surprisingly, as we show in Figure 1, none of the prominent existing methods produces the fully correct credit assignment (see Appendix G for a detailed discussion).

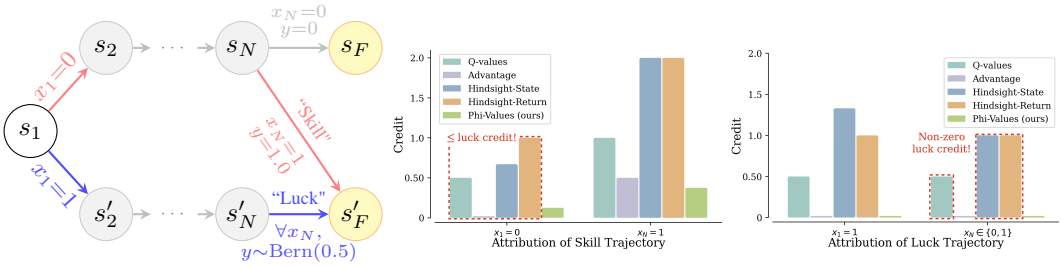

Figure 1: **Among all tested methods, only causal credit assignment correctly distinguishes Skill from Luck.** *Left:* A SkillLuck MDP under a uniform random policy. The first action selects the branch; on the top branch, $x_N{=}1$ yields reward 1 with certainty ("Skill," red) while $x_N{=}0$ yields 0 (gray). On the bottom branch, the reward is drawn from $\text{Bern}(0.5)$ regardless of x_N ("Luck," blue). *Center:* Only our causal credit assignment assigns more credit to the first action on the Skill trajectory ($x_1{=}0$) than on the Luck trajectory ($x_1{=}1$). *Right:* On the Luck trajectory, x_N has no causal effect on the reward. Only the advantage and our method correctly assign zero credit to x_N.

Humans, on the other hand, routinely perform credit assignment which does not seem to be challenging as it appears to machines. One possible reason is that we are endowed with causal reasoning capabilities that enable us to allocate credit counterfactually. Chess masters, for instance, replay key positions after every game, analyzing what would have happened under different moves. As Kasparov (2007) writes, "it is so important to question success as vigorously as you question failure." Structural causal models (SCMs) (Pearl, 2009; Bareinboim et al., 2022) formalize this type of counterfactual reasoning: given an observed trajectory, we can simulate the outcome under a different action at any step while holding all other factors fixed, isolating the causal effect of that decision. This type of counterfactual analysis lays the foundation for a fine-grained attribution for individual actions in each trajectory regardless of delayed effect, stochasticity and sparsity of the reward signals.

In this work, building on the Counterfactual Shapley Value (ϕ-value) framework of Lee et al. (2025), we distribute the total causal effect across actions using Shapley values (Shapley, 1953), which provide the unique allocation satisfying efficiency, symmetry, and null-player axioms. The resulting ϕ-MDP replaces the environment's reward signal with per-step ϕ-values, providing dense causal learning signals even when the original reward is delayed or sparse. We prove this redistribution preserves the optimal policy (Theorem 3.7) and derive a consistent estimator computing all T values from a single coalition sample in $O(T)$ time (Theorem 4.2). We integrate these into ϕ-PPO, which replaces per-step rewards with ϕ-values and uses Prioritized Trajectory Replay (PTR). PTR selects subtrajectories by causal importance, in contrast to PER (Schaul et al., 2015b) which prioritizes individual transitions by TD error. Our contributions are summarized as follows.

- **Credit Assignment via ϕ-values**. We show that credit assignment via ϕ-values is equivalent to a reward redistribution that preserves the optimal policy and spikes only on causal actions.
- **Tractable ϕ-value Estimation**. We propose a consistent ϕ-value estimator that achieves linear time complexity, with a tunable bias–variance tradeoff via λ-bootstrapping.
- **Efficient and Explainable Policy Optimization**. We propose ϕ-PPO and demonstrate its strong empirical performance in several challenging credit assignment testbeds.

2 Preliminaries

In this section, we establish foundational concepts to facilitate our discussion. Throughout the paper, we use uppercase letters to denote random variables (X), lowercase for their realizations (x) and bold letters for sets ($\mathbf{V}$).

2.1 Structural Causal Models and MDPs

Structural Causal Models (SCMs) lay the mathematical foundations for counterfactual reasoning needed to isolate causal contributions (Bareinboim et al., 2022).

Definition 2.1 (Structural Causal Model)**.** An SCM is $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ where $\mathbf{V}$ is endogenous variables, $\mathbf{U}$ is exogenous variables, $\mathcal{F} = \{f_i\}$ are structural equations $V_i = f_i(\mathbf{Pa}_i, \mathbf{U}_i)$, and $P(\mathbf{U})$ is exogenous distribution.

Under this definition, the counterfactual $V_{\mathbf{x}}(\mathbf{u})$ is the value of V when $\mathbf{X}$ is set to $\mathbf{x}$ while $\mathbf{U} = \mathbf{u}$. Then we cast MDPs as SCMs to ground the credit assignment discussion with causal semantics.

Definition 2.2 (MDP-SCM)**.** An MDP-SCM is an SCM with endogenous variables $\mathbf{V} = \{S_t, X_t, Y_t\}_{t=1}^{T} \cup \{Y\}$ (states, actions, per-step rewards, and outcome $Y = f_Y(Y_{1:T})$), mutually independent exogenous variables $\mathbf{U} = \{U_{S_t}, U_{X_t}, U_{Y_t}\}_{t=1}^{T}$, and structural equations defining system dynamics: $S_1 = f_S(U_{S_1}), S_{t+1} = f_S(S_t, X_t, U_{S_{t+1}}), X_t = \pi(S_t, U_{X_t}),\ Y_t = r(S_t, X_t, U_{Y_t})$.

We use $P(S)$ to denote the initial state distribution induced by U_{S_1} and the discounted total reward as the outcome $Y = \sum_t \gamma^{t-1} Y_t$. Value functions $V(s)$, Q-values $Q(s, x)$, and advantages $A(s, x)$ under MDP-SCM follow the standard RL definitions (Sutton & Barto, 2018, Sec. 3.1).

2.2 Policy Gradients

The goal of reinforcement learning is to find the policy that maximizes the cumulative reward (outcome), $\pi^* = \arg\max_\pi \mathbb{E}_{\tau \sim \mathcal{M}^\pi}[Y]$, where trajectory $\tau \sim \mathcal{M}^\pi$ is sampled from the environment MDP $\mathcal{M}^\pi$ under policy π. Optimizing this objective with respect to a parameterized policy π_θ leads to the basic form of policy gradients (Williams, 1992),

$$\mathbb{E}\Big[\sum_{t\geq 0} \gamma^t \nabla_\theta \log \pi(x_t|s_t)\big(\sum_{t'\geq t} \gamma^{t'-t} Y_{t'} - V(s_t)\big)\Big] \tag{1}$$

where $V(s_t) = \mathbb{E}\big[\sum_{t'\geq t} \gamma^{t'-t} Y_{t'} | s_t\big]$ is the value function. Here it serves as a baseline with respect to which the advantage of taking the specific action x_t instead of others is strengthened. Subtracting the baseline will not bias the gradient as long as the baseline is a function of state s_t. Modern policy gradient methods use generalized advantages (Schulman et al., 2016) and gradient clipping (Schulman et al., 2017) to reduce variance. In this work, we base our learning algorithm on proximal policy gradient (PPO) (Schulman et al., 2017), whose objective is to maximize the following,

$$\min\big(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\hat{A}_t\big) \tag{2}$$

where $\hat{A}_t = \sum_{\ell=0}^{T-t+1} (\gamma\lambda)^\ell \delta_{t+\ell},\ \delta_t = Y_t + \gamma V(s_{t+1}) - V(s_t)$ is the generalized advantage and $r_t(\theta) = \frac{\pi_\theta(x_t|s_t)}{\pi_{\theta_{\text{old}}}(x_t|s_t)}$ is the importance ratio between the current policy distribution and the previous policy collecting the data.

2.3 Counterfactual Shapley Values

Shapley values (Shapley et al., 1953) fairly distribute credit among players in cooperative games. Given players $\mathbf{X}$ and value function $f : 2^{\mathbf{X}} \to \mathbb{R}$, a *coalition* $\mathbf{Z} \subseteq \mathbf{X}$ is a subset of players, and $f(\mathbf{Z})$ measures their joint contribution. The Shapley value ϕ_t quantifies player X_t's fair share of the total value $f(\mathbf{X}) - f(\emptyset)$:

$$\phi_t = \sum_{\mathbf{Z}\subseteq\mathbf{X}\setminus\{X_t\}} \frac{|\mathbf{Z}|!(T - |\mathbf{Z}| - 1)!}{T!}\big[f(\mathbf{Z}\cup\{X_t\}) - f(\mathbf{Z})\big]. \tag{3}$$

Shapley values satisfy desirable properties (Shapley et al., 1953), including *efficiency*: $\sum_t \phi_t = f(\mathbf{X}) - f(\emptyset)$. For credit assignment, actions $\mathbf{X} = \{X_1, \ldots, X_T\}$ are players, and given trajectory $\tau = (s_{1:T}, x_{1:T}, y_{1:T})$ (the *evidence* for counterfactual reasoning), the Shapley value ϕ_t quantifies how much X_t contributed to outcome Y.

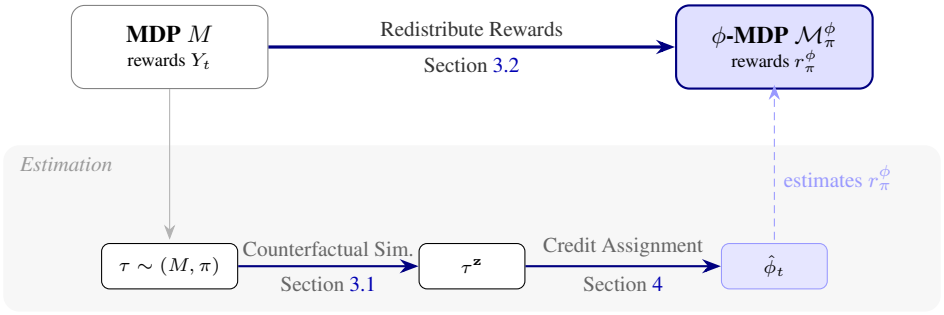

Figure 2: Our counterfactual Shapley values assigns causal credits to each action via reward redistribution. *Top:* The MDP M is transformed into ϕ-MDP $\mathcal{M}_\pi^\phi$ by redistributing rewards Y_t into $r_\pi^\phi(s, x) = \mathbb{E}[\phi_t \mid S_t{=}s, X_t{=}x]$. Both MDPs share the same optimal policy (Theorem 3.7). *Bottom:* The estimation pipeline samples a trajectory $\tau \sim \pi$ and a coalition mask $\mathbf{z} \sim Q^*$, then simulates the counterfactual trajectory $\tau^{\mathbf{z}}$ (Sec. 3.1). Action credits are estimated based on sampled counterfactual trajectories (Algo. 1) and used for policy optimization (Sec. 4).

Causal Contributions. We use counterfactual natural total effect (Lee et al., 2025) to measure a set of actions' causal contributions to the outcome. It is the expected difference between the observed outcome and a counterfactual outcome where a subset of actions are replaced. "Natural" means the comparison uses the posterior $P(\mathbf{U} \mid \tau)$ rather than the prior, ensuring the counterfactual is consistent with the observed trajectory:

$$f(\mathbf{Z}) = \mathrm{NTE}(\mathbf{Z}, Y \mid \tau) = \mathbb{E}_{\mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U} \mid \tau)} \big[Y(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \big], \tag{4}$$

where $Y(\mathbf{u})$ is the observed outcome and $Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})$ is the counterfactual with coalition actions replaced by baseline alternatives $\mathbf{Z}(\mathbf{u}')$ sampled from a baseline policy π_{base} (Sec. 3.1), while exogenous $\mathbf{u}$ is fixed.

Kernel Formulation. The standard Shapley formula sums over $T!$ permutations; the kernel formulation (Lundberg & Lee, 2017) sums over 2^T coalitions, enabling amortization (Section 4). Let $\mathbf{z} \in \{0, 1\}^T$ denote the coalition mask ($z_t = 1$ iff $X_t \in \mathbf{Z}$):

$$\phi_t = \sum_{\mathbf{z} \in \{0,1\}^T} \kappa_t(\mathbf{z}) \cdot f(\mathbf{z}), \tag{5}$$

where $|\mathbf{z}| = \sum_t z_t$ and $\kappa_t(\mathbf{z}) = \frac{1}{T\binom{T-1}{|\mathbf{z}|-1}}$ if $z_t = 1$ while $\kappa_t(\mathbf{z}) = -\frac{1}{T\binom{T-1}{|\mathbf{z}|}}$ if $z_t = 0$.

3 Counterfactual Credit Assignment

In the remainder of the text, we introduce our proposed causal solution to CAP. In Section 3, we formalize the credit assignment problem from a causal perspective, propose ϕ-values as a principled solution and its related calculations in the RL context. In Section 4, we develop efficient estimators for ϕ-values and incorporate them into policy learning. Figure 2 visualizes the conceptual structures.

First, we define the Credit Assignment Problem (CAP) as a causal inference problem: only actions that causally contribute to the outcome should receive credit, proportional to their contributions.

Definition 3.1 (Causal Credit Assignment, Informal)**.** A causal credit assignment function is a mapping ϕ that maps a trajectory, an action at a time step and a baseline policy to a real number denoting its causal contribution to the outcome observed in the trajectory,

$$\phi : \tau \times \mathcal{X} \times T \times \Pi \mapsto \mathbb{R}^T \tag{6}$$

where $\tau \sim \mathcal{M}_\pi$ is a trajectory sampled from the baseline policy and Π is the policy space. The causal assignment ϕ should satisfy the following desiderata,

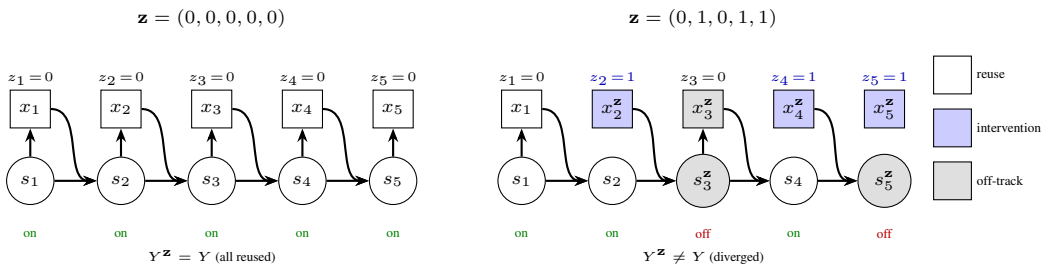

Figure 3: Counterfactual simulation ($T = 5$). *Left:* $\mathbf{z} = \mathbf{0}$, all values reused, $Y^{\mathbf{z}} = Y$. *Right:* $\mathbf{z} = (0, 1, 0, 1, 1)$; intervention at $t = 2$ causes divergence, returns on-track at $t = 4$ ($s_4^{\mathbf{z}} = s_4$ by chance), diverges again at $t = 5$.

- $\mathbf{D_1}$ **Causal Admissibility** : Non-cause actions should be assigned zero credit;
- $\mathbf{D_2}$ **Causal Power** : Actions causing the rewards should be assigned non-zero credit;
- $\mathbf{D_3}$ **Causal Normality** : Actions against more probable baseline policies get more credit;
- $\mathbf{D_4}$ **Causal Effect Scaling** : Given the same baseline policy, actions affecting the outcome more should be allocated more credit.

A complete formal version of the causal credit assignment desiderata is presented in App. C, Def. C.3. Here, we measure each action's causal contribution to the outcome as the counterfactual natural total effect (Eq. 4). Unlike discounted summation of rewards, the counterfactual natural total effect is the difference between the returns under baseline policy and returns had we acted differently given an already realized trajectory. Advantage function shares a similar idea but they cannot capture counterfactual outcomes with respect to specific trajectories or quantify an action's full potential under different action selections at other time steps.

Counterfactual natural total effect provides a fair source of credit pool as the basis for allocation and these desiderata are axioms embedded in human credit assignment systems, which a causal credit assignment function should satisfy. $\mathbf{D_1}, \mathbf{D_2}$ together guarantee that the credit assignment function generates both sufficient and necessary allocations. That is, actions get credit if and only if they cause the outcome. $\mathbf{D_3}, \mathbf{D_4}$ require that the scale of the allocated credit should reflect an action's relative contribution compared against a baseline policy. We show that counterfactual Shapley values (ϕ-values, Eq. 5) calculated from counterfactual natural total effect satisfy the above desiderata (Lee et al., 2025) and is a desirable credit assignment method.

Theorem 3.2 (ϕ-Values are Causal)**.** *ϕ-values satisfy the causal credit assignment criteria, $\mathbf{D_1} - \mathbf{D_4}$.*

Next, we present the details in calculating ϕ-values and ϕ-MDP, an MDP that redistributes rewards according to Counterfactual Shapley values, as our solution to the Causal Credit Assignment Problem.

3.1 Counterfactual Simulation

To compute Shapley values ϕ_t, we need a causal contribution function $f : 2^{\mathbf{X}} \to \mathbb{R}$ measuring $\mathbf{X}$'s contribution under a given coalition $\mathbf{z}$. The NTE from Section 2.3 averages over both posterior and prior exogenous variables; here, we condition on the observed trajectory, giving the *conditional* causal contribution: $f(\mathbf{z}) = Y - Y_{\sigma_{\mathbf{z}}}$, where Y is the observed return and $Y_{\sigma_{\mathbf{z}}}$ is the counterfactual return under intervention $\sigma_{\mathbf{z}}$ which specifies counterfactual actions based on coalition membership $\mathbf{z} \in \{0, 1\}^T$ and trajectory state:

$$\sigma_{\mathbf{z}}(t) = \begin{cases} x_t & z_t = 0,\ s_t^{\mathbf{z}} = s_t \\ x \sim \pi(\cdot \mid s_t^{\mathbf{z}}) & z_t = 0,\ s_t^{\mathbf{z}} \neq s_t \\ x \sim \pi_{\text{base}}(\cdot \mid s_t^{\mathbf{z}}) & z_t = 1 \end{cases} \tag{7}$$

where $s_t^{\mathbf{z}}$ is the counterfactual state at time t. The baseline π_{base} defines the "default" actions for intervened time steps. Options include: uniform random ($\pi_{\text{base}}(x \mid s) = 1/|\mathcal{A}|$), measuring

contribution relative to random behavior; a fixed reference policy π_0, measuring improvement over π_0; or self-baseline ($\pi_{\text{base}} = \pi$), isolating whether the *specific* action x_t matters versus what the same policy typically does. We use self-baseline throughout training (Section 5.3 illustrates the effect of alternative baselines on attributions). The intervention design rests on the following assumption.

Assumption 3.3 (Counterfactual Independence)**.** Exogenous noise is independent across steps: $(S_{t+1})_{S_t,X_t} \perp (S_{t+1})_{S'_t,X'_t}$ for $(S_t, X_t) \neq (S'_t, X'_t)$, and $(X_t)_{S_t} \perp (X_t)_{S'_t}$ for $S_t \neq S'_t$.

This assumption does not constrain the MDP, only its SCM representation. Since any distribution can be sampled via inverse CDF with independent uniform noise, any MDP admits an SCM with independent exogenous variables and the same optimal policy; the assumption selects this representation. The consequence is simple: *when parents match the observed trajectory, the observed value is the counterfactual; when parents differ, we resample*. For actions (parent S_t): when $z_t = 0$ and $s_t^{\mathbf{z}} = s_t$, the counterfactual equals the observed x_t; when $s_t^{\mathbf{z}} \neq s_t$, we resample from $\pi(\cdot \mid s_t^{\mathbf{z}})$. When $z_t = 1$, we resample from π_{base}. For transitions and rewards (parents (S_t, X_t)): when $(s_t^{\mathbf{z}}, x_t^{\mathbf{z}}) = (s_t, x_t)$, we reuse (s_{t+1}, y_t); otherwise, we resample. See Fig. 3 for a visualization of the counterfactual simulation process.

The full pseudo-code is provided in Algo. 2 in appendix Sec. A; Section 4 makes exact computation tractable via coalition sampling.

3.2 Reward Redistribution

With the counterfactual Shapley values, we can define a new rewarding system based on ϕ_t while keeping the system dynamics unchanged. This leads to the ϕ-MDP, a reward-redistributed counterpart of the original MDP.

Definition 3.4 (ϕ-MDP)**.** Given an MDP-SCM $\mathcal{M}$ and baseline policy π_{base}, the *ϕ-MDP* $\mathcal{M}_\pi^\phi$ replaces the reward function r with the *policy-induced Shapley reward*:

$$r_\pi^\phi(s, x) = \mathbb{E}_{\tau\sim\pi}[\phi_t \mid S_t = s, X_t = x], \tag{8}$$

where the expectation is over trajectories τ sampled from π, conditioned on visiting $(S_t, X_t) = (s, x)$.

Lemma 3.5 (Gradient Equivalence)**.** *For any baseline policy π_{base}:* $\mathbb{E}_\pi[\sum_t \phi_t \cdot \nabla \log \pi(X_t \mid S_t)] = \mathbb{E}_\pi[Y \cdot \sum_t \nabla \log \pi(X_t \mid S_t)]$.

Remark 3.6 (Independence under self-baseline). The proof uses $Y' \perp X_t \mid S_t$ where $Y' = Y_{\sigma_\mathbf{1}}$ is the full-baseline outcome. Under self-baseline, Y' is computed on a *counterfactual trajectory* using fresh action samples $X'_{1:T} \sim \pi(\cdot \mid S^{\mathbf{1}}_{1:T})$, where $S^{\mathbf{1}}_{1:T}$ denotes the counterfactual state sequence. Crucially, Y' carries no information about which specific action X_t was taken at the observed state S_t: the counterfactual uses fresh action samples independent of X_t. The score function identity $\mathbb{E}_{X_t\sim\pi}[\nabla \log \pi(X_t \mid S_t)] = 0$ then gives $\mathbb{E}[Y'\nabla \log \pi_t] = 0$.

Gradient equivalence means policy gradient algorithms receive the same update direction in expectation whether maximizing returns from MDP $\mathcal{M}$ or its dual ϕ-MDP.

Theorem 3.7 (Optimal Policy Equivalence)**.** *ϕ-MDP preserves optimal policies as the original MDP.*

More specifically, $\arg\max_\theta \mathbb{E}_{\pi_\theta}[\sum_t \phi_t] = \arg\max_\theta \mathbb{E}_{\pi_\theta}[Y]$, so ϕ-MDP is a drop-in replacement: any policy gradient algorithm applied to the redistributed rewards converges to the same optimal policy. One caveat: the ϕ-MDP is *undiscounted* ($\gamma_\phi = 1$). Since the NTE game outcome $Y = \sum_t \gamma^{t-1} r_t$ already incorporates temporal discounting, applying γ to ϕ_t would double-discount. The undiscounted ϕ-return $\sum_t \phi_t$ equals $Y - Y^{\mathbf{1}}$ by Shapley efficiency.

3.3 How Causal Credit Assignment Helps Learning

We have shown that the ϕ-MDP provides a principled credit assignment while preserving the optimal policy, but does it also improve learning efficiency? As an overview, Fig. 4 illustrates each mechanism

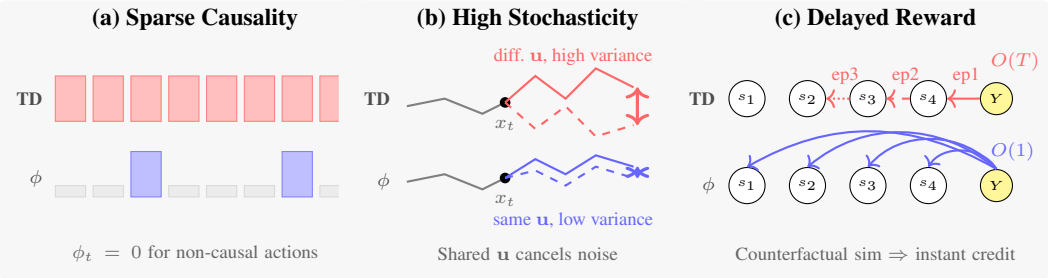

Figure 4: Three dimensions where ϕ-redistribution reduces variance. *(a) Sparse causality.* TD updates all T actions; ϕ updates only the k' causal actions, reducing variance by factor k'/T. *(b) High stochasticity.* TD compares returns under different noise realizations (high variance); ϕ uses shared exogenous noise $\mathbf{u}$, so the noise cancels in the difference. *(c) Delayed reward.* TD(0) propagates credit one step per episode, requiring $O(T)$ episodes; ϕ assigns credit to all steps in one episode via counterfactual simulation.

visually. Below, we present the formal conditions our method helps policy optimization under the assumption of a finite second moment.

Assumption 3.8 (Finite Second Moment)**.** $\forall t, \mathbb{E}[Y_t^2] \leq \sigma^2$.

Proposition 3.9 (Sparse Causality)**.** *Let* $k' = |\{t : \phi_t \neq 0\}|$, $\sigma_\phi^2 = \max_t \mathrm{Var}[\phi_t \nabla \log \pi_t]$, *and* $\bar{\rho} \in [0, 1]$ *the maximal pairwise correlation among causal terms. Then* $\mathrm{Var}[\hat{g}^\phi] \leq k'(1 + (k'-1)\bar{\rho})\, \sigma_\phi^2$.

When few actions influence the outcome ($k' \ll T$), ϕ-values assign zero credit to non-causal actions, so only k' terms contribute to gradient variance. The cross-correlation $\bar{\rho}$ is small in practice: the score function identity $\mathbb{E}[\nabla \log \pi_s \mid S_s] = 0$ decorrelates terms at distinct timesteps, and $\bar{\rho} = 0$ exactly in deterministic environments (Appendix C).

Proposition 3.10 (High Stochasticity)**.** *Let* $\mathbf{z}' = \mathbf{z}$ *with* $z_t = 1$ *and* $\rho_Y = \mathrm{Corr}[Y^{\mathbf{z}'}, Y^{\mathbf{z}}]$ *under shared exogenous randomness. Then* $\mathrm{Var}[\phi_t] \propto (1 - \rho_Y)$.

Proposition 3.10 holds in expectation over coalitions $\mathbf{z} \sim Q^*$ and trajectories. Define the *effective coupling* $\bar{\rho}_Y = \mathbb{E}_{\mathbf{z} \sim Q^*}[\mathrm{Corr}[Y^{\mathbf{z}'}, Y^{\mathbf{z}}]]$, averaging over the optimal coalition distribution. When a counterfactual trajectory diverges (off-track), fresh noise is sampled, reducing $\mathrm{Corr}[Y^{\mathbf{z}'}, Y^{\mathbf{z}}]$ for that sample. The effective coupling $\bar{\rho}_Y$ is high when: (1) coalitions are small; (2) transitions are nearly deterministic; or (3) rewards depend on shared exogenous factors (unobserved confounders). Variance reduction is proportional to $(1 - \bar{\rho}_Y)$.

Proposition 3.11 (Direct Propagation)**.** *Tabular TD(0) with terminal-only reward requires* $O(T)$ *episodes for credit to reach the initial state.* ϕ*-redistribution provides credit in* $O(1)$ *episodes via* $O(TM_{\delta,\epsilon})$ *counterfactual simulations.*

Remark 3.12 (Propagation Complexity). For terminal-only reward: TD(0) requires $O(T)$ episodes and $O(T^2)$ updates; ϕ-redistribution requires $O(1)$ episodes and $O(TM_{\delta,\epsilon})$ counterfactual simulations.

This shows that ϕ-values avoid the delayed-effect and inefficient propagation problem of TD(0) methods. Though n-step returns, TD(λ) and the use of replay buffers mitigate this problem (Daley, 2025, Sec. 1.2-1.3), their core credit assumption rests on temporal proximity (Pignatelli et al., 2024). As the trajectory rolls out, reward in a distant step is deemed to be exponentially less likely caused by a given action (by the discounting factor). This clearly does not reflect an action's true causal contribution as we have demonstrated in Fig. 1. And we will also see in experiment Sec. 5 that even with all the modern DRL techniques (Schulman et al., 2017; Schaul et al., 2016), existing algorithms still cannot solve those seemingly naive tabular MDPs due to the implicit time proximity assumption in credit assignment. And the most important implication of Prop. 3.11 is that our proposed ϕ-value can allocate credit in a lossless and faithful way w.r.t each action's actual causal contribution.

To sum, the "sweet spot" occurs when all three dimensions favor ϕ-values: sparse causality ($k' \ll T$), high stochasticity ($\rho_Y \approx 1$), and delayed reward. When both sparsity and stochasticity benefits apply,

the variance reduction factors multiply: sparsity reduces contributing terms from T to k', while coupled comparison reduces each term's variance by $(1-\rho_Y)$, giving combined ratio $\frac{k'}{T}(1-\rho_Y)$. On the other hand, there are cases when ϕ-methods offer marginal advantage in terms of learning efficiency other than explanatory benefits over the standard TD methods when: (1) rewards are dense ($k' \approx T$); (2) environment is deterministic (ρ_Y is high); or (3) actions are rewarded immediately.

4 Estimating Counterfactual credit for Policy Optimization

This section develops efficient estimators for counterfactual Shapley values with $O(TM_{\delta,\epsilon})$ total complexity, where $M_{\delta,\epsilon} = \sigma^2/((1-\gamma)^2\delta\epsilon^2)$, and integrates them into policy optimization as ϕ-PPO.

Importance Sampling Equivalence. Each coalition evaluation $f(\mathbf{z})$ requires $O(T)$ simulation time (Algo. 2). Computing T Shapley values independently would require M samples per value, yielding $O(T^2M)$ total cost. The kernel formulation (Eq. 5) amortizes this: a single coalition evaluation updates all T values via different kernel weights $\kappa_t(\mathbf{z})$, reducing cost to $O(TM)$.

Definition 4.1 (Coalition Estimator)**.** Given samples $\mathbf{z}^{(m)} \sim Q$ and NTE $f(\mathbf{z}) = Y - Y^{\mathbf{z}}$, where $Y^{\mathbf{z}} \equiv Y_{\sigma_{\mathbf{z}}}$ is the counterfactual return, the *coalition estimator* is $\hat{\phi}_t = \frac{1}{M}\sum_{m=1}^{M} \frac{\kappa_t(\mathbf{z}^{(m)})}{Q(\mathbf{z}^{(m)})} \cdot f(\mathbf{z}^{(m)})$.

Under a mild assumption, we can derive the variance-reduction optimal proposal distribution $Q^*(\mathbf{z})$ for coalition sampling.

Theorem 4.2 (Optimal Proposal)**.** *Under Assumption 3.8, the variance-minimizing proposal is* $Q^*(\mathbf{z}) = q_k^*/\binom{T}{k}$ *where* $k = |\mathbf{z}|$ *and* $q_k^* \propto \sqrt{c_k}$ *with:*

$$c_k = \begin{cases} \frac{1}{T}\left(\frac{1}{k} + \frac{1}{T-k}\right) & k \in \{1, \ldots, T-1\} \\ \frac{1}{T^2} & k = T \end{cases} \tag{9}$$

Under Q^*, the estimator achieves (ϵ, δ)-accuracy with $M = O(M_{\delta,\epsilon})$ samples.

The optimal $q_k^* \propto \sqrt{c_k}$ concentrates on extreme coalition sizes ($k \approx 1$ or $k \approx T-1$) where kernel weights $|\kappa_t|$ are largest. In contrast to uniform sampling, which induces variance exponential in T (extreme sizes have probability $O(2^{-T})$), the optimal proposal achieves $O(1)$ second moment. Variance is low because Y and $Y^{\mathbf{z}}$ share exogenous noise via common random numbers (Glasserman & Yao, 1992): the difference $f(\mathbf{z}) = Y - Y^{\mathbf{z}}$ isolates the effect of action choices while canceling environmental stochasticity. In practice, we also use antithetic sampling (Covert & Lee, 2021): each coalition $\mathbf{z}$ is paired with its complement $\mathbf{1} - \mathbf{z}$ to further reduce variance.

Bootstrapped Estimation. The basic estimator Def. (4.1) requires full trajectory simulation for each coalition. Bootstrapping with a learned value function $V \approx V_\pi$ reduces variance by replacing future counterfactual rewards with value estimates. We use TD-λ (Sutton & Barto, 2018, Sec. 12.2) that interpolates between Monte Carlo and bootstrapped estimates:

$$G_t^{\mathbf{z},\lambda} = y_t^{\mathbf{z}} + \gamma[(1-\lambda)V(s_{t+1}^{\mathbf{z}}) + \lambda G_{t+1}^{\mathbf{z},\lambda}], \tag{10}$$

where $\lambda = 1$ yields pure Monte Carlo ($G_t^{\mathbf{z},1} = Y_t^{\mathbf{z}}$) and $\lambda = 0$ yields one-step TD ($G_t^{\mathbf{z},0} = y_t^{\mathbf{z}} + \gamma V(s_{t+1}^{\mathbf{z}})$). Then we define λ-bootstrapped NTE to decompose $Y^{\mathbf{z}}$ into exact rewards before step t and λ-returns from t onwards. Let $R_t^{\mathbf{z}} = \sum_{s<t}\gamma^{s-1}y_s^{\mathbf{z}}$ denote the cumulative counterfactual reward before step t. The per-step treatment effect is

$$f_t^{\mathbf{z}} = Y - (R_t^{\mathbf{z}} + \gamma^{t-1}G_t^{\mathbf{z},\lambda}). \tag{11}$$

When $\lambda = 1$, we have $R_t^{\mathbf{z}} + \gamma^{t-1}G_t^{\mathbf{z},1} = Y^{\mathbf{z}}$, so $f_t^{\mathbf{z}} = f(\mathbf{z})$ recovers the exact NTE. For $\lambda < 1$, the identity $\mathbb{E}[G_t^{\mathbf{z},\lambda} \mid s_t^{\mathbf{z}}] = V_\pi(s_t^{\mathbf{z}}) + O(\epsilon_V)$ implies $\mathbb{E}[f_t^{\mathbf{z}}] = f(\mathbf{z}) + O(\epsilon_V)$ where $\epsilon_V = \|V - V_\pi\|_\infty$.

Algorithm 1 L3EST: Counterfactual Shapley Estimation

Input: Trajectory τ of length T; value network V_ψ; coalition samples M
Output: Attributions $\hat{\phi}_{1:T}$, TD errors $\delta_{1:T}$, $\delta^{\mathbf{z}}_{1:T,1:M}$

```
for m = 1 to M do
    Sample coalition z^(m) ~ Q* for T players {Thm. 4.2}
    Simulate counterfactual τ^z^(m) via CTFSIM {Algo. 2}
    for t = 1 to T do
        Compute δ_t^z^(m) ← G_t^z^(m),λ − V(s_t^z^(m))
    end for
end for
Compute TD-error δ_{1:T} from observed trajectory
Compute φ̂_{1:T} by averaging over M samples {Def. 4.3}
return φ̂_{1:T}, δ_{1:T}, δ^z_{1:T,1:M}
```

Definition 4.3 (λ-Bootstrapped Estimator)**.** Given the per-step treatment effect $f^{\mathbf{z}}_t$ Eq. (11), the *λ-bootstrapped estimator* is $\hat{\phi}^\lambda_t = \frac{1}{M}\sum_{m=1}^{M} \frac{\kappa_t(\mathbf{z}^{(m)})}{Q^*(\mathbf{z}^{(m)})} \cdot f_t^{\mathbf{z}^{(m)}}$.

By incorporating the contraction rate of λ-return (Tsitsiklis & Van Roy, 1997), we derive the following bias bound for our ϕ-value estimator.

Theorem 4.4 (Bias Bound for $\hat{\phi}^\lambda$)**.** *Under Assumptions 3.3, 3.8:* $|\mathbb{E}[\hat{\phi}^\lambda_t] - \phi_t| \leq \frac{2\gamma(1-\lambda)}{1-\gamma\lambda}\|V - V_\pi\|_\infty$.

The estimator is unbiased when $\lambda = 1$ (pure MC) or $\epsilon_V = 0$ (perfect value function). Decreasing λ trades bias for variance by truncating the effective horizon (Sutton & Barto, 2018, Ch. 12). Algo. 1 summarizes the overall procedure for counterfactual Shapley estimation.

ϕ-PPO. ϕ-PPO integrates Shapley credit assignment into PPO (Schulman et al., 2017) via two mechanisms. First, the ϕ-MDP $\mathcal{M}^\phi_\pi$ (Definition 3.4) reformulates credit as per-step reward $r^\phi_\pi(s,x) = \mathbb{E}[\phi_t \mid S_t = s, X_t = x]$, enabling standard actor-critic methods. Second, Prioritized Trajectory Replay (PTR) focuses updates on high-$|\phi|$ subtrajectories rather than individual transitions as in PER (Schaul et al., 2016). See Algo. 3 and Sec. B in appendix for the details. In each iteration, ϕ-PPO samples sub-trajectories via PTR, computes ϕ-estimates, and updates both values and policy networks. Advantages are normalized by batch standard deviation (not z-scored, to preserve $\phi = 0$ as the no-effect baseline).

We admit that Algo. 1 requires $O(TM)$ value network evaluations per trajectory (one per counterfactual state per coalition sample) while standard PPO requires $O(T)$ evaluations. But this overhead is acceptable when: (1) value inference is cheaper than environment simulation for complex environments; (2) the M evaluations can be batched for GPU efficiency. In practice, we use $M = 1$ coalition sample per trajectory, matching standard actor-critic overhead while still benefiting from counterfactual variance reduction.

5 Experiments

We evaluate ϕ-PPO on two training benchmarks: SkillLuck (Sec. 5.1) and Combinatorial Lock (Sec. 5.2), illustrate ϕ-value attributions on DoorKey (Sec. 5.3), and provide a controlled ablation on the Fork MDP in Appendix E.1. Throughout, a run *succeeds* iff $P(\text{optimal}) > 0.9$ at all causal states for 200 consecutive episodes. We compare all methods on total environment steps: ϕ-PPO uses $(1+M)T$ steps per episode (1 observed trajectory plus M counterfactual rollouts); all baselines use T steps per episode. No method receives extra simulator access (full hyperparameters in Table 3). We compare against PPO, REINFORCE, and eight prior credit assignment methods: HCA-state, HCA-return (Harutyunyan et al., 2019b), CCA (Mesnard et al., 2021), CoCoA (Meulemans et al., 2023),

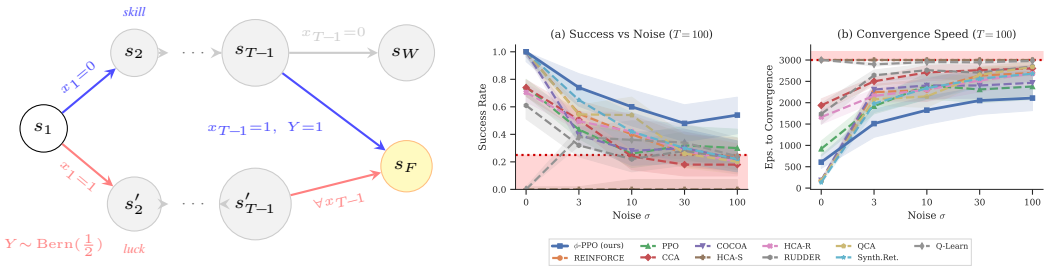

Figure 5: **Skill vs. Luck results** (T=100, γ=0.999, M=1, n=50 seeds per condition). *Left:* MDP structure; the first action selects skill or luck, and only the final action on the skill branch affects the reward. *(a)* Success rate with Wilson 95% CIs across noise levels $\sigma \in \{0, 3, 10, 30, 100\}$. *(b)* Mean episodes to convergence with bootstrap 95% CIs; failures imputed at the budget limit (3,000 episodes).

QCA (Mesnard et al., 2023), RUDDER (Arjona-Medina et al., 2019), Synthetic Returns (Raposo et al., 2021), and H-DICE (Velu et al., 2024).

Q1: Do ϕ-values assign credit accurately, and does this translate into a learning advantage?
Q2: Can ϕ-PPO learn the optimal policy under delayed, sparse, and highly stochastic rewards?
Q3: How sample efficient is ϕ-PPO compared to vanilla PPO and prior credit assignment methods?

5.1 Skill vs. Luck

The SkillLuck MDP separates skill-based from luck-based rewards to test whether correct causal attributions translate into learning (**Q1**, **Q2**, **Q3**). The agent chooses from $\mathcal{A} = \{0, 1\}$ at each of T=100 steps (γ=0.999, M=1). Every reward includes i.i.d. noise $\mathcal{N}(0, \sigma^2)$; we sweep $\sigma \in \{0, 3\}$ to test robustness to stochasticity. The first action x_1 selects one of two branches. Actions at intermediate steps ($2 \leq t \leq T-2$) do not affect the trajectory or reward. On the *skill* branch, the final action determines the reward: x_{T-1}=1 yields reward 1, while x_{T-1}=0 yields 0. On the *luck* branch, the agent receives a reward of $\text{Bern}(0.5) \cdot \gamma^{T-3}$ at step 2, regardless of any action. The γ^{T-3} scaling ensures both branches yield returns of comparable magnitude. Under the optimal final action, the skill branch has expected discounted return γ^{T-2} while the luck branch has $0.5\gamma^{T-2}$, so the optimal policy chooses skill. On the skill branch, two actions are causal: x_1 and x_{T-1}. On the luck branch, only x_1 is causal because it determines the branch; all subsequent actions have zero causal effect.

Predictions. ϕ-values should assign zero credit to x_{T-1} on luck (**D1**) and more credit to x_1 on skill than luck (**D4**); Figure 1 confirms this. HCA fails because skill and luck branches share the same terminal observation. At T=100, methods that cannot propagate delayed credit default to the luck branch.

Results (Figure 5). At σ=0, several baselines match or exceed ϕ-PPO because the noiseless setting poses no variance challenge; ϕ-PPO's slight shortfall reflects M=1 approximation noise. HCA fails as predicted. At σ=3, the ranking reverses: ϕ-PPO outperforms all baselines because shared noise cancels in the counterfactual difference (Proposition 3.10), while all other methods degrade. REINFORCE and PPO both fail, consistent with predictions. Among successful runs (right panel), ϕ-PPO also converges fastest.

5.2 Combinatorial Lock

The Combinatorial Lock is a T-step chain with tempting immediate rewards that tests delayed credit assignment (**Q2**, **Q3**). The agent chooses from $\mathcal{A} = \{0, 1\}$ at each position $p \in \{0, \ldots, T-1\}$: action 0 (CORRECT) advances to p+1 with zero immediate reward; action 1 (SHORTCUT) terminates

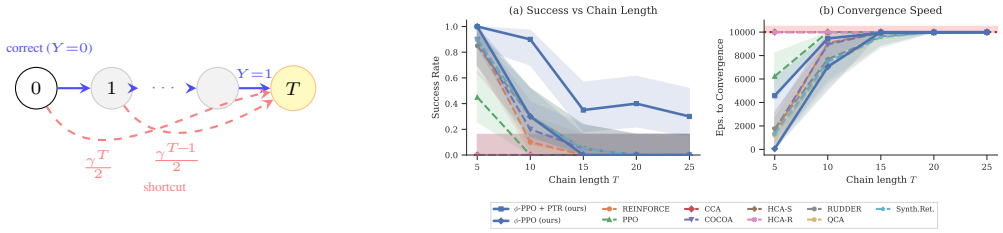

Figure 6: **Combinatorial Lock results** ($T \in \{5, 10, 15, 20, 25\}$, n=20 seeds, each seeded with one optimal trajectory). *Left:* MDP structure; CORRECT advances one step (zero reward) until the final transition yields Y=1 (discounted return γ^{T-1}, twice any shortcut's $\gamma^T/2$), while SHORTCUT from position p terminates with $\gamma^{T-p}/2$. *(a)* Success rate with Wilson 95% CIs. *(b)* Mean episodes to convergence with bootstrap 95% CIs; failures imputed at the budget limit (10,000 episodes). Solid lines are ϕ-PPO variants; dashed lines are baselines.

with immediate reward $\gamma^{T-p}/2$. Reaching position T via T consecutive CORRECT actions yields reward $+1$ (discounted return γ^{T-1}). The ratio of shortcut to optimal return is $\gamma/2 < 1$ at every position, so the shortcut is always suboptimal despite positive immediate reward. We sweep $T \in \{5, 10, 15, 20, 25\}$ with $\gamma = 0.99$ and M=16 coalition samples. Every run begins with one optimal trajectory in the replay buffer so that all methods start from the same initial information. This isolates *exploitation* (can the method propagate credit from a known success?) from exploration, which becomes vanishingly unlikely as T grows.

Predictions. Every action along the optimal path is causal ($k'/T = 1$), so the sparse causality benefit (Proposition 3.9) does not apply. Combinatorial Lock isolates the delayed reward benefit (Proposition 3.11): (1) ϕ-PPO should propagate credit from the terminal reward to position 0 in $O(1)$ episodes by evaluating full counterfactual returns; (2) TD-based methods require $O(T)$ bootstrap steps, and the shortcut temptation should lock the policy before credit reaches position 0, causing baseline success to collapse as T grows.

Results (Figure 6). Both predictions are confirmed. In the left panel, ϕ-PPO + PTR (solid blue) is the only method that scales beyond T=10; every baseline (dashed) collapses by T=15. The right panel confirms that ϕ-PPO + PTR converges at a roughly constant episode count across chain lengths, consistent with $O(1)$ propagation, while baselines flatline at the budget ceiling. Without PTR (solid orange), the seed trajectory is diluted by failures; PPO + PTR/PER (dashed) also fail, confirming that neither prioritized replay nor ϕ-values alone suffice—both are necessary. Performance degrades at large T because random coalition swaps are exponentially unlikely to preserve the full optimal path, so most counterfactual rollouts return zero and estimator variance increases.

5.3 DoorKey

We compute ϕ-value attributions along an optimal trajectory in the 5×5 DoorKey environment from MiniGrid (Chevalier-Boisvert et al., 2018) (**Q1**). The agent selects from 7 actions (navigation, pickup, toggle, etc.) across $|\mathcal{S}| = 400$ states (5×5 grid × 4 orientations × 2 key states × 2 door states). The optimal trajectory reaches the goal in 13 steps.

Baseline choice determines the causal story. Under a uniform random baseline (top right), ϕ_t is nearly uniform because DoorKey is a serial dependency chain in which every step must be correct for success. Values taper after the door toggle (t=8) because a random agent increasingly reaches the goal by chance from nearby states. Under a navigation-only baseline (bottom right), the causal story changes entirely. The navigator already takes the correct action at post-door states, so those actions receive $\phi_t \approx 0$ (D1, causal admissibility). Pickup, toggle, and first-room navigation steps that the navigator cannot replicate receive nonzero ϕ_t (D2, causal power), splitting the total credit

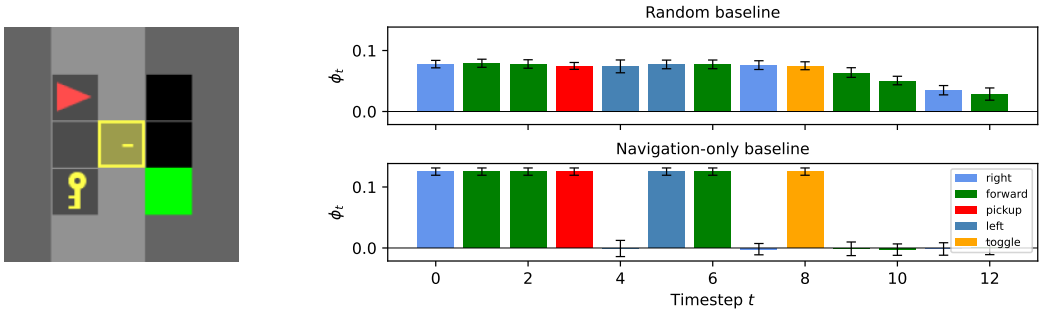

Figure 7: **DoorKey attributions** (M=4096, γ=0.99, 20 seeds). *Left:* the agent (red) must collect the key, unlock the door, and reach the goal (green). *Top right:* uniform random baseline. *Bottom right:* navigation-only baseline that follows the shortest path ignoring walls.

equally. ϕ-values thus measure credit *relative to the baseline policy*, not in absolute terms. Notably, the key pickup (t=3) receives comparable ϕ_t to the door toggle (t=8) under both baselines, despite zero immediate reward; its value reflects the downstream effect of enabling the toggle five steps later, confirming that ϕ-values propagate credit through dependent subgoal chains (Proposition 3.11).

Across all benchmarks, ϕ-PPO converges in fewer episodes than every baseline. At M=1 (SkillLuck), the 2× per-episode cost is offset by ≥4× fewer episodes; at M=16 (Lock), ϕ-PPO wins on episode efficiency but not total simulation budget, motivating future variance reduction work.

6 Conclusion

We introduced Counterfactual Shapley Credit Assignment, a principled framework that addresses the fundamental temporal credit assignment problem by grounding reward attribution in causal theory via Structural Causal Models. By utilizing the Counterfactual Shapley Value (ϕ-value), we isolate an agent's true causal contribution from environmental stochasticity and spurious correlations, inducing a dual ϕ-MDP that preserves the original optimal policy while providing significantly clearer learning signals. Our framework specifically addresses the challenges of high stochasticity, sparse causality, and delayed rewards through a computationally efficient linear-time estimator and the resulting ϕ-PPO algorithm. Empirical results across challenging benchmarks demonstrate that ϕ-PPO achieves superior sample efficiency and convergence compared to existing baselines, offering a scalable and theoretically sound path toward more efficient and explainable reinforcement learning.

Acknowledgments

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www.anthropic.com/news/claude-3-family.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1806.07857.

Ilze Amanda Auzina, Joschka Strüber, Sergio Hernández-Gutiérrez, Shashwat Goel, Ameya Prabhu, and Matthias Bethge. Intrinsic credit assignment for long horizon interaction, 2026. URL https://arxiv.org/abs/2202.06793.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pp. 507–556. Association for Computing Machinery and Morgan & Claypool, 2022. ISBN 9781450395861. URL https://dl.acm.org/doi/book/10.1145/3501714.

Sander Beckers. The counterfactual ness definition of causation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6210–6217, May 2021. DOI: 10.1609/aaai.v35i7.16772. URL https://ojs.aaai.org/index.php/AAAI/article/view/16772.

Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, 195(2):835–862, 2018. DOI: 10.1007/s11229-016-1247-1.

Michael Chang, Sid Kaushik, Sergey Levine, and Tom Griffiths. Modularity in reinforcement learning via algorithmic independence in credit assignment. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1452–1462. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/chang21b.html.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 15084–15097, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for gymnasium, 2018. URL https://github.com/Farama-Foundation/Minigrid.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *International conference on artificial intelligence and statistics*, pp. 3457–3465. PMLR, 2021.

Brett Daley. *Multistep Credit Assignment in Deep Reinforcement Learning*. PhD thesis, University of Alberta, Edmonton, AB, Canada, 2025. Supervisor: Marlos C. Machado.

Haoyou Deng, Keyu Yan, Chaojie Mao, Xiang Wang, Yu Liu, Changxin Gao, and Nong Sang. DenseGRPO: From sparse to dense reward for flow matching model alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nIwFge9nW0.

Johan Ferret, Raphael Marinier, Matthieu Geist, and Olivier Pietquin. Self-attentional credit assignment for transfer in reinforcement learning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2655–2661. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Manage. Sci.*, 38(6):884–908, June 1992. ISSN 0025-1909.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li,

H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. DOI: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Rémi Munos. Hindsight credit assignment. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12467–12476, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight credit assignment. *Advances in neural information processing systems*, 32, 2019b.

Garry Kasparov. *How Life Imitates Chess: Making the Right Moves, from the Board to the Boardroom*. Bloomsbury USA, 2007. ISBN 978-1596913875.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kai-Zhan Lee, Drago Plecko, and Elias Bareinboim. Causal explanations through counterfactual variable attributions. Technical Report R-135, Causal Artificial Intelligence Lab, Columbia University, May 2025. URL https://causalai.net/r135.pdf.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning.

In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Thomas Mesnard, Wenqi Chen, Alaa Saade, Yunhao Tang, Mark Rowland, Theophane Weber, Clare Lyle, Audrunas Gruslys, Michal Valko, Will Dabney, Georg Ostrovski, Eric Moulines, and Remi Munos. Quantile credit assignment. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24517–24531. PMLR, 23–29 Jul 2023.

Alexander Meulemans, Simon Schug, Seijin Kobayashi, Nathaniel Daw, and Greg Wayne. Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=yvqqkOn9Pi.

Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49:8–30, 1961.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. DOI: 10.1038/nature14236.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand. arxiv, 2019. URL http://arxiv.org/abs/1910.07113.

Hsiao-Ru Pan and Bernhard Schölkopf. Skill or luck? return decomposition via advantage functions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZFMiHfZwIf.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. DOI: 10.1017/CBO9780511803161.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bNtr6SLgZf. Survey Certification.

Aditya A Ramesh, Jiamin He, Jürgen Schmidhuber, and Martha White. Improving reward-based hindsight credit assignment. In *European Workshop on Reinforcement Learning (EWRL)*, 2025.

David Raposo, Sam Ritter, Adam Santoro, Greg Wayne, Theophane Weber, Matt Botvinick, Hado van Hasselt, and Francis Song. Synthetic returns for long-term credit assignment, 2021. URL https://arxiv.org/abs/2102.12425.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1312–1320, Lille, France, 07–09 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v37/schaul15.html.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015b.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.05952.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2019. ISSN 1476-4687. DOI: 10.1038/s41586-020-03051-4. URL https://www.nature.com/articles/s41586-020-03051-4.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1506.02438.

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

Kun Shao, Zhentao Tang, Yuanheng Zhu, Nannan Li, and Dongbin Zhao. A Survey of Deep Reinforcement Learning in Video Games. arxiv, 2019. URL http://arxiv.org/abs/1912.10944.

Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton, 1953.

Lloyd S Shapley et al. A value for n-person games. *Annals of Mathematics Studies*, 28:307–318, 1953.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. ISSN 1476-4687. DOI: 10.1038/nature24270. URL https://www.nature.com/articles/nature24270;.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44, August 1988. ISSN 0885-6125. DOI: 10.1023/A:1022633531479. URL https://doi.org/10.1023/A:1022633531479.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, second edition, 2018.

Richard S. Sutton, Doina Precup, and Satinder P. Singh. Between MDPs and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2): 181–211, 1999. DOI: 10.1016/S0004-3702(99)00052-1. URL https://doi.org/10.1016/S0004-3702(99)00052-1.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '11, pp. 761–768, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0982657161.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. DOI: 10.1109/9.580874.

Akash Velu, Skanda Vaidyanath, and Dilip Arumugam. Hindsight-DICE: Stable credit assignment for deep reinforcement learning, 2024. URL https://openreview.net/forum?id=xIKgGWH2jI.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey, 2020. URL https://arxiv.org/abs/1908.08796.

Supplementary Materials

The following content was not necessarily subject to peer review.

Appendix Contents

A Counterfactual Simulation
B ϕ-PPO
C Theory Details
D Proofs
E Additional Experiments
F Limitations and Future Work
G Credit Assignment Methods: Detailed Comparison

A Counterfactual Simulation

Algorithm 2 CTFSIM: Counterfactual Trajectory Simulation

Input: Coalition $\mathbf{z} \in \{0,1\}^T$, trajectory $\tau = (s_{1:T}, x_{1:T}, y_{1:T})$, policies π, π_{base}
Output: $(s^{\mathbf{z}}_{1:T}, y^{\mathbf{z}}_{1:T}, Y^{\mathbf{z}})$: counterfactual states, rewards, and total return

$s^{\mathbf{z}}_1 \leftarrow s_1$
for $t = 1$ to T **do**
  if $z_t = 0$ **and** $s^{\mathbf{z}}_t = s_t$ **then**
    $x^{\mathbf{z}}_t \leftarrow x_t$ {On-track: reuse observed}
  else if $z_t = 0$ **then**
    $x^{\mathbf{z}}_t \sim \pi(\cdot \mid s^{\mathbf{z}}_t)$ {Off-track: sample from π}
  else
    $x^{\mathbf{z}}_t \sim \pi_{\text{base}}(\cdot \mid s^{\mathbf{z}}_t)$ {Intervention}
  end if
  if $(s^{\mathbf{z}}_t, x^{\mathbf{z}}_t) = (s_t, x_t)$ **then**
    $(s^{\mathbf{z}}_{t+1}, y^{\mathbf{z}}_t) \leftarrow (s_{t+1}, y_t)$ {Reuse observed}
  else
    $s^{\mathbf{z}}_{t+1} \sim P(\cdot \mid s^{\mathbf{z}}_t, x^{\mathbf{z}}_t)$
    $y^{\mathbf{z}}_t \sim r(s^{\mathbf{z}}_t, x^{\mathbf{z}}_t, \cdot)$
  end if
end for
$Y^{\mathbf{z}} \leftarrow \sum_{t=1}^{T} \gamma^{t-1} y^{\mathbf{z}}_t$ {Discounted return}
return $(s^{\mathbf{z}}_{1:T}, y^{\mathbf{z}}_{1:T}, Y^{\mathbf{z}})$

B ϕ-PPO

ϕ-PPO integrates Shapley credit assignment into PPO (Schulman et al., 2017) via two mechanisms. First, the ϕ-MDP $\mathcal{M}^{\phi}_{\pi}$ (Definition 3.4) reformulates credit as per-step reward $r^{\phi}_{\pi}(s,x) = \mathbb{E}[\phi_t \mid S_t = s, X_t = x]$, enabling standard actor-critic methods. Second, Prioritized Trajectory Replay (PTR) focuses policy updates on high-$|\phi|$ actions when credit is sparse.

Intuition. Compared to standard PPO, ϕ-PPO makes three changes: (1) replace per-step rewards r_t with Shapley values $\hat{\phi}_t$, which isolate causal contribution from return noise; (2) add a second value network V_ψ for bootstrapping counterfactual simulations (Algo. 1); the standard PPO value network V^{ϕ}_{ω} remains, now estimating advantages on the ϕ-MDP; (3) use PTR to select trajectories for replay. Compared to standard PER (Schaul et al., 2015b), PTR prioritizes by $|\phi_t|$ (causal impact) rather than

TD error $|\delta_t|$. Consider Pong: when the ball moves away from the paddle, actions have no causal effect on the outcome, yet TD error may be high if the value network is inaccurate. TD-error PER would wastefully prioritize these non-causal states; PTR correctly assigns low priority since $|\phi_t| \approx 0$. Conversely, if the policy network learns slower than the value network, TD error is low but the policy remains suboptimal at causally important states. PTR correctly prioritizes these states because $|\phi_t|$ reflects whether the action matters, not whether the value estimate is accurate.

Realized estimates $\hat{\phi}_t$ yield unbiased policy gradients when substituted for r_π^ϕ. By Theorem 4.4 with $\lambda = 1$, $\mathbb{E}[\hat{\phi}_t \mid \tau] = \phi_t$, where $\tau = (s_{1:T}, x_{1:T}, y_{1:T})$ denotes the trajectory. The law of iterated expectations then gives $\mathbb{E}[\hat{\phi}_t \mid s_t, x_t] = r_\pi^\phi(s_t, x_t)$.

The algorithm maintains two value networks: $V_\psi \approx V_\pi$ for bootstrapping counterfactual simulations in Algo. 1, and $V_\omega^\phi \approx V_\pi^\phi$ for computing ϕ-advantages.

ϕ-Advantages. Let $\hat{\phi}_t$ denote the Shapley estimate from Algo. 1. The *ϕ-TD error* measures one-step prediction error using $\hat{\phi}_t$ as reward:

$$\delta_t^\phi = \hat{\phi}_t + V_\omega^\phi(s_{t+1}) - V_\omega^\phi(s_t). \tag{12}$$

The *ϕ-advantage* generalizes GAE (Schulman et al., 2015) to the ϕ-MDP:

$$A_t^\phi = \sum_{\ell=0}^{T-t} (\lambda^\phi)^\ell \delta_{t+\ell}^\phi, \tag{13}$$

where $\lambda^\phi \in [0, 1]$ controls the bias-variance tradeoff (distinct from λ in Algo. 1, which controls bootstrapping depth). The ϕ-return is

$$G_t^\phi = \sum_{\ell=0}^{T-t} \hat{\phi}_{t+\ell}, \tag{14}$$

which serves as the training target for V_ω^ϕ (Eq. 22). Shapley efficiency ensures $\sum_t \phi_t = Y - Y_\emptyset$, where $Y_\emptyset$ is the baseline return (all-default actions); the sum of per-step credit equals the total reward improvement.

Adaptive fresh/replay ratio. Each training iteration mixes fresh rollouts with replay from the priority buffer. A static fresh probability ξ is suboptimal: early in training, fresh rollouts discover high-$|\phi|$ states; later, replay exploits known high-impact actions. We adapt ξ via Thompson sampling (Russo et al., 2018).

Let $p_j = \max_t p_{j,t}$ denote the trajectory-level priority, where $p_{j,t}$ is the priority from Eq. (20) below. A fresh trajectory "wins" if $p_j > \bar{p}_\mathcal{B}$. Here $\bar{p}_\mathcal{B} = \sum_{i \in \mathcal{B}} p_i^2 / \sum_{i \in \mathcal{B}} p_i$ is the expected priority under proportional sampling. Let $u \in [0, 1]$ denote the fraction of fresh trajectories that win. We track (α, β), initialized to $(1, 1)$, via exponential moving averages with decay $\gamma_\xi < 1$:

$$\alpha \leftarrow \gamma_\xi \alpha + u, \quad \beta \leftarrow \gamma_\xi \beta + (1 - u). \tag{15}$$

The decay discounts old observations to adapt to non-stationarity. At steady state $\alpha + \beta \to 1/(1-\gamma_\xi)$, bounding the effective sample size.

At each iteration, sample $\xi \sim \mathrm{Beta}(\alpha + 1, \beta + 1)$ and collect $n_{\text{fresh}} = \mathrm{clamp}(\lfloor \xi B \rfloor, 1, B-1)$ fresh rollouts. Stratified sampling reduces variance in the fresh/replay ratio compared to per-trajectory Bernoulli sampling.

Importance-sampling corrections. The IS correction depends on the trajectory source. Both fresh and replay trajectories use the PPO ratio for multi-epoch updates:

$$r_{j,t} = \pi_\theta(x_{j,t} \mid s_{j,t}) / \pi_{\theta_{\text{old}}}(x_{j,t} \mid s_{j,t}), \tag{16}$$

where $\pi_{\theta_{\text{old}}}$ is the policy snapshot at the start of the current training step. The clipped surrogate objective (Schulman et al., 2017) prevents large policy updates:

$$\ell(r, A) = -\min\big(rA,\ \text{clip}(r, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})A\big). \tag{17}$$

Replay trajectories additionally require a PTR importance-sampling weight to correct for priority-based sampling. Let $P_j = p_j / \sum_{i\in\mathcal{B}} p_i$ denote the sampling probability for trajectory j from buffer $\mathcal{B}$. The normalized IS weight is:

$$w_j = (|\mathcal{B}| \cdot P_j)^{-\beta_{\text{is}}} \Big/ \max_{j'\in\mathcal{B}} (|\mathcal{B}| \cdot P_{j'})^{-\beta_{\text{is}}}. \tag{18}$$

The exponent β_{is} anneals from 0 to 1 over training; at $\beta_{\text{is}} = 1$, the weights fully correct for non-uniform sampling. Since ϕ-values are *recomputed* for replay trajectories using the current policy and value function, no policy importance weight is needed.

The full policy loss applies the PTR weight only to replay trajectories:

$$\mathcal{L}_\theta^{\text{PTR}} = \frac{1}{B}\sum_{j=1}^{B} \tilde{w}_j \sum_{t=1}^{T} \ell(r_{j,t}, A_t^\phi), \quad \tilde{w}_j = \begin{cases} 1 & \text{if } j \text{ is fresh} \\ w_j & \text{if } j \text{ is replay} \end{cases}. \tag{19}$$

PTR: Priority computation. When credit is sparse, PTR focuses updates on high-impact actions. The priority for state s_t is

$$p_t = (\phi_t^{(2)})^{\alpha_{\text{ptr}}/2} + \epsilon_{\text{ptr}}, \tag{20}$$

where $\phi_t^{(2)}$ is a bias-corrected exponential moving average (Kingma & Ba, 2015) of squared Shapley estimates. The priority $\sqrt{\phi^{(2)}}$ is an exponentially-weighted root-mean-square (RMS), capturing both mean magnitude and variance: $\sqrt{\mathbb{E}[\phi^2]} = \sqrt{\mu^2 + \sigma^2}$. This prioritizes states where (i) actions have consistent causal impact (high $|\mu|$), and (ii) the action choice matters but the policy has not yet converged (high σ with $\mu \approx 0$). The exponent α_{ptr} controls priority sharpness; ϵ_{ptr} prevents starvation.

Priority decay. Priorities become stale as the policy improves. Each iteration, all priorities decay by $p_j \leftarrow \gamma_p p_j$, where $\gamma_p \in [0.99, 0.9999]$. Trajectories with $p_j < \epsilon_{\text{evict}}$ are evicted, implicitly bounding the buffer size to $O(\log(\epsilon_{\text{evict}})/\log(\gamma_p))$ iterations of data.

Training. Three components train jointly. The value network V_ψ for counterfactual bootstrapping trains on both actual and counterfactual TD errors from Algo. 1:

$$\mathcal{L}_V = \tfrac{1}{(M+1)BT}\Big(\textstyle\sum_{j,t} \delta_{j,t}^2 + \sum_{j,t,m} (\delta_{j,t}^{\mathbf{z}^{(m)}})^2\Big). \tag{21}$$

The ϕ-critic V_ω^ϕ regresses to the ϕ-return:

$$\mathcal{L}_{V^\phi} = \tfrac{1}{BT}\textstyle\sum_{j,t} (G_{j,t}^\phi - V_\omega^\phi(s_{j,t}))^2. \tag{22}$$

Following Schulman et al. (2017), all three components train via a single combined loss with gradient clipping (norm $g_{\max}$):

$$\mathcal{L}^* = \mathcal{L}_\theta^{\text{PTR}} + c_1\mathcal{L}_V + c_2\mathcal{L}_{V^\phi} - c_3 H(\pi_\theta), \tag{23}$$

where c_1, c_2 weight the value losses and c_3 weights the entropy bonus $H(\pi_\theta) = -\mathbb{E}_{s\sim d_\pi}\big[\sum_x \pi_\theta(x \mid s)\log\pi_\theta(x \mid s)\big]$ (Schulman et al., 2017) (Table 3).

Algorithm. Algo. 3 summarizes ϕ-PPO. Each iteration samples subtrajectories via PTR, computes ϕ-estimates, and updates all three networks. Advantages are normalized by batch standard deviation (not z-scored, to preserve $\phi = 0$ as the no-effect baseline).

Algorithm 3 ϕ-PPO

Input: Policy π_θ; value networks V_ψ, V_ω^ϕ; batch size B; coalition samples M; epochs K; decays γ_ξ, γ_p
Output: Updated parameters θ, ψ, ω

Initialize priority buffer $\mathcal{B} \leftarrow \emptyset$; pseudo-counts $\alpha, \beta \leftarrow 0$
for iteration $= 1, 2, \ldots$ **do**
  Sample $\xi \sim \mathrm{Beta}(\alpha+1, \beta+1)$; set $n_{\text{fresh}} \leftarrow \mathrm{clamp}(\lfloor \xi B \rfloor, 1, B-1)$
  $\forall j \in \mathcal{B}$: $p_j \leftarrow \gamma_p p_j$; evict if $p_j < \epsilon_{\text{evict}}$ {priority decay}
  for $j = 1$ to B **do**
    if $\mathcal{B} = \emptyset$ **or** $j \le n_{\text{fresh}}$ **then**
      Roll out fresh $\tau \sim \pi_\theta$; set $\tilde{w}_j \leftarrow 1$
    else
      Sample $\tau \propto p_\tau$ from $\mathcal{B}$; set $\tilde{w}_j$ via Eq. 18
    end if
    Compute $\hat{\phi}_{1:T}, \delta_{1:T}, \delta^{\mathbf{z}}_{1:T,1:M} \leftarrow \text{L3Est}(\tau, M)$ {Algo. 1}
    Compute ϕ-advantages $A^\phi_{1:T}$ {Eq. 13}
    $A^\phi_{1:T} \leftarrow A^\phi_{1:T} / \mathrm{std}(A^\phi)$ {normalize, preserve ϕ=0}
    Compute priorities $p^\tau_{1:T}$; update $\mathcal{B}$ {Eq. 20}
  end for
  Update $\alpha \leftarrow \gamma_\xi \alpha + u$, $\beta \leftarrow \gamma_\xi \beta + (1-u)$ {u: fresh win rate}
  Optimize $\mathcal{L}^*$ w.r.t. θ, ψ, ω for K epochs {Eq. 23}
end for

Guarantees. ϕ-PPO inherits theoretical properties from its components.

Gradient equivalence. By Lemma 3.5, ϕ-values produce the same expected policy gradient as original returns: $\mathbb{E}[\sum_t \phi_t \nabla \log \pi_t] = \mathbb{E}[Y \sum_t \nabla \log \pi_t]$. Thus ϕ-PPO targets the same optimum as standard PPO.

Bootstrap bias. The λ-return (Eq. 10) introduces bias bounded by $O(\gamma(1-\lambda)\epsilon_V/(1-\gamma\lambda))$ where $\epsilon_V = \|V - V_\pi\|_\infty$ (Theorem 4.4). This bias vanishes as $V_\psi \to V_\pi$.

Trajectory replay. Because ϕ-values are defined over trajectories (Definition 3.4), PTR replays entire subtrajectories rather than individual transitions. This is sound by Corollary C.4: for any state s and horizon n, the optimal policy in the undiscounted truncated ϕ-MDP $\mathcal{M}^\phi_n(s)$ equals the optimal policy in $\mathcal{M}$. Since $V^{\phi,*} = 0$ under self-baseline at optimality, no bootstrap is needed in the exact case; in practice, $\hat{V}^\phi \approx 0$ during training. IS weights (Eq. 18) correct for the prioritized sampling distribution, ensuring unbiased gradient estimates.

Hyperparameters are in Table 3; proofs in Appendix D.4.

C Theory Details

Definition C.1 (Functional Dependence)**.** Given world $(\mathcal{M}, \mathbf{u})$, the functional dependence of Y on X under baseline $z = (x', \mathbf{z}')$ is:

$$c(\mathcal{M}, \mathbf{u}, X, Y, z) = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}', x'}(\mathbf{u}) \tag{24}$$

The baseline space is $\mathcal{Z} = \{(x', \mathbf{z}') : x' \in \mathcal{D}_X, \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{z}' \in \mathcal{D}_{\mathbf{Z}}\}$ equipped with probability measure $P^{\mathcal{M}}(\mathbf{Z}, X)$.

This measures the outcome difference when X changes from baseline x' to its actual value, with other variables held at $\mathbf{z}'$. When $\mathbf{z}' = \emptyset$, this reduces to the unit-level total effect $Y(\mathbf{u}) - Y_{x'}(\mathbf{u})$. Functional dependence captures the counterfactual test: if changing X changes Y, then X is

responsible for Y. This condition is necessary for causation across major definitions in the actual causation literature (Halpern, 2016; Beckers & Vennekens, 2018; Beckers, 2021).

To formalize the condition, we introduce the *why-query* $\mathrm{Why}(Y \mid \mathbf{e}; \mathbf{X})$, composed of three parts:

(1) the event being explained, or *explanandum*, $Y = y$, inducing alternate values, or *foils*, $Y = y' \neq y$;
(2) the *evidence* $\mathbf{E} = \mathbf{e}$, the observed context used to infer causes; and
(3) the *explanatory variables* $\mathbf{X} \subseteq \mathbf{V}$, which prescribe the subset of variables to attribute to.[1]

We write $\mathrm{Why}(y \mid \mathbf{e})$ as shorthand for $\mathrm{Why}(Y{=}y \mid \mathbf{e}; \mathbf{X})$ when $\mathbf{X}$ is clear from context.

Given a why-query, an Explanatory Variable Attribution (EVA) assigns a numerical attribution to each explanatory variable in $\mathbf{X}$.

Definition C.2 (Explanatory Variable Attribution)**.** An explanatory variable attribution (EVA) is a function $\phi : \Omega \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}$, where Ω is the space of SCMs over $\mathbf{V}$ and $\mathbb{W}$ is the space of why-queries $\mathrm{Why}(Y \mid \mathbf{e}; \mathbf{X})$ asking why $Y = y \in \mathbf{e}$, given evidence $\mathbf{E} = \mathbf{e}$, in terms of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$.

With EVA, we are ready to formally define the Causal Credit Assignment criteria from Def. 3.1.

Definition C.3 (Causal Credit Assignment, formal)**.** We define four desiderata as mappings $D_{1:4} : \Omega \times \mathbb{C} \times \Phi \to \{0, 1\}$, where $\mathbb{C}$ is the space of causal measures and Φ is the space of EVAs. Given SCM $\mathcal{M} \in \Omega$, causal measure $c \in \mathbb{C}$, and EVA $\phi \in \Phi$, we say $D_i(\mathcal{M}, c, \phi) = 1$ when:

$$
\begin{aligned}
D_1: &\quad c = 0 \implies \phi = \mathbf{0} && \text{(Causal Admissibility)}\\
D_2: &\quad \exists \mathbf{u}, z : c \neq 0 \implies \phi \neq \mathbf{0} && \text{(Causal Power)}\\
D_3: &\quad c_{\mathcal{M}} = c_{\mathcal{M}'} \wedge P^{\mathcal{M}}_{\bar{c}}(Z) \not\equiv P^{\mathcal{M}'}_{\bar{c}}(Z) \wedge P^{\mathcal{M}}_{+}(z) \geq P^{\mathcal{M}'}_{+}(z)\\
&\quad \implies \phi_{\mathcal{M}} > \phi_{\mathcal{M}'} && \text{(Causal Normality)}\\
D_4: &\quad c_{\mathcal{M}} \geq c_{\mathcal{M}'} \wedge P^{\mathcal{M}}(Z) \equiv P^{\mathcal{M}'}(Z) \wedge P^{\mathcal{M}}(c_Z) \not\equiv P^{\mathcal{M}'}(c_Z)\\
&\quad \implies \phi_{\mathcal{M}} > \phi_{\mathcal{M}'} && \text{(Causal Effect Scaling)}
\end{aligned}
$$

Quantification: all premises universally quantified unless marked $\exists$. Notation: c, $c_{\mathcal{M}}$ abbreviate $c(\mathcal{M}, \mathbf{u}, X, Y, z)$; ϕ, $\phi_{\mathcal{M}}$ abbreviate $\phi_X(\mathcal{M}, w)$; $P \equiv P'$ denotes distributional equality; $P^{\mathcal{M}}_{\bar{c}}(Z) := P^{\mathcal{M}}(Z \mid c \neq 0)$; $P^{\mathcal{M}}_{+}(z) := \mathrm{sign}(c_{\mathcal{M}}) \cdot P^{\mathcal{M}}(z)$.

Chunk-size Updates. Theorem 3.7 establishes optimality equivalence for full-length trajectories optimization. In practice, we chunk trajectories: each chunk begins at an arbitrary state s (sampled from a replay buffer) and extends for n steps with a value bootstrap at the end. The following corollary shows that both modifications—arbitrary starting state and finite-horizon truncation—preserve the optimal policy.

Corollary C.4 (Chunked ϕ-Learning)**.** *For any state s and horizon n, let $\mathcal{M}^{\phi}_{n}(s)$ denote the undiscounted ϕ-MDP starting from s with actions $x_{1:n}$ and terminal value $V^{\phi}(s_n)$. The optimal policy in $\mathcal{M}^{\phi}_{n}(s)$ equals the optimal policy in $\mathcal{M}$.*

The corollary holds for *any* state s, not only the initial state distribution. This justifies experience replay: sampling (s_t, x_t) from a buffer and computing $\phi_{t:t+n}$ yields correct policy updates regardless of trajectory origin. Under self-baseline at optimality, $V^{\phi,*}(s) = 0$ for all s (Shapley efficiency), so no bootstrap is needed in the exact case. In practice, $\hat{V}^{\phi} \approx 0$ during training, and the bootstrap error vanishes as the policy converges. This enables Prioritized Trajectory Replay (PTR) using $|\phi_t|$ as priority. TD-error prioritization samples states where the value function is inaccurate, but accurate values do not guarantee optimal actions. In contrast, $|\phi_t|$ measures causal contribution: large $|\phi_t|$

[1] Explicitly selecting the set of explanatory variables as a subset of observed variables allows us to exclude irrelevant or unobserved causes such as "Alice arriving" or "the Big Bang", even though it is true that if Alice did not arrive, or the Big Bang did not occur, the funding outcome would be different!

Table 1: Three dimensions where ϕ-methods improve sample complexity.

Dimension	Improvement	Condition	TD Failure
Sparse causality	$\times\, k'/T$ variance	$k' \ll T$	Wasted updates
High stochasticity	$\times\, (1-\rho_Y)$ variance	$\rho_Y \approx 1$	Noise swamps signal
Delayed reward	$O(T) \to O(1)$ episodes	Large delay	Slow propagation

indicates the action significantly affected the outcome. Policy improvement at states with large $|\phi_t|$ affects returns the most. When credit is sparse ($k' \ll T$ causal actions), non-causal state-action pairs have $\phi_t \approx 0$ and receive minimal priority, concentrating updates on the causally-relevant subset.

D Proofs

This appendix contains proofs and additional material omitted from the main text.

D.1 Notation

D.2 Problem Setting (Section 3)

Lemma D.1 (Variance of Discounted Sum)**.** *If* $\mathbb{E}[Y_t^2] \leq \sigma^2$ *for all* t*, then* $\mathrm{Var}[Y] \leq \sigma^2/(1-\gamma)^2$ *where* $Y = \sum_{t=1}^{T} \gamma^{t-1} Y_t$.

Proof. By Cauchy-Schwarz and $\mathrm{Var}[Y_t] \leq \mathbb{E}[Y_t^2] \leq \sigma^2$, we have $|\mathrm{Cov}[Y_t, Y_s]| \leq \sigma^2$. Thus:

$$\mathrm{Var}[Y] = \sum_{t,s} \gamma^{t+s-2} \mathrm{Cov}[Y_t, Y_s] \leq \sigma^2 \left(\sum_{t=1}^{\infty} \gamma^{t-1} \right)^2 = \frac{\sigma^2}{(1-\gamma)^2}. \tag{25}$$

□

Proposition D.2 (Sample Complexity is γ-Independent)**.** *For relative error targets* $\epsilon_{\mathrm{rel}} = \epsilon/|\phi|$, $M_{\delta,\epsilon_{\mathrm{rel}}} = O(\sigma^2/(\delta \epsilon_{\mathrm{rel}}^2 R_{\max}^2))$.

Proof. Returns scale as $O(R_{\max}/(1-\gamma))$, so $\epsilon = O(\epsilon_{\mathrm{rel}} R_{\max}/(1-\gamma))$. Substituting: $M_{\delta,\epsilon} = \sigma^2/((1-\gamma)^2 \delta \epsilon^2) = \sigma^2/(\delta \epsilon_{\mathrm{rel}}^2 R_{\max}^2)$. □

D.3 Credit Assignment (Section 3)

Theorem 3.2 (ϕ-Values are Causal)**.** *ϕ-values satisfy the causal credit assignment criteria,* $\mathbf{D_1} - \mathbf{D_4}$.

Proof. We show ϕ-values (Shapley values of the NTE game $f(\mathbf{z}) = Y - Y_{\sigma_{\mathbf{z}}}$) satisfy each desideratum (formal definitions in Appendix C).

D1 (Causal Admissibility). Suppose $c = 0$: the functional dependence of Y on X_t is zero for all contexts $(\mathbf{u}, z)$. Then changing X_t's intervention status does not affect the counterfactual outcome: $Y^{\mathbf{z} \cup \{t\}} = Y^{\mathbf{z}}$ for every coalition $\mathbf{z}$. All marginal contributions vanish: $f(\mathbf{z} \cup \{t\}) - f(\mathbf{z}) = Y^{\mathbf{z}} - Y^{\mathbf{z} \cup \{t\}} = 0$. By the Shapley null player axiom, $\phi_t = 0$.

D2 (Causal Power). Suppose $\exists\, \mathbf{u}, z$ such that $c(\mathcal{M}, \mathbf{u}, X_t, Y, z) \neq 0$. Then there exists a coalition $\mathbf{z}^*$ (corresponding to context z) where the marginal contribution is nonzero: $f(\mathbf{z}^* \cup \{t\}) - f(\mathbf{z}^*) \neq 0$. Since all Shapley weights $w(\mathbf{z}) = \frac{|\mathbf{z}|!(T-|\mathbf{z}|-1)!}{T!} > 0$, the weighted sum $\phi_t = \sum_{\mathbf{z}: z_t = 0} w(\mathbf{z})[f(\mathbf{z} \cup \{t\}) - f(\mathbf{z})]$ includes a nonzero term with a strictly positive weight. Therefore $\phi_t \neq 0$.

Table 2: Notation glossary

Symbol	Meaning
SCM and MDP	
$\mathbf{V}, \mathbf{U}$	Endogenous, exogenous variables
τ	Trajectory: $(s_{1:T}, x_{1:T}, y_{1:T})$
S_t, X_t, Y_t	State, action, reward at time t
$\pi(x \mid s)$	Policy (action distribution given state)
$\gamma \in (0, 1)$	Discount factor
T	Horizon (episode length)
Coalitions and Counterfactuals	
$\mathbf{z} \in \{0, 1\}^T$	Coalition mask ($z_t = 1 \Leftrightarrow X_t$ intervened)
$\lvert\mathbf{z}\rvert = k$	Coalition size
$s^{\mathbf{z}}, Y^{\mathbf{z}}$	Counterfactual state, outcome under mask $\mathbf{z}$
Shapley Values and Credit	
ϕ_t	Counterfactual Shapley value of action X_t
$\kappa_t(\mathbf{z})$	Shapley kernel weight
G_t	Observed return from time t: $\sum_{k \geq 0} \gamma^k y_{t+k}$
G_t^ϕ	ϕ-return from time t: $\sum_{\tau \geq t} \gamma^{\tau - t} \phi_\tau$
$\mathcal{M}_\pi^\phi$	ϕ-MDP (reward-redistributed MDP)
$r_\pi^\phi(s, x)$	ϕ-MDP reward: $\mathbb{E}[\phi_t \mid S_t = s, X_t = x]$
ϕ-PPO (Algo. 3)	
V_ψ	Value network for original MDP (bootstrapping)
V_ω^ϕ	Value network for ϕ-MDP (advantages)
B	Batch size (number of trajectories)
w_j	PTR importance-sampling weight for sample j
Analysis	
(ϵ, δ)	Accuracy parameters
$M_{\delta,\epsilon}$	Sample complexity: $\sigma^2/((1-\gamma)^2 \delta \epsilon^2)$
k'	Number of causal actions (with $\phi_t \neq 0$)
λ	NTE mixing parameter (0: bootstrap, 1: Monte Carlo)
λ^ϕ	GAE mixing for ϕ-advantage estimation
ρ_Y	Outcome correlation under shared exogenous noise

D3 (Causal Normality). Consider two SCMs $\mathcal{M}, \mathcal{M}'$ with identical causal effects ($c_{\mathcal{M}} = c_{\mathcal{M}'}$) but different baseline distributions satisfying $P_+^{\mathcal{M}}(z) \geq P_+^{\mathcal{M}'}(z)$. The NTE game value $f(\mathbf{z})$ depends on the SCM through the counterfactual simulation, which samples baseline actions from π_{base}. When $P_+^{\mathcal{M}}(z) \geq P_+^{\mathcal{M}'}(z)$, the baseline under $\mathcal{M}$ assigns weakly higher probability to baselines that produce positive causal effects. Each marginal contribution $f(\mathbf{z} \cup \{t\}) - f(\mathbf{z})$ measures the additional effect of intervening at step t. Under $\mathcal{M}$, the "default" behavior (baseline) is more aligned with positive outcomes, so the departure from baseline caused by X_t receives greater credit. Formally, the Shapley value is the expected marginal contribution over uniformly random orderings; since each marginal contribution under $\mathcal{M}$ weakly exceeds that under $\mathcal{M}'$ (by the monotonicity of the NTE in baseline probabilities when the causal effect sign is fixed), we obtain $\phi_{\mathcal{M}} > \phi_{\mathcal{M}'}$.

D4 (Causal Effect Scaling). Consider two SCMs with identical baseline distributions ($P^{\mathcal{M}}(Z) \equiv P^{\mathcal{M}'}(Z)$) but $c_{\mathcal{M}} \geq c_{\mathcal{M}'}$ (larger causal effect in $\mathcal{M}$). Larger causal effects directly increase the magnitude of counterfactual outcome differences: for each coalition $\mathbf{z}$, $|Y^{\mathbf{z}} - Y^{\mathbf{z} \cup \{t\}}|$ is weakly larger under $\mathcal{M}$. Since baseline distributions are identical, the Shapley weights are the same in both games. By the Shapley monotonicity property (if player t's marginal contribution in game v weakly exceeds that in game v' for every coalition, then $\phi_t(v) \geq \phi_t(v')$), we obtain $\phi_{\mathcal{M}} > \phi_{\mathcal{M}'}$. □

Lemma 3.5 (Gradient Equivalence)**.** *For any baseline policy* π_{base}*:* $\mathbb{E}_\pi[\sum_t \phi_t \cdot \nabla \log \pi(X_t \mid S_t)] = \mathbb{E}_\pi[Y \cdot \sum_t \nabla \log \pi(X_t \mid S_t)]$.

Proof. By Shapley efficiency, $\sum_t \phi_t = Y - Y^{\mathbf{1}}$ where $Y^{\mathbf{1}} = Y_{\sigma_{\mathbf{1}}}$ is the full-baseline outcome. We show $\mathbb{E}[Y^{\mathbf{1}} \cdot \sum_t \nabla \log \pi_t] = 0$.

Under any baseline policy π_{base} (including self-baseline), the counterfactual $Y^{\mathbf{1}}$ replaces all observed actions with fresh samples $X^{\mathbf{1}}_{1:T} \sim \pi_{\text{base}}(\cdot \mid S^{\mathbf{1}}_{1:T})$. These samples are drawn independently of the observed actions $X_{1:T}$. We claim $Y^{\mathbf{1}} \perp X_t \mid S_t$. By the Markov property of the MDP, $X_t \sim \pi(\cdot \mid S_t)$ is conditionally independent of all preceding random variables given S_t. The baseline outcome $Y^{\mathbf{1}}$ is a deterministic function of the shared initial state S_1 and the counterfactual noise (actions from π_{base} and resampled transitions), all of which are independent of X_t. Although $Y^{\mathbf{1}}$ depends on S_1, which is an ancestor of S_t, the Markov property ensures $S_1 \perp X_t \mid S_t$. Therefore $Y^{\mathbf{1}} \perp X_t \mid S_t$ for any baseline policy. By the score function identity:

$$\mathbb{E}[Y^{\mathbf{1}} \cdot \nabla \log \pi(X_t \mid S_t)] = \mathbb{E}_{S_t}[\mathbb{E}[Y^{\mathbf{1}} \mid S_t] \cdot \mathbb{E}_{X_t}[\nabla \log \pi(X_t \mid S_t)]] = 0. \tag{26}$$

□

Theorem 3.7 (Optimal Policy Equivalence)**.** *ϕ-MDP preserves optimal policies as the original MDP.*

Proof. By Lemma 3.5, $\nabla_\theta J_\phi(\theta) = \nabla_\theta J(\theta)$ for all θ. Two functions with identical gradients everywhere have identical stationary points (where the gradient vanishes). Therefore both objectives share the same critical points, including maxima.

Explicitly, by Shapley efficiency, $J_\phi(\theta) = J(\theta) - \mathbb{E}_{\pi_\theta}[Y^{\mathbf{1}}]$. The term $\mathbb{E}_{\pi_\theta}[Y^{\mathbf{1}}]$ (expected baseline return under self-baseline) has zero gradient by the same argument as Lemma 3.5: $Y^{\mathbf{1}}$ is independent of the realized actions X_t, so $\mathbb{E}[Y^{\mathbf{1}} \nabla \log \pi] = 0$. Thus J_ϕ and J differ by a function with zero gradient, and $\arg\max J_\phi = \arg\max J$. □

Corollary D.3 (ϕ-Value Function)**.** *The ϕ-MDP $\mathcal{M}^\phi_\pi$ has a well-defined value function $V^\phi(s) = \mathbb{E}[\sum_{\ell \geq 0} \phi_{t+\ell} \mid S_t = s]$ satisfying the Bellman equation $V^\phi(s) = \mathbb{E}_{x \sim \pi}[r^\phi_\pi(s, x) + \mathbb{E}_{s' \sim P(\cdot \mid s, x)}[V^\phi(s')]]$. The ϕ-MDP is undiscounted ($\gamma_\phi = 1$): discounting is already incorporated into the Shapley values via the NTE game outcome $Y = \sum_t \gamma^{t-1} r_t$.*

Proof. The ϕ-MDP $\mathcal{M}^\phi_\pi$ is a valid MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition $P(s' \mid s, x)$, discount $\gamma_\phi = 1$, and reward $r^\phi_\pi(s, x) = \mathbb{E}[\phi_t \mid S_t = s, X_t = x]$. Discounting is not applied to ϕ-rewards because the Shapley game outcome $Y = \sum_t \gamma^{t-1} r_t$ already incorporates temporal discounting; applying γ again would double-discount. Since $\mathbb{E}[|\phi_t|] < \infty$ (bounded by $\mathbb{E}[|Y|] < \infty$ via efficiency) and episodes are finite ($T < \infty$), the undiscounted return $\sum_t \phi_t$ is integrable. Standard MDP theory applies: value functions exist and satisfy Bellman equations. □

Corollary C.4 (Chunked ϕ-Learning)**.** *For any state s and horizon n, let $\mathcal{M}^\phi_n(s)$ denote the undiscounted ϕ-MDP starting from s with actions $x_{1:n}$ and terminal value $V^\phi(s_n)$. The optimal policy in $\mathcal{M}^\phi_n(s)$ equals the optimal policy in $\mathcal{M}$.*

Proof. The proof proceeds in two steps, corresponding to the two truncations.

Step 1: Arbitrary initial state. Theorem 3.7 shows that the optimal policy in $\mathcal{M}^\phi$ equals the optimal policy in $\mathcal{M}$. By the Markov property, the optimal policy $\pi^*(x \mid s)$ depends only on the current state s, not on the initial state distribution or the trajectory history. Therefore, π^* is optimal in $\mathcal{M}^\phi(s)$ for any starting state s. This justifies experience replay: transitions sampled from a buffer correspond to different starting states, but all yield the same optimal policy.

Step 2: Finite horizon with value bootstrap. Consider the n-step truncated ϕ-MDP $\mathcal{M}_n^\phi(s)$ with terminal value $V^\phi(s_n)$. Since the ϕ-MDP is undiscounted ($\gamma_\phi = 1$, Corollary D.3), the objective is:

$$J_n^\phi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{n} \phi_t + V^\phi(s_n) \mid S_1 = s\right]. \tag{27}$$

Under self-baseline at the optimal policy π^*, Shapley efficiency gives $\mathbb{E}_{\pi^*}[\sum_t \phi_t \mid S_1 = s] = \mathbb{E}_{\pi^*}[Y - Y^{\mathbf{1}} \mid S_1 = s] = V^*(s) - V^*(s) = 0$ for all s. Therefore $V^{\phi,*}(s) = 0$ for all s, and the n-step objective simplifies to $J_n^\phi(s) = \mathbb{E}_\pi[\sum_{t=1}^n \phi_t \mid S_1 = s]$. By Lemma 3.5 applied to the n-step game (the proof uses Assumption 3.3 step-by-step and carries through for any finite horizon), $\nabla_\theta \mathbb{E}_\pi[\sum_{t=1}^n \phi_t] = \nabla_\theta \mathbb{E}_\pi[\sum_{t=1}^n \gamma^{t-1} r_t]$. The RHS is the gradient of the standard n-step return, whose maximizer is π^*. Since J_n^ϕ and the standard objective share the same gradient for all θ, they have the same maximizer.

Approximate case. In practice, V^ϕ is learned, not zero, so we bootstrap with $\hat{V}^\phi(s_n) \approx V^{\phi,*}(s_n) = 0$. The bootstrap error $|\hat{V}^\phi(s_n)|$ introduces bias; as training converges ($\hat{V}^\phi \to 0$), the bias vanishes. This parallels the standard n-step bias-variance tradeoff: shorter chunks reduce variance from trajectory noise at the cost of bootstrap bias. □

D.4 Estimation (Section 4)

Theorem 4.2 (Optimal Proposal)**.** *Under Assumption 3.8, the variance-minimizing proposal is* $Q^*(\mathbf{z}) = q_k^*/\binom{T}{k}$ *where* $k = |\mathbf{z}|$ *and* $q_k^* \propto \sqrt{c_k}$ *with:*

$$c_k = \begin{cases} \frac{1}{T}\left(\frac{1}{k} + \frac{1}{T-k}\right) & k \in \{1, \ldots, T-1\} \\ \frac{1}{T^2} & k = T \end{cases} \tag{9}$$

Under Q^*, the estimator achieves (ϵ, δ)-accuracy with $M = O(M_{\delta,\epsilon})$ samples.

Proof. The second moment of the IS estimator is $\mathbb{E}[(\kappa_t(\mathbf{z})/Q(\mathbf{z}))^2 f(\mathbf{z})^2]$. Since $\mathrm{Var}[f(\mathbf{z})] \leq \sigma^2/(1-\gamma)^2$ (Assumption 3.8), we minimize:

$$\min_Q \sum_{\mathbf{z}} \frac{\kappa_t(\mathbf{z})^2}{Q(\mathbf{z})} \quad \text{subject to} \quad \sum_{\mathbf{z}} Q(\mathbf{z}) = 1. \tag{28}$$

The $k = 0$ coalition ($\mathbf{z} = \mathbf{0}$, no interventions) has $f(\mathbf{0}) = Y - Y = 0$ deterministically, so it contributes nothing and is excluded from sampling. Using two-level sampling over $k \in \{1, \ldots, T\}$ (size k with probability q_k, then uniform within size k): $Q(\mathbf{z}) = q_k/\binom{T}{k}$. The objective becomes $\sum_{k=1}^T c_k/q_k$ where $c_k = \frac{1}{T}(\frac{1}{k} + \frac{1}{T-k})$ for $k < T$ and $c_T = 1/T^2$. By Cauchy-Schwarz, $q_k^* \propto \sqrt{c_k}$ minimizes this. Chebyshev gives (ϵ, δ)-accuracy with $M = O(M_{\delta,\epsilon})$ samples. □

Theorem 4.4 (Bias Bound for $\hat{\phi}^\lambda$)**.** *Under Assumptions 3.3, 3.8:* $|\mathbb{E}[\hat{\phi}_t^\lambda] - \phi_t| \leq \frac{2\gamma(1-\lambda)}{1-\gamma\lambda}\|V - V_\pi\|_\infty$.

Proof. The λ-bootstrapped estimator (Definition 4.3) replaces the exact counterfactual return from step t onward with a λ-return $G_t^{\mathbf{z},\lambda}$ (Eq. 10). We bound the bias introduced by this substitution.

Step 1: λ-return bias. The λ-return is a mixture of n-step returns: $G_t^{\mathbf{z},\lambda} = (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} G_t^{(n)}$, where $G_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k y_{t+k}^{\mathbf{z}} + \gamma^n V(s_{t+n}^{\mathbf{z}})$. Taking conditional expectations given $s_t^{\mathbf{z}}$:

$$\mathbb{E}[G_t^{(n)} \mid s_t^{\mathbf{z}}] = V_\pi(s_t^{\mathbf{z}}) + \gamma^n \mathbb{E}[V(s_{t+n}^{\mathbf{z}}) - V_\pi(s_{t+n}^{\mathbf{z}}) \mid s_t^{\mathbf{z}}]. \tag{29}$$

Since $|V(s) - V_\pi(s)| \leq \epsilon_V$ for all s:

$$|\mathbb{E}[G_t^{\mathbf{z},\lambda} \mid s_t^{\mathbf{z}}] - V_\pi(s_t^{\mathbf{z}})| \leq (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1}\gamma^n \epsilon_V = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\epsilon_V. \tag{30}$$

Step 2: Per-coalition NTE bias. The exact NTE is $f(\mathbf{z}) = Y - Y^{\mathbf{z}} = Y - R_t^{\mathbf{z}} - \gamma^{t-1} G_t^{\mathbf{z},\mathrm{MC}}$, where $G_t^{\mathbf{z},\mathrm{MC}} = \sum_{s \geq t} \gamma^{s-t} y_s^{\mathbf{z}}$ is the Monte Carlo return. The bootstrapped NTE is $f_t^{\mathbf{z}} = Y - R_t^{\mathbf{z}} - \gamma^{t-1} G_t^{\mathbf{z},\lambda}$. Their difference is:

$$|E[f_t^{\mathbf{z}}] - f(\mathbf{z})| = \gamma^{t-1} |\mathbb{E}[G_t^{\mathbf{z},\mathrm{MC}} - G_t^{\mathbf{z},\lambda}]| \leq \gamma^{t-1} \cdot \frac{\gamma(1-\lambda)}{1-\gamma\lambda} \epsilon_V \leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda} \epsilon_V, \tag{31}$$

where the last inequality uses $\gamma^{t-1} \leq 1$ for $t \geq 1$.

Step 3: Shapley bias. The estimator is $\hat{\phi}_t^{\lambda} = \sum_{\mathbf{z}} \kappa_t(\mathbf{z}) f_t^{\mathbf{z}}$ (in expectation over the proposal Q^*). The true Shapley value is $\phi_t = \sum_{\mathbf{z}} \kappa_t(\mathbf{z}) f(\mathbf{z})$. Using the marginal contribution form with non-negative Shapley weights $w(\mathbf{z}) \geq 0$ summing to 1:

$$|\mathbb{E}[\hat{\phi}_t^{\lambda}] - \phi_t| = \Big| \sum_{\mathbf{z}: z_t = 0} w(\mathbf{z}) \big(\mathbb{E}[f_t^{\mathbf{z} \cup \{t\}} - f_t^{\mathbf{z}}] - [f(\mathbf{z} \cup \{t\}) - f(\mathbf{z})] \big) \Big| \tag{32}$$

$$\leq \sum_{\mathbf{z}: z_t = 0} w(\mathbf{z}) \big(|\mathbb{E}[f_t^{\mathbf{z} \cup \{t\}}] - f(\mathbf{z} \cup \{t\})| + |\mathbb{E}[f_t^{\mathbf{z}}] - f(\mathbf{z})| \big) \tag{33}$$

$$\leq \frac{2\gamma(1-\lambda)}{1-\gamma\lambda} \epsilon_V. \qquad \square$$

The last line uses Step 2 for each of the two bias terms and $\sum w(\mathbf{z}) = 1$. Both coalitions $\mathbf{z}$ and $\mathbf{z} \cup \{t\}$ share the same counterfactual states up to step t (the intervention at step t only affects states from $t+1$ onward), so both bootstrap from the same starting state $s_t^{\mathbf{z}}$, and the bound from Step 2 applies to each. The factor of 2 arises because the continuation trajectories diverge after step t (different actions), so the two bootstrap biases do not cancel in general.

D.5 PTR Buffer Design (Section B)

PTR stores full trajectories with per-timestep priorities. The priority formula uses a smoothed estimate:

$$p_t = (\phi_t^{(2)})^{\alpha_{\mathrm{ptr}}/2} + \epsilon_{\mathrm{ptr}}, \tag{34}$$

where $\phi_t^{(2)}$ is a bias-corrected exponential moving average (Kingma & Ba, 2015) of squared Shapley estimates, α_{ptr} controls priority sharpness, and ϵ_{ptr} prevents starvation. Sampling selects (trajectory, starting timestep) pairs proportionally to p_t using the Gumbel-max trick for efficient vectorized sampling. When the buffer reaches capacity, the trajectory with lowest total priority $\sum_t p_t$ is evicted.

Following standard PER (Schaul et al., 2015b), we apply importance sampling (IS) correction to account for the non-uniform sampling distribution. Let p_i denote the priority of sample i in the buffer. Each sampled subtrajectory receives weight $w_i = (|\mathcal{B}| \cdot p_i / \sum_j p_j)^{-\beta_{\mathrm{is}}}$, normalized by $\max_i w_i$ for stability. The exponent β_{is} anneals from β_0 (e.g., 0.4) to 1 over training, providing full bias correction at convergence while allowing faster early learning.

D.6 When Credit Helps (Section 3.3)

Proposition 3.9 (Sparse Causality)**.** *Let* $k' = |\{t : \phi_t \neq 0\}|$, $\sigma_\phi^2 = \max_t \mathrm{Var}[\phi_t \nabla \log \pi_t]$, *and* $\bar{\rho} \in [0,1]$ *the maximal pairwise correlation among causal terms. Then* $\mathrm{Var}[\hat{g}^{\phi}] \leq k'(1 + (k'-1)\bar{\rho})\, \sigma_\phi^2$.

Proof. By the Shapley null player axiom (D1), $\phi_t = 0$ whenever X_t has no causal effect on Y, so $\hat{g}^{\phi} = \sum_{t \in \mathcal{C}} g_t$ where $g_t = \phi_t \nabla \log \pi(X_t \mid S_t)$ and $|\mathcal{C}| = k'$. Expanding the variance of the sum:

$$\mathrm{Var}[\hat{g}^{\phi}] = \sum_{t \in \mathcal{C}} \mathrm{Var}[g_t] + 2 \sum_{\substack{t,s \in \mathcal{C} \\ t<s}} \mathrm{Cov}[g_t, g_s]. \tag{35}$$

The diagonal terms are bounded by $k'\sigma_\phi^2$. For the cross-terms, $|\mathrm{Cov}[g_t, g_s]| \leq \bar{\rho}\,\sigma_\phi^2$ by definition of $\bar{\rho}$ and Cauchy-Schwarz on covariance, and there are $\binom{k'}{2}$ pairs. Combining: $\mathrm{Var}[\hat{g}^\phi] \leq k'\sigma_\phi^2 + k'(k'-1)\bar{\rho}\,\sigma_\phi^2 = k'(1+(k'-1)\bar{\rho})\,\sigma_\phi^2$.

Why $\bar{\rho}$ is small. For $t < s$, condition on (S_s, S_t, X_t) and integrate over $X_s \sim \pi(\cdot \mid S_s)$:

$$\mathrm{Cov}[g_t, g_s] = \mathbb{E}\big[g_t \cdot \mathbb{E}[\phi_s \nabla \log \pi_s \mid S_s, S_t, X_t]\big] - \mathbb{E}[g_t]\mathbb{E}[g_s]. \tag{36}$$

The score function identity $\mathbb{E}[\nabla \log \pi(X_s \mid S_s) \mid S_s] = 0$ drives the inner expectation to zero whenever ϕ_s is independent of X_s given S_s. Under self-baseline, ϕ_s depends on X_s through the on-track condition ($s_s^{\mathbf{z}} = s_s$ when the observed action is reused), introducing a residual correlation. This dependence is weak because: (i) the on-track condition at step s is primarily determined by interventions at earlier steps, not by X_s itself; and (ii) the optimal proposal Q^* concentrates on extreme coalition sizes where the on-track probability is either very high ($|\mathbf{z}| \approx 0$) or very low ($|\mathbf{z}| \approx T$). In deterministic environments, $\bar{\rho} = 0$ for on-track coalitions since ϕ_s given S_s depends on X_s only through the on-track indicator, which is determined by prior interventions.

In contrast, standard REINFORCE has $\hat{g} = Y \sum_{t=1}^{T} \nabla \log \pi_t$ where all T terms contribute regardless of causal structure, giving $\mathrm{Var}[\hat{g}] = O(T\sigma_Y^2)$. The variance ratio is $k'(1+(k'-1)\bar{\rho})/T \approx k'/T$ when $\bar{\rho} \ll 1$. □

Proposition 3.10 (High Stochasticity)**.** *Let $\mathbf{z}' = \mathbf{z}$ with $z_t = 1$ and $\rho_Y = \mathrm{Corr}[Y^{\mathbf{z}'}, Y^{\mathbf{z}}]$ under shared exogenous randomness. Then* $\mathrm{Var}[\phi_t] \propto (1 - \rho_Y)$.

Proof. The Counterfactual Shapley value is a weighted sum of marginal contributions:

$$\phi_t = \sum_{\mathbf{z}: z_t = 0} w(\mathbf{z}) \cdot [f(\mathbf{z} \cup \{t\}) - f(\mathbf{z})], \tag{37}$$

where $w(\mathbf{z}) \geq 0$ are Shapley weights summing to 1. Each marginal contribution compares outcomes $Y^{\mathbf{z}\cup\{t\}}$ and $Y^{\mathbf{z}}$ under shared exogenous noise $\mathbf{U}$. Let $\sigma^2 = \mathrm{Var}[Y^{\mathbf{z}}]$ (approximately equal for nearby coalitions) and $\rho_Y = \mathrm{Corr}[Y^{\mathbf{z}\cup\{t\}}, Y^{\mathbf{z}}]$. The variance of a single marginal contribution is:

$$\mathrm{Var}[Y^{\mathbf{z}\cup\{t\}} - Y^{\mathbf{z}}] = \mathrm{Var}[Y^{\mathbf{z}\cup\{t\}}] + \mathrm{Var}[Y^{\mathbf{z}}] - 2\mathrm{Cov}[Y^{\mathbf{z}\cup\{t\}}, Y^{\mathbf{z}}] \tag{38}$$

$$= 2\sigma^2 - 2\sigma^2\rho_Y = 2\sigma^2(1 - \rho_Y). \tag{39}$$

When environment stochasticity dominates action effects, $\rho_Y \approx 1$ (both outcomes are driven by the same exogenous noise), and variance vanishes. By convexity of variance, $\mathrm{Var}[\phi_t] \leq \max_{\mathbf{z}} \mathrm{Var}[f(\mathbf{z} \cup \{t\}) - f(\mathbf{z})] = O(\sigma^2(1-\rho_Y))$.

In contrast, independent sampling (different $\mathbf{U}$ for each outcome) gives $\rho_Y = 0$ and variance $2\sigma^2$, a factor of $1/(1-\rho_Y)$ larger. □

Proposition 3.11 (Direct Propagation)**.** *Tabular TD(0) with terminal-only reward requires $O(T)$ episodes for credit to reach the initial state. ϕ-redistribution provides credit in $O(1)$ episodes via $O(TM_{\delta,\epsilon})$ counterfactual simulations.*

Proof. Consider tabular TD(0) with terminal-only reward R_T (all $R_t = 0$ for $t < T$). Initialize $V(s) = 0$ for all states. The TD(0) update at step t is: $V(S_t) \leftarrow V(S_t) + \alpha[R_t + \gamma V(S_{t+1}) - V(S_t)]$.

Episode 1: Only the transition (S_{T-1}, S_T) has nonzero TD target ($R_T + \gamma \cdot 0 = R_T$). After episode 1: $V(S_{T-1}) > 0$, all other values remain 0.

Episode n: The TD target at step $T - n$ becomes nonzero (since $V(S_{T-n+1}) > 0$ from previous episodes). Credit propagates backward by one step per episode.

After T episodes: $V(S_1)$ receives its first nonzero update. This requires T episodes for credit to reach the initial state.

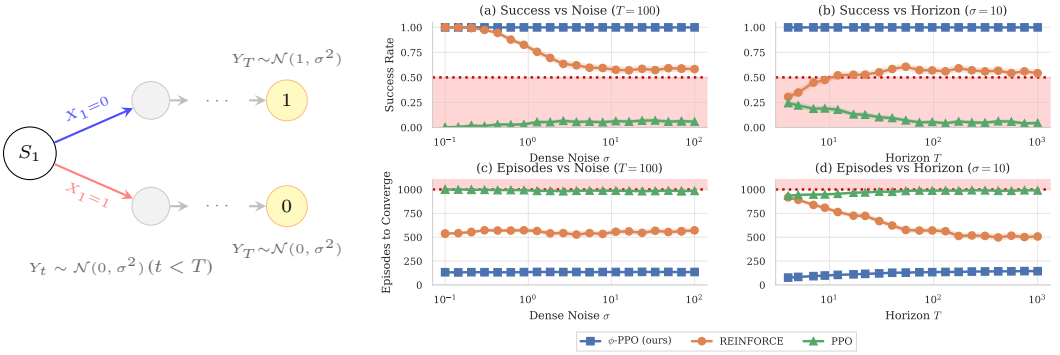

Figure 8: **Fork MDP results** (n=1,000 runs per condition; shaded bands are 95% CIs). *Left:* MDP structure; only the first c actions affect the terminal reward. *Top row (a–c):* Success rate across noise σ, horizon T, and causal steps c. *Bottom row (d–f):* Mean episodes to convergence.

In contrast, ϕ-redistribution computes ϕ_t for all t from a single trajectory via $O(T \cdot M_{\delta,\epsilon})$ counterfactual simulations. All timesteps receive credit signal in $O(1)$ episodes. The trade-off: fewer episodes but $O(T \cdot M)$ simulation cost per episode. □

Remark 3.12 (Propagation Complexity). For terminal-only reward: TD(0) requires $O(T)$ episodes and $O(T^2)$ updates; ϕ-redistribution requires $O(1)$ episodes and $O(TM_{\delta,\epsilon})$ counterfactual simulations.

Proof. Follows directly from Proposition 3.11. For TD(0), each episode updates one state; propagating terminal reward across T states requires $O(T)$ episodes with $O(T)$ updates each, totaling $O(T^2)$. For ϕ-redistribution, a single trajectory yields ϕ_t for all T timesteps via $O(T \cdot M_{\delta,\epsilon})$ counterfactual simulations, so one episode suffices. □

E Additional Experiments

This section contains additional experiments validating the theoretical predictions.

E.1 Fork MDP

The Fork MDP isolates the three variance reduction benefits of ϕ-PPO. The agent chooses from $\mathcal{A} = \{0, 1\}$ at each of T steps (γ=0.99). Only the first c actions affect the terminal reward: $Y_T \sim \mathcal{N}(1, \sigma^2)$ if $x_1 = \cdots = x_c = 0$ (the optimal actions), $Y_T \sim \mathcal{N}(0, \sigma^2)$ otherwise. Intermediate rewards are i.i.d. noise $Y_t \sim \mathcal{N}(0, \sigma^2)$ for $t < T$. The environment combines sparse causality ($k' = c \ll T$, Proposition 3.9), high stochasticity (shared noise cancels in the counterfactual difference, Proposition 3.10), and delayed reward (Y_T only, Proposition 3.11).

Results (Figure 8). At the default configuration (T=100, σ=10, c=1), ϕ-PPO achieves 100% success, while REINFORCE degrades to ∼58% and PPO fails at ∼5%. Each column isolates one benefit: noise cancellation (a), horizon independence (b), and causal isolation (c). The bottom row confirms that ϕ-PPO also converges fastest: median episodes remain low across all sweeps, while PPO and REINFORCE slow dramatically as conditions worsen. PPO's clipping saturates under high-variance advantages, preventing meaningful policy updates. REINFORCE drifts randomly in parameter space, converging roughly half the time. With M=1, ϕ-PPO requires $2T$ steps per episode (one counterfactual rollout), yielding 4× episode efficiency but 2× simulation efficiency over PPO.

Table 3: Hyperparameters for ϕ-PPO (Section B) and Fork MDP experiments (Section E.1).

Parameter	Symbol	Value
ϕ-PPO (Algo. 3)		
Discount factor	γ	0.99
Trajectory length (chunk size)	T_{chunk}	128
Batch size (trajectories)	B	64
Policy epochs	K	4
Coalition samples	M	1
NTE mixing	λ	1.0 (pure MC)
GAE mixing	λ^{ϕ}	0.95
PPO clip	ϵ_{clip}	0.2
Value loss coefficient	c_1	0.5
ϕ-value loss coefficient	c_2	0.5
Entropy coefficient	c_3	0.01
Gradient clip norm	$g_{\max}$	0.5
Exploration decay	γ_{ξ}	0.99
PTR (Appendix D.5)		
Priority exponent	α_{ptr}	0.6
UCB coefficient	β_{ucb}	1.0
Priority floor	ϵ_{ptr}	0.01
EMA decay	α_{ema}	0.99
IS exponent	β_{is}	$\beta_0 \to 1$
IS exponent start	β_0	0.4
IS annealing iters	M_{β}	100000
Priority decay	γ_p	0.999
Eviction threshold	ϵ_{evict}	10^{-6}
Optimization		
Learning rate	η	0.1
Fork MDP environment		
Episode horizon	T_{ep}	100
Causal actions	c	1
Noise std	σ	10
Network architecture	–	MLP [64, 64]

E.2 Coin Flip MDP: High Stochasticity

A single-step MDP: $U \sim \text{Bernoulli}(0.5)$, action $X \in \{0, 1\}$, return $Y = 100U + X$. The action's causal effect is $\phi = 1$ with zero variance, while $\text{Var}[Y] = 2500$. Correlation $\rho = \text{Corr}[Y(0, U), Y(1, U)] \approx 0.9998$.

E.3 Treasure Hunt MDP: Delayed Reward

A $T = 4$ step MDP with terminal reward only. Each step: collect treasure or skip. Reward $Y_3 = \#\{\text{collected}\}$. All actions are causal ($k' = 4$). ϕ-TD provides signal in 1 episode; standard TD requires $O(T) = 4$ episodes for credit propagation.

F Limitations and Future Work

Scope. Our approach requires a simulator to compute counterfactual trajectories, limiting applicability to environments with resettable state.

Limitations. Our approach requires simulator access for counterfactual trajectories. Sample complexity scales with $\sigma^2/(1-\gamma)^2$, which can be large when $\gamma \approx 1$. We assume counterfactual independence (Assumption 3.3), excluding correlated exogenous noise.

Future work. Finite-sample bounds via Bernstein concentration; relaxing counterfactual independence; empirical validation on Atari/MuJoCo benchmarks.

G Credit Assignment Methods: Detailed Comparison

To facilitate the discussion, here we entail the additional notations we use for related work comparison. $\alpha_{i\to j}$: learned attention weight from token j to token i. $g(\Delta_{0:t})$: LSTM hidden state after processing trajectory up to step t. $P^\pi(X_t{=}x \mid S_t, S_T)$: hindsight posterior over actions given a future state. $P^\pi(X_t{=}x \mid S_t, Z)$: hindsight posterior over actions given observed return Z. $\phi^\star(s,x,z)$: optimal DICE density-ratio estimator; $\chi^\pi(z \mid s)$: return distribution under π from state s. Φ_t: learned hindsight statistic of the future trajectory, constrained to satisfy $X_t \perp \Phi_t \mid S_t$. U': encoding of a future step (S', X', Y'); choices include $U'{=}S'$ (state), $U'{=}Y'$ (reward), or a learned encoding. $\mathcal{Q}(s,x,\tau)$: quantile function—inverse CDF of the return distribution $F^{-1}_{\eta_{s,x}}(\tau)$ at quantile level $\tau \in [0,1]$. $\hat{\tau}$: inferred quantile level of the observed return, $\hat{\tau} = F_{\eta_{S_t,X_t}}(Z)$. $B^\pi(s,x,s') = V(s') - \mathbb{E}_{S'\sim p(\cdot|s,x)}[V(S')]$: luck (nature's contribution to the transition).

First we provide an overview of the credit assignment methods we compare against in table 4 with their credit expressions and descriptions. Then table 5 evaluates each method against desirable credit assignment properties (3.1) on the Fork MDP (Section E.1: $T{=}100$, $\sigma{=}10$, only X_1 causal). Below we discuss the details of each category of credit assignment methods in the literature.

A. Temporal: discounted reward. In the existing credit assignment literature, action values/advantages often serve as a key proxy for actions' influence on the rewards (Watkins & Dayan, 1992; Sutton, 1988; Sutton et al., 1999; Schaul et al., 2015a; Sutton et al., 2011; Chang et al., 2021; Pan & Schölkopf, 2024; Auzina et al., 2026). Although action values and advantages yield unbiased gradients, they assign credit by actions' temporal proximity to the reward, rather than their causal effects. This is both inefficient and biased in terms of credit assignment in that recent actions with no causal effect on the reward add noise to the gradient, while causally relevant actions that occur earlier receive discounted credit. Intuitively, credit should correspond to actions cause the outcome: we credit studying hard for a good grade, rather than flipping over the returned test to view the grade.

B. Memory: learned models Another line of work utilizes learnable sequence models to automatically distribute the credit over actions (Chen et al., 2021; Arjona-Medina et al., 2019; Ferret et al., 2020; Raposo et al., 2021; Deng et al., 2026), dropping the reliance on the time contiguity assumption. However, such methods conflate agent's actions influence over rewards with environment stochasticity. As a result, they are unable to distinguish skill from luck in learning. Also, learned models suffer from the same problem as black box models that they are also in lack of explainability.

C. Hindsight: posterior/prior ratios and conditional baselines The line of hindsight conditioning (Harutyunyan et al., 2019a; Velu et al., 2024; Mesnard et al., 2021; 2023; Meulemans et al., 2023; Ramesh et al., 2025)[2] partially addresses the skill-luck issue by conditioning on future events and credit each action by how relevant it is to the given event. The fundamental limitation of those is that isolating each single action during attribution overlooks subsequent actions' collective influence on the rewards. Thus, the skill-luck issue still persists as we demonstrate in Figure 1.

[2]For clarity, prior uses of the term 'counterfactual' in this literature refer to hindsight conditioning not counterfactuals defined in the causal theory (Pearl, 2009).

Table 4: Credit assignment methods: expressions and descriptions.

Method	Expression	Description
A. Temporal: discounted reward		
TD (Williams, 1992), Options (Sutton et al., 1999), UVFA (Schaul et al., 2015a), Horde (Sutton et al., 2011)	$\gamma^{T-t}Y_T$	Reward Y_T discounted by γ^{T-t}, the temporal distance from action X_t to reward
Direct Adv. Estimation (Pan & Schölkopf, 2024)	$\hat{A}(s,x)$	Fit $Y = V(S_0) + \sum_t A(S_t, X_t) + \sum_t B^\pi(S_t, X_t, S_{t+1})$ by least squares. $\hat{A}$: agent's causal effect (skill); B^π: environment's effect (luck)
QCA (Mesnard et al., 2023)	$\mathcal{Q}(s,x,\hat{\tau}) - \sum_{x'} \pi(x') \mathcal{Q}(s,x',\hat{\tau})$	Quantile level $\hat{\tau}{=}F_{\eta_{s,x}}(Z)$ indexes "luck." Credit: action's quantile value minus the policy-averaged quantile value at the same luck level
B. Memory: learned models		
DT (Chen et al., 2021)	$\alpha_{T\to t}$	Transformer trained via cross-entropy on actions, conditioned on returns-to-go $\hat{Y}_t{=}\sum_{t'\geq t} Y_{t'}$. Credit implicit via attention weights $\alpha_{T\to t}$
RUDDER (Arjona-Medina et al., 2019)	$g(\Delta_{0:t}) - g(\Delta_{0:t-1})$	LSTM g trained via MSE on total return Y. Credit to step t: change in predicted return after observing (S_t, X_t)
SECRET (Ferret et al., 2020)	$\alpha_{t\leftarrow T}\, Y_T$	Transformer trained via weighted cross-entropy on $\mathrm{sign}(Y_T)$. Credit: attention $\alpha_{t\leftarrow T}$ on (S_t, X_t), scaled by Y_T
C. Hindsight: posterior/prior ratios and conditional baselines		
HCA (Harutyunyan et al., 2019a)	$\dfrac{P^\pi(X_t{=}x \mid S_t, S_T)}{\pi(x \mid S_t)}$	Posterior over X_t given future state S_T, divided by prior π. Ratio >1: action became more likely given the future
H-DICE (Velu et al., 2024)	$\dfrac{\pi(x \mid s)}{P^\pi(X_t{=}x \mid S_t, Z)}$	Same hindsight ratio as HCA (conditioned on return Z instead of state). Estimated via DICE variational objective $\phi^\star \cdot \chi^\pi$ instead of direct density ratio
CCA (Mesnard et al., 2021)	$G_t - \mathbb{E}[G_t \mid S_t, \Phi_t]$	Φ_t: learned compression of future trajectory $(S, X, Y)_{t+1:T}$, trained to predict G_t while satisfying $X_t \perp \Phi_t \mid S_t$. Baseline $\mathbb{E}[G_t \mid S_t, \Phi_t]$ fit by regression
COCOA (Meulemans et al., 2023)	$\dfrac{P^\pi(U' \mid S_t, X_t)}{P^\pi(U' \mid S_t)} - 1$	U': encoding of a future step (S', X', Y'); typically $U'{=}Y'$ (reward). Ratio measures how much X_t increased the probability of observing U'. Estimated by supervised learning
D. Ours: counterfactual simulation		
ϕ-**values**	$\mathbf{Y} - \mathbf{Y^z}$, **shared** $\mathbf{U}$	Replay trajectory in simulator with alternative action $X'_t{\sim}\pi$ and same exogenous noise $\mathbf{U}$. Credit: difference in outcomes $Y - Y^{\mathbf{z}}$

Table 5: Desirable credit assignment properties, following the axiomatic framework of Lee et al. (2025). **D1** (Admissibility): non-causal actions receive zero credit. **D2** (Power): causal actions receive nonzero credit. **D3** (Normality): given equal causal effects, actions whose counterfactual baselines are more probable under π receive higher credit (actions π would have taken anyway get less credit). **D4** (Effect scaling): given equal baselines, actions with larger causal effects receive proportionally more credit. **D5** (Efficiency): credit sum to total return, $\sum_t \phi_t = Y - Y^{\mathbf{0}}$. ✓: satisfied. $\sim$: partially or in principle but not in practice. ×: violated.

Method	**D1** Admiss.	**D2** Power	**D3** Normal.	**D4** Scaling	**D5** Effic.
A. Temporal					
TD	×	$\sim$	×	$\sim$	×
Options	×	$\sim$	×	$\sim$	×
UVFA	×	$\sim$	×	$\sim$	×
Horde	×	$\sim$	×	$\sim$	×
Direct Advantage Estimation	$\sim$	×	×	×	×
QCA (exact)	✓	✓	×	✓	×
QCA (learned)	✓	×	×	×	×
B. Memory					
DT	×	×	×	×	×
RUDDER	×	$\sim$	×	×	✓
SECRET	×	×	×	×	$\sim$
C. Hindsight					
HCA	$\sim$	×	$\sim$	×	×
H-DICE	$\sim$	×	$\sim$	×	×
CCA	$\sim$	$\sim$	×	$\sim$	×
COCOA	✓	×	$\sim$	×	×
D. Ours					
ϕ-**values**	✓	✓	✓	✓	✓