LEARNING FROM SUCCESSES WITH MULTI-STEP IMPLICIT VALUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering a successful trajectory does not ensure that a reinforcement learning agent learns effectively from it. In long-horizon, sparse-reward tasks, evaluating the current policy can assign low values even to states on successful trajectories, weakening the learning signal from rare successes. Building on Implicit Q Learning, we propose **multi-step implicit value estimation**, which combines expectile-based value learning with recursive temporal propagation. Rather than evaluating the behavior policy alone, the formulation targets an implicit policy that favors higher-valued actions while retaining unbiased expectations over rewards and transitions. We establish contraction of a joint action–state value operator and show that its multi-step compositions share a unique fixed point equal to the expected-return value of this implicit policy under the original MDP. We introduce $\mathrm{IQ}(\tau, \lambda)$, a coupled backward recursion for action and state values, and $\mathrm{IV}(\tau, \lambda)$, a value-only specialization for deterministic environments who recovers generalized advantage estimation (GAE) at the symmetric expectile. Experiments show improvements over PPO with GAE on multiple sparse- and dense-reward tasks, with complementary gains from exploration bonuses. Ablations highlight the importance of multi-step propagation and the versality of our methods.

1 INTRODUCTION

Reinforcement learning requires both discovering useful behavior and converting that experience into a better policy. These are related but distinct challenges. Even after observing a high-return trajectory, an agent may learn little from it if its value targets predominantly reflect poor subsequent behavior. Multi-step estimators such as generalized advantage estimation (GAE) accelerate propagation, but do not change which continuation policy is being evaluated (Schulman et al., 2016; 2017). We use *successful experience* to mean relatively high-return behavior, not only a binary terminal success; the proposed estimators require neither a success label nor sparse rewards.

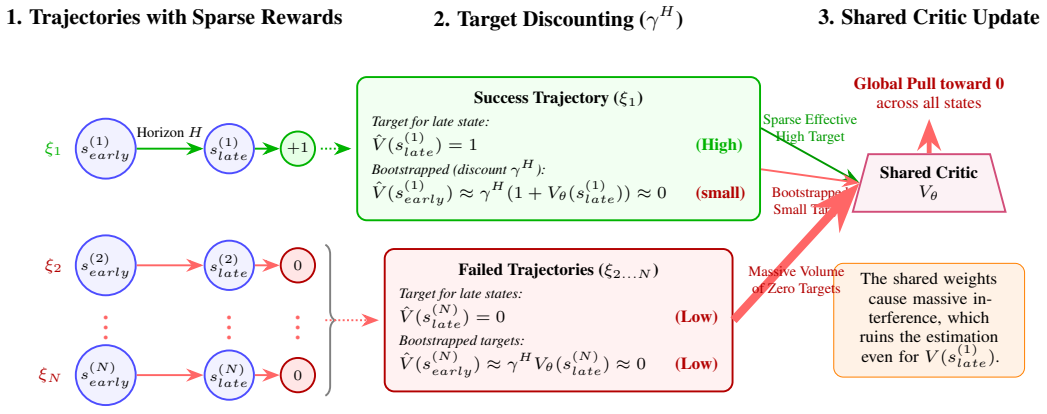


Figure 1: Sparse rewards lead to a massive volume of zero-value targets, further discounted by long horizon, biasing the shared critic globally toward zero. The shared network interference degrades the estimation of rare successful states.

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

The difficulty is particularly visible in long-horizon, sparse-reward tasks (Lee et al., 2022; Shin & Kim, 2023; Wang et al., 2020). As outlined in Figure 1, a critic can assign low values even to states along a successful route, when the policy usually fails. Such values need not be inaccurate: they may correctly describe the behavior policy. Shared nonlinear function approximation can add a second difficulty: fitting numerous low-return targets can alter predictions on successful trajectories through shared parameters. Bootstrapping then reuses these predictions in targets such as $r+\gamma V_\psi(s')$, passing their errors to earlier states and potentially reinforcing the interference. This mechanism is consistent with evidence that bootstrapping affects gradient interference and target coherence in temporal-difference learning (Bengio et al., 2020). It motivates a broader question: *which policy's value provides a useful learning signal?*

Our formulation builds on *implicit Q-learning* (IQL) (Kostrikov et al., 2022) that introduce expectile value to offline value estimation. Its recursive equations separate environmental averaging from an expectile over actions. IDQL further characterizes an associated implicit actor for a given action-value function (Hansen-Estruch et al., 2023). However, a fundamental convergence theorem that established tabular Q learning converges have not been worked out for the implicit Q learning. We answer three questions: 1) whether the recursive value pair is well defined, 2) which policy it evaluates to and how policy improves, 3) whether *coupled partial updates* converge to that pair. The last question is not answered by characterizing an exact statewise expectile: both critics move, and each changes the other's target. We establish a joint contraction without requiring either regression problem to be solved fully at each update. This permits multi-step propagation of an implicit policy's value without first fitting an explicit actor to that policy.

Our method performs a backward recursion over each rollout. $\mathrm{IQ}(\tau,\lambda)$ retains separate action and state critics, with parameters λ_q and λ_v controlling the action-value correction and asymmetric state-value propagation. Setting $\lambda_q = 1$ removes the action critic from the recursion and yields $\mathrm{IV}(\tau,\lambda)$. Positive corrections receive more weight than negative corrections, but failures are not removed from training. The estimators supply critic targets and actor scores to a PPO-style update. Minimum ensembles help control approximation-driven overestimation.

Our contributions are threefold. **First**, we prove that the population implicit Bellman operator is a contraction whose iterates converge to a unique fixed point, thereby establishing the existence and uniqueness of recursive implicit values. We identify this fixed point as the expected-return value of an implicit policy under the original MDP, and prove that coupled partial updates and expected backward propagation converge to the same target. **Second**, we introduce coupled multi-step implicit-value estimators and characterize their relationship to GAE, explicitly distinguishing critic regression targets from actor advantage scores. **Third**, we conduct an extensive empirical evaluation: covering sparse-reward navigation, randomized environments, and dense-reward control. Beyond comparisons with PPO (Schulman et al., 2017) using GAE, we examine learning from rare successes in DeepSea, compatibility with exploration bonuses, and the effects of multi-step propagation, and critic ensembles. These experiments demonstrate gains across a broad range of tasks and clarify which components contribute to effective learning from successful experience.

2 Preliminaries

We consider a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_r, p, \gamma, \rho_0)$ with bounded rewards, $0 \le \gamma < 1$, and discrete or continuous actions. For a fixed policy π, V_π and Q_π denote its ordinary expected discounted returns. Online rollouts are fragments $\xi = (s_0, a_0, r_0, \dots, s_H)$ collected under the current behavior policy. When discussing population operators, we hold this policy fixed. Neither this convention nor the use of rollout fragments asserts arbitrary off-policy evaluation without correction.

GAE and critic targets. For frozen rollout values $V_h = V_\psi(s_h)$, GAE satisfies

$$\delta_h = r_h + \gamma \hat{V}_{h+1} - V_h, \qquad \hat{V}_h = V_h + \lambda_{\mathrm{GAE}}\delta_h, \qquad A_h^{\mathrm{GAE}} = \delta_h, \quad \hat{V}_H = V_\psi(s_H). \tag{1}$$

The usual critic regression target is $\hat{V}_h$ (Schulman et al., 2016). We ignored termination conditions.

Expectiles and IQL. For $0 < \tau < 1$, define $L_2^\tau(u) = |\tau - \mathbb{I}\{u < 0\}|u^2$ and

$$\mathrm{E}_\tau(X) = \arg\min_v \mathbb{E}[L_2^\tau(X - v)], \qquad \Delta_\tau(u) = \tau \max(u, 0) - (1-\tau)\max(-u, 0). \tag{2}$$

The expectile is characterized by $\mathbb{E}[\Delta_\tau(X - \mathrm{E}_\tau(X))] = 0$. At $\tau = 1/2$ it is the mean; for larger τ it gives greater influence to values above the fitted location. IQL regresses V toward an action-value expectile and Q toward $r + \gamma V(s')$ using squared error (Kostrikov et al., 2022). This construction motivates the population operators

Definition 1 (Implicit Bellman operators). *For a fixed policy π and $0 < \tau < 1$, define the conditional Bellman backup and the action-expectile map by*

$$\begin{aligned}(\mathcal{T}_Q V)(s,a) &= \mathbb{E}[r + \gamma V(S') \mid s, a], \\ (\mathcal{T}_V^\tau Q)(s) &= \mathrm{E}_\tau(Q(s,A)), \qquad A \sim \pi(\cdot \mid s).\end{aligned} \tag{3}$$

The first map averages over the original reward and transition laws; the second takes an expectile over actions under π.

Remark 1. While various papers have been using the value and construct estimators, its' convergent property in population has not been worked out.

3 From Implicit Values to Convergent Joint Updates

We first establish the recursive implicit-value target, identify the policy it evaluates, and establish an exact policy-improvement condition. We then show that coupled partial updates and their multi-step compositions converge to this same target.

3.1 Existence of Implicit Values

The action values inside an IQL expectile are themselves bootstrapped from the state values. Thus the target must solve both equations simultaneously.

Definition 2 (Recursive implicit values). *Using the operators in Definition 1, define $\mathcal{T}_\pi^\tau = \mathcal{T}_Q \circ \mathcal{T}_V^\tau : \mathcal{Q} \to \mathcal{Q}$. An implicit value pair for (π, τ) is a pair $(Q_\pi^\tau, V_\pi^\tau) \in \mathcal{Z}$ satisfying*

$$Q_\pi^\tau = \mathcal{T}_Q V_\pi^\tau, \qquad V_\pi^\tau = \mathcal{T}_V^\tau Q_\pi^\tau. \tag{4}$$

Equivalently, Q_π^τ is a fixed point of $\mathcal{T}_\pi^\tau$ and $V_\pi^\tau = \mathcal{T}_V^\tau Q_\pi^\tau$.

Theorem 3 (Existence, uniqueness, and exact value iteration). *The operator $\mathcal{T}_\pi^\tau$ is a γ-contraction. Consequently, Equation (4) has a unique bounded solution. For every $Q_0 \in \mathcal{Q}$, the iteration $Q_{k+1} = \mathcal{T}_\pi^\tau Q_k$ satisfies, for all $k \geq 0$,*

$$\|Q_k - Q_\pi^\tau\|_\infty \leq \gamma^k \|Q_0 - Q_\pi^\tau\|_\infty . \tag{5}$$

Proof sketch. Expectiles are nonexpansive in the supremum norm (Lemma 13), whereas $\mathcal{T}_Q$ is γ-Lipschitz. Their composition is therefore a γ-contraction on the complete space $\mathcal{Q}$. The contraction mapping theorem gives the unique fixed point and geometric rate; Equation (4) then determines V_π^τ uniquely. Appendix A.2 gives the details. □

Remark 2 (Relation to Q-learning convergence). This is the population value-iteration guarantee for the implicit Bellman backup: it plays the role of Bellman contraction in classical Q-learning analysis. As the Bellman contraction does not imply convergence of Deep Q learning, neither do our theorems do to our stochastic estimator.

3.2 Implicit Policy and Exact Policy Improvement

The expectile reweighting characterized in IDQL (Hansen-Estruch et al., 2023) identifies an action distribution. At the recursive fixed point, this distribution also identifies the continuation policy being evaluated.

Definition 4 (Induced implicit policy). *For the pair in Theorem 3, define*

$$\pi_\tau(da \mid s) = \frac{w_\tau(s,a)}{Z_\tau(s)} \pi(da \mid s), \qquad w_\tau(s,a) = \begin{cases} \tau, & Q_\pi^\tau(s,a) \geq V_\pi^\tau(s), \\ 1-\tau, & \textit{otherwise}, \end{cases} \tag{6}$$

where $Z_\tau(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[w_\tau(s,a)]$. Since $0 < \min\{\tau, 1-\tau\} \leq Z_\tau(s)$, this defines a probability kernel absolutely continuous with respect to π.

Theorem 5 (An attainable implicit-policy value)**.** *The recursive pair equals the ordinary expected-return value pair of π_τ:*

$$(Q_\pi^\tau, V_\pi^\tau) = (Q^{\pi_\tau}, V^{\pi_\tau}). \tag{7}$$

For fixed π and $0 < \tau_1 \le \tau_2 < 1$, $V_\pi^{\tau_1} \le V_\pi^{\tau_2}$ pointwise, and $V_\pi^{1/2} = V_\pi$. Thus $V^{\pi_\tau} \ge V_\pi$ whenever $1/2 \le \tau < 1$.

Proof sketch. Expectile stationarity gives $\mathbb{E}_\pi[w_\tau(Q_\pi^\tau - V_\pi^\tau)] = 0$. Dividing by Z_τ yields $V_\pi^\tau = \mathbb{E}_{\pi_\tau} Q_\pi^\tau$. Together with $Q_\pi^\tau = \mathcal{T}_Q V_\pi^\tau$, these are the ordinary Bellman equations for π_τ. Monotonicity follows by comparing the monotone contractions $\mathcal{T}_V^\tau \circ \mathcal{T}_Q$ across τ; at $\tau = 1/2$ the expectile is the mean. See Appendix A.3. □

Remark 3 (Value improvement before actor fitting). The upper-expectile target evaluates a policy at least as valuable as π before an explicit actor has been fitted to it. The improvement comes from reweighting actions; rewards and transitions retain their original distributions. This establishes an attainable value target without requiring binary success labels.

Definition 6 (Implicit-policy advantage)**.** *Define $A_\pi^\tau(s,a) = Q_\pi^\tau(s,a) - V_\pi^\tau(s)$. By Theorem 5, this is the ordinary advantage A^{π_τ} and satisfies $\mathbb{E}_{a\sim\pi_\tau(\cdot|s)} A_\pi^\tau(s,a) = 0$. A policy π^+ satisfies the implicit improvement condition if*

$$\mathbb{E}_{a\sim\pi^+(\cdot|s)} A_\pi^\tau(s,a) \ge 0 \quad \textit{for every } s. \tag{8}$$

Theorem 7 (Policy improvement from an exact implicit advantage)**.** *Let $1/2 \le \tau < 1$. Every stationary Markov policy π^+ satisfying Equation (8) obeys*

$$V^{\pi^+} \ge V^{\pi_\tau} = V_\pi^\tau \ge V^\pi. \tag{9}$$

Proof sketch. The performance-difference identity (Kakade & Langford, 2002) allows characterizing the improvement. See Appendix A.4 for a complete proof. □

Remark 4 (The advantage and the practical actor score). The guarantee concerns A_π^τ, the advantage of the implicit policy. The theorem justify the use of A_π^τ to guide policy improvement.

3.3 Coupled Partial Updates Preserve the Target

Exact value iteration solves an expectile subproblem at each step. We now replace that solution by a partial score update, while allowing both value functions to change.

Definition 8 (Joint λ-adjusted implicit Bellman operator)**.** *For $\lambda = (\lambda_q, \lambda_v)$ with $0 < \lambda_q \le 1$ and $\lambda_v > 0$, define $\mathcal{J}_\lambda^\tau : \mathcal{Z} \to \mathcal{Z}$ by $\mathcal{J}_\lambda^\tau(Q,V) = (Q^+, V^+)$, computed sequentially as*

$$Q^+ = (1-\lambda_q)Q + \lambda_q \mathcal{T}_Q V, \tag{10}$$

$$V^+(s) = V(s) + \lambda_v \mathbb{E}_{a\sim\pi(\cdot|s)}[\Delta_\tau(Q^+(s,a) - V(s))]. \tag{11}$$

The dependence on the fixed policy π is suppressed in the notation. The environmental expectation in $\mathcal{T}_Q V$ is taken before applying Δ_τ; the action expectation in Equation (11) remains outside the score.

In general $V^+ \ne \mathcal{T}_V^\tau Q^+$. Thus exact expectile minimization alone does not establish convergence of this joint map.

Theorem 9 (Target-preserving convergence of coupled partial updates)**.** *Let $m = \min\{\tau, 1-\tau\}$ and $M = \max\{\tau, 1-\tau\}$. If $0 < \lambda_q \le 1$ and $0 < \lambda_v \le 1/M$, then $\mathcal{J}_\lambda^\tau$ is a contraction in the joint norm, with factor at most*

$$\kappa = \max\{1 - \lambda_q(1-\gamma),\ 1 - \lambda_v m \lambda_q (1-\gamma)\} < 1. \tag{12}$$

For every admissible λ, its unique fixed point is the same pair $z_* = (Q_\pi^\tau, V_\pi^\tau)$ as in Equation (4). For every $z_0 \in \mathcal{Z}$, $z_{k+1} = \mathcal{J}_\lambda^\tau z_k$ satisfies $\|z_k - z_*\|_\infty \le \kappa^k \|z_0 - z_*\|_\infty$ for all $k \ge 0$.

Proof sketch. Positive update weights make joint stationarity equivalent to the two equations in (4). For two inputs at joint distance D, their Q^+ components differ by at most $\kappa_q D$, where $\kappa_q =$

$1-\lambda_q(1-\gamma)$. Every secant slope of Δ_τ lies in $[m, M]$. Writing its action average as $\bar{\alpha}(s)$, the state-value difference is bounded by

$$\big[1-\lambda_v\bar{\alpha}(s)+\lambda_v\bar{\alpha}(s)\kappa_q\big]D \le \big[1-\lambda_v m(1-\kappa_q)\big]D.$$

Here $1-\lambda_v\bar{\alpha}(s) \ge 0$ by the step-size restriction. The two component bounds give κ; completeness then gives convergence. Appendix A.5 supplies the full secant-slope argument. □

Remark 5 (Partial fitting, unchanged policy value). The expectile level τ determines the target. The admissible correction weights λ_q, λ_v determine how both critics approach it, without changing the implicit policy value. The theorem establishes the foundation for extending the algorithm to multi-step version.

4 Multi-Step Implicit Value Estimators

Coupled backward estimation. $\mathrm{IQ}(\tau,\lambda)$ uses frozen rollout predictions $Q_h = Q_\phi(s_h,a_h)$ and $V_h = V_\psi(s_h)$. Initialize $\bar{V}_H = V_H$, with true-terminal bootstrapping suppressed by c_h. For $h = H-1,\dots,0$, compute

$$\delta_h^q = r_h + c_h\bar{V}_{h+1} - Q_h, \qquad \bar{Q}_h = Q_h + \lambda_q\delta_h^q, \tag{13}$$

$$\delta_h^v = \bar{Q}_h - V_h, \qquad \bar{V}_h = V_h + \lambda_v\Delta_\tau(\delta_h^v), \qquad \widehat{A}_h = \delta_h^v\tau. \tag{14}$$

Here $\lambda = (\lambda_q,\lambda_v)$ in $\mathrm{IQ}(\tau,\lambda)$. The actor score $\widehat{A}_h$ is an estimate of the implicit policy advantage as introduce in Definition 6 with lagged value base for convenience. The state recursion propagates these asymmetric corrections to preceding steps.

A value-only specialization. Setting $\lambda_q = 1$ cancels Q_h in Equation (13). The resulting $\mathrm{IV}(\tau,\lambda)$ recursion is

$$\delta_h = r_h + c_h\bar{V}_{h+1} - V_h, \qquad \bar{V}_h = V_h + \lambda_v\Delta_\tau(\delta_h), \qquad \widehat{A}_h = \delta_h. \tag{15}$$

The identity holds algebraically in any MDP. Deterministic rewards and transitions motivate omitting Q, because a separate head is then unnecessary for averaging the immediate environmental outcome.

Critic and actor updates. We regress to detached rollout targets as in PPO,

$$\mathcal{L}_V = \mathbb{E}_h[(V_\psi(s_h) - \mathrm{sg}(\bar{V}_h))^2], \qquad \mathcal{L}_Q = \mathbb{E}_h[(Q_\phi(s_h,a_h) - \mathrm{sg}(\bar{Q}_h))^2], \tag{16}$$

with $\mathcal{L}_Q$ omitted for $\mathrm{IV}(\tau,\lambda)$. PPO's clipped surrogate uses the detached score $\widehat{A}_h$ with the behavior-policy likelihood ratio, plus the baseline entropy regularizer and advantage normalization (Schulman et al., 2017). Algorithm 1 specifies the update order. The induced implicit policy explains the value target; distinguishes the PPO surrogate from an exact implicit-value policy gradient.

Proposition 10 (Exact reductions). $\mathrm{IQ}(\tau,\lambda)$ *at* $\lambda_q = 1$ *gives exactly the* $\mathrm{IV}(\tau,\lambda)$ *state targets and actor scores. For* $\mathrm{IV}(\tau,\lambda)$ *with* $\tau = 1/2$ *and* $\lambda_v = 2\lambda_{\mathrm{GAE}}$, *and* $\lambda_v\Delta_\tau(\delta_h^v) = \lambda_{\mathrm{GAE}}\delta_h^v$.

$$\widehat{A}_h = \delta_h = A_h^{\mathrm{GAE}}, \qquad \bar{V}_h = V_h + \lambda_{\mathrm{GAE}}A_h^{\mathrm{GAE}}. \tag{17}$$

Thus the IV recovers GAE exactly.

See Appendix A.6 for a proof by calculation.

4.1 Convergence of Multi Step Operator

Corollary 11 (Multi-step joint compositions). *Under Theorem 9, define* $\mathcal{J}_{\lambda,H}^\tau = (\mathcal{J}_\lambda^\tau)^{\circ H}$ *for each integer* $H \ge 1$. *For all* $z, \widetilde{z} \in \mathcal{Z}$,

$$\|\mathcal{J}_{\lambda,H}^\tau z - \mathcal{J}_{\lambda,H}^\tau\widetilde{z}\|_\infty \le \kappa^H\|z-\widetilde{z}\|_\infty. \tag{18}$$

Its unique fixed point is z_*. *Repeated updates* $z_{k+1} = \mathcal{J}_{\lambda,H}^\tau z_k$ *satisfy* $\|z_k - z_*\|_\infty \le \kappa^{Hk}\|z_0 - z_*\|_\infty$.

Proof sketch. Apply the joint contraction H times. Since $\mathcal{J}_\lambda^\tau z_* = z_*$, its composition fixes z_*; strict contraction ensures uniqueness. Iterating the composition gives the stated rate. □

Remark 6 (Composition and frozen-anchor propagation). Corollary 11 updates both critics after each population step. Backward targets instead freeze the critics and apply nonlinear scores to sampled continuations, so their expected update is generally not $(\mathcal{J}_\lambda^\tau)^{\circ H}$. We leave a more complete theoretical characterization to the future, as they are not that insightful, and we perform sampling anything.

Overestimation and minimum ensembles. Asymmetry can reinforce positive approximation errors. We therefore use minimum predictions from an ensemble: $Q_{\min}(s,a) = \min_j Q_{\phi_j}(s,a)$ for the $\mathrm{IQ}(\tau,\lambda)$ action-value anchors and $V_{\min}(s) = \min_j V_{\psi_j}(s)$ for the $\mathrm{IV}(\tau,\lambda)$ state-value anchors and bootstrap. The resulting common targets supervise the individual heads. This follows the practical use of multiple critics to reduce overestimation (Fujimoto et al., 2018; An et al., 2021); it is neither a certified lower bound nor part of the exact contraction theorem. Larger ensembles can also introduce pessimism. We use $\tau = 0.6$, $\lambda_v = \lambda_{\mathrm{GAE}}/\tau$ with $\lambda_{\mathrm{GAE}} = 0.95$, and $\lambda_q = 0.9$ for $\mathrm{IQ}(\tau,\lambda)$ unless an ablation varies them. Ensemble size and the $\tau \to 1$ regime are examined empirically.

5 Experiment

Our experiments address four questions. **First**, can multi-step implicit value propagation better exploit successful trajectories, and does it complement exploration? **Second**, do its benefits extend beyond sparse-reward tasks to dense-reward settings? **Third**, under stochastic transitions, does explicitly modeling Q provide an advantage over the value-only formulation? **Finally**, can a small value ensemble control optimistic overestimation without becoming overly pessimistic? We study these questions using DeepSea, sparse- and dense-reward benchmarks, stochastic environments, and ensemble-size ablations. Our benchmarks are drawn from Navix (Pignatelli et al., 2024), JaxGCRL (Bortkiewicz et al., 2025), Pointax (Radji, 2025), Gymnax (Lange, 2022), Brax (Freeman et al., 2021), and Atari (Farebrother & Castro, 2024; Weng et al., 2022), with PPO as the baseline. Unless otherwise specified, we use $\tau = 0.6$, $N = 2$ for IV, $\lambda_q = 0.9$ for IQ, and $\lambda_v = \lambda_{\mathrm{GAE}}/\tau$ with $\lambda_{\mathrm{GAE}} = 0.95$. Additional implementation details and hyperparameter analyses are provided in Appendix F.

5.1 Combining Implicit Q with RND Exploration

The *DeepSea* environment (Osband et al., 2019) is an $N \times N$ long-horizon, sparse-reward task in which the agent receives a reward of 1 only after making all N binary decisions correctly. It therefore separates two challenges: discovering rare successful trajectories and propagating their reward back over a long horizon. We use DeepSea as a controlled testbed for studying these two effects.

We combine **IQ** with Random Network Distillation (RND) to test whether multi-step value propagation is complementary to intrinsic exploration. Figure 2 compares **PPO**, **PPO+RND**, **IQ**, **IQ+RND (1-step)**, and **IQ+RND** on $N \in \{16, 32, 48\}$..

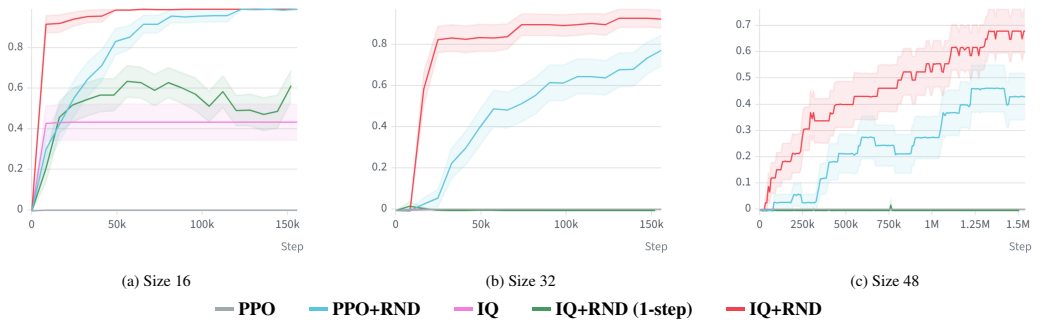


Figure 2: Combining Implicit Q (IQ) with RND exploration on DeepSea with $N \in \{16, 32, 48\}$. IQ+RND provides the strongest performance as the horizon increases: on size 48, it reaches approximately 0.65 return versus about 0.43 for PPO+RND, while IQ+RND (1-step) remains near zero. Curves report mean episodic return, and shaded regions indicate ± 1 standard deviation.

The advantage of IQ+RND becomes increasingly clear as the horizon grows. On DeepSea-16, IQ+RND reaches near-optimal return substantially faster than PPO+RND. On DeepSea-32, IQ+RND reaches approximately 0.9 return versus about 0.75 for PPO+RND, while IQ and IQ+RND (1-step) remain near zero. On DeepSea-48, IQ+RND reaches approximately 0.65, compared with roughly 0.43 for PPO+RND—about a $1.5\times$ higher final return—whereas the one-step variant again makes little progress. These results show that RND helps discover successful trajectories, while multi-step IQ is crucial for effectively propagating their rewards over long horizons.

5.2 More Sparse Reward

Figure 3 compares IV with PPO across ten sparse-reward environments. IV outperforms PPO on **7 out of 10 tasks**, with especially large gains on several long-horizon navigation problems. On DoorKey-8×8, IV reaches above 0.5 return while PPO remains near zero. On Empty-16×16, IV reaches approximately 0.9 versus about 0.5 for PPO, and on Ant Big Maze it achieves roughly 0.06 compared with less than 0.01 for PPO.

The gains are not universal: PPO performs slightly better on *Large Diverse GR* and *LavaGap*. We conjecture that diverse-goal tasks increase value interference, while the hard constraints in LavaGap may make optimistic updates more sensitive to estimation noise. Additional results and discussion are provided in Appendix F.4.

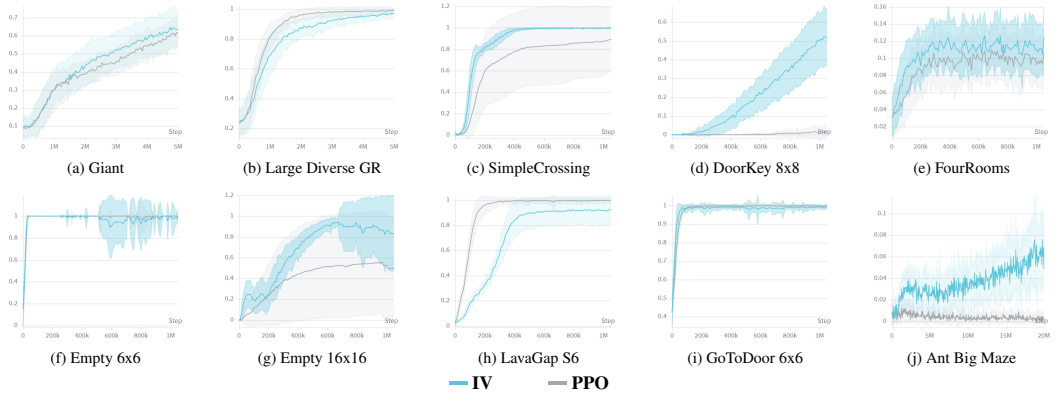


Figure 3: Performance comparison between IV and PPO across ten sparse-reward environments. IV outperforms PPO on 7 of 10 tasks, with especially large gains on DoorKey-8×8, Empty-16×16, and Ant Big Maze. Curves show the mean episodic return over 32 random seeds, with shaded regions indicating ± 1 standard deviation.

5.3 Stochastic MDPs

Figure 4 shows that explicitly modeling Q is beneficial under stochastic transitions. The largest gain appears on Atari-BeamRider, where IQ reaches roughly $5{,}800$ return, compared with about $4{,}200$ for IV and $2{,}200$ for PPO—more than $2.5\times$ the PPO return and about $1.4\times$ the IV return. On Breakout, Pong, Dynamic Obstacles, and Empty Random 8×8, the differences are smaller, but IQ still generally achieves slightly higher final returns or faster convergence. Overall, these results support explicit Q modeling for stochastic dynamics, with the benefit becoming especially pronounced on harder tasks.

5.4 Dense-Reward Environments

To test whether implicit value estimation also helps beyond sparse rewards, we evaluate IV on eight dense-reward environments. As shown in Figure 5, IV outperforms PPO on **7 out of 8 tasks**.

The gains are substantial on several environments. IV reaches roughly 70 return on Breakout-MinAtar versus about 38 for PPO ($1.8\times$), and about 45 on Humanoid U-Maze versus 15 for PPO ($3\times$). On Humanoid Big Maze, IV also improves the return from roughly -260 to -70. Gains are smaller on the remaining tasks, with one environment showing comparable performance. These results indicate that the benefit of optimistic implicit value estimation extends beyond sparse-reward credit assignment.

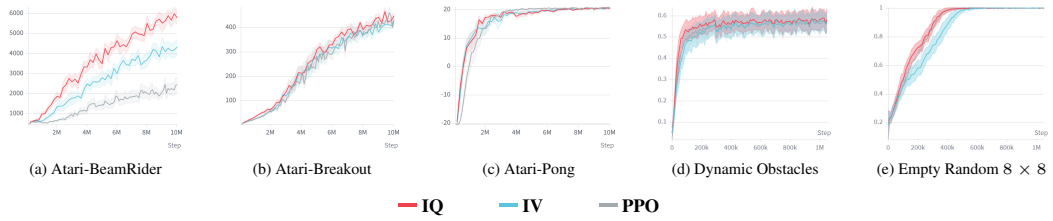


Figure 4: Performance comparison of IQ, IV, and PPO on five environments with stochastic transition dynamics. The largest gain appears on Atari-BeamRider, where IQ reaches roughly 5,800 return, more than 2.5× the PPO return. Curves show mean episodic return over 8 random seeds for Atari tasks and 32 random seeds for the remaining tasks; shaded regions indicate ±1 standard error.

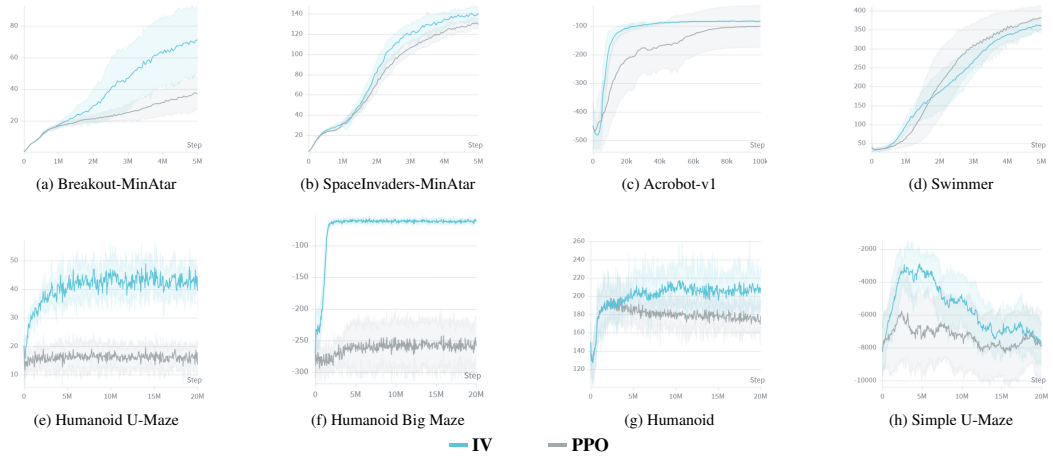


Figure 5: Performance comparison between IV and PPO on 8 dense-reward environments. IV outperforms PPO on 7 of 8 tasks, including approximately 1.8× higher return on Breakout-MinAtar and 3× higher return on Humanoid U-Maze. The top row (a–d) contains Gymnax tasks and the bottom row (e–h) contains JaxGCRL tasks. Curves show the mean return over 32 random seeds, with shaded regions indicating ±1 standard deviation.

5.5 Value-Ensemble Size

Figure 6 studies the effect of the number of value networks N. Larger ensembles are not uniformly better, and their effect depends on both the method and environment.

For PPO, increasing N provides little benefit and can hurt performance. For IV, a small ensemble can help: on Empty-16 × 16, $N = 2$ clearly outperforms $N = 1$, while larger ensembles become more conservative. On Ant Big Maze, however, larger ensembles perform better. IQ is comparatively less sensitive to N on Asterix-MinAtar and Dynamic Obstacles. Overall, these results support using a small ensemble to control optimistic overestimation without introducing excessive pessimism; we therefore use $N = 2$ in the main experiments.

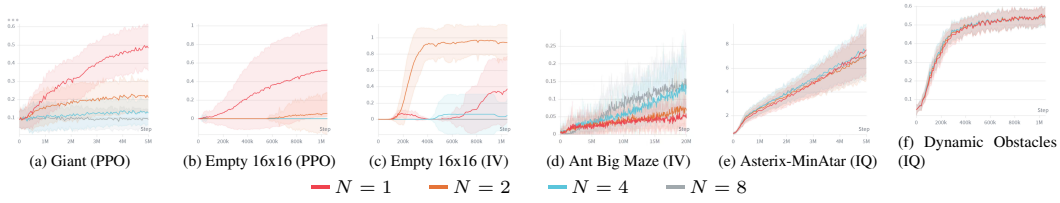


Figure 6: Ablation on the number of value networks N. Larger ensembles are not uniformly better: PPO often degrades as N increases, IV benefits from a small ensemble on some tasks but can become overly conservative for large N, while IQ is relatively insensitive on the two shown environments. Subfigures (a–b) show PPO, (c–d) IV, and (e–f) IQ. Shaded regions indicate ±1 standard deviation.

6 Related Work

Implicit values and expectile regression. IQL supplies the recursive expectile formulation and monotonicity/optimality-limit results (Kostrikov et al., 2022); IDQL characterizes the implicit actor of a value-loss minimizer for fixed Q (Hansen-Estruch et al., 2023). Our analysis separates these ingredients from convergence of the coupled partial-update dynamics. Uni-RL develops online and offline learning through an f-divergence-regularized Bellman system (Xu et al., 2025); its regularized objective is distinct from our ordinary-return fixed-point identity. ENQ applies an expectile loss to an n-step action-value error (Ghanem & Ghogho, 2026), rather than our coupled Q/V recursion. Appendix C compares the targets, update structures, and guarantees explicitly.

Multi-step learning and critic failure modes. Classical multi-step methods address propagation and off-policy mismatch through trace cutting, expected backups, or importance-weight truncation (Watkins, 1989; Peng & Williams, 1994; Precup et al., 2000; Munos et al., 2016; Espeholt et al., 2018). We change the value target toward an implicit improved policy rather than provide a new unbiased correction for arbitrary behavior–target mismatch. Self-imitation learning reuses an agent's past good decisions (Oh et al., 2018). Highway RL also uses selective multi-step propagation (Wang et al., 2024); it lacks theoretical guarantee as it is overestimating for MDP randomness. Bootstrapping can also affect interference and generalization under shared nonlinear function approximation (Bengio et al., 2020). Recent work on PPO value flattening studies insufficient value separation and modifies critic supervision (Li et al., 2026). It is closely related to our motivation, but changes which states supervise the critic rather than introducing our asymmetric target recursion.

Distributional Reinforcement Learning. The expectile value estimation is also used in the distributional setting (Bellemare et al., 2023), where the randomness is defined over MDPs. However, their focus is on being conservative w.r.t the environmental distribution rather than being optimistic. Quantile regression is another strategy for estimation values rather than the mean (Dabney et al., 2017).

Exploration and uncertainty. Intrinsic-reward methods such as RND promote the discovery of experience (Burda et al., 2018); our experiments combine exploration bonuses with altered value propagation. The expectile parameter controls action reweighting, not an exploration bonus or epistemic confidence radius. Unlike a return-distribution representation, the population construction applies its expectile to action values after conditional environmental averaging (Dabney et al., 2017; Kostrikov et al., 2022). These are complementary approaches to learning and exploration.

7 Conclusion and Limitations

Building on IQL, we introduced multi-step implicit value estimators $\mathrm{IQ}(\tau, \lambda)$ and $\mathrm{IV}(\tau, \lambda)$ for learning from successful experience through multi-step implicit value estimation, recovering GAE as a special case. Our theory connects optimistic value estimation to policy improvement under the original expected-return objective: coupled partial updates and expected backward propagation converge to a common, attainable implicit-policy value, and we establish a sufficient condition for its exact advantage to improve the current policy. This provides a principled foundation for combining optimism over action choice with multi-step credit assignment. Extensive experiments demonstrate improvements over PPO with GAE across sparse- and dense-reward tasks, with complementary gains from exploration bonuses. These findings highlight the continuation policy encoded in the value target as an important design choice: learning effectively from successful experience depends on both propagating rewards and valuing better future decisions.

AI use statement

In this work, we used generative AI tools for aiding and polish writing (sometimes rewrite by AI, and we re-edit, this include theorem and proofs), for retrieval and discovery related work and for drafting sections (the Frequently Asked Questions in Appendix and Ethics Statement are drafted by AI). We have not used generative AI tools for research ideation or execution, and for proving mathematical claims. and generating synthetic datasets are not applicable to this work. Additionally,

we used generative AI tools for re-organize code, and submit bash jobs. No new experimental results were generated through this manuscript-editing process. We have reviewed all AI-assisted work. We checked LLM-retrieved additional related work are all manually read and checked for relevance. LLM-generated code was verified and tested for correctness by one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies reinforcement learning in simulated benchmark environments. Implicit value estimates should not be interpreted as safety guarantees, calibrated uncertainty, or evidence supporting deployment in safety-critical settings. The method can overestimate values and can underweight unfavorable outcomes; deployment would require application-specific validation beyond the experiments reported here. However, intuitively, it is less dangerous than exploration bonus based method that could lure agents to explore uncertainty actions.

REPRODUCIBILITY STATEMENT

The rollout recursions, target losses, and update order are given in Section 4 and Algorithm 1. Appendix A provides proofs with explicit assumptions; Appendices E–F describe the reported setup and include task-wise curves. The implementation is available at `https://anonymous.4open.science/r/dig-rejax-2B47/README.md`. The appendix distinguishes documented global hyperparameters from per-environment settings inherited from the experiment configurations.

REFERENCES

Hongjoon Ahn, Heewoong Choi, Jisu Han, and Taesup Moon. Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In *Neural Information Processing Systems*, 2021.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. *ArXiv*, abs/1609.05140, 2016. URL https://arxiv.org/abs/1609.05140.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Fabio Bellini, Bernhard Klar, Alfred Müller, and Emanuela Rosazza Gianin. Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48, 2014. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2013.10.015. URL https://www.sciencedirect.com/science/article/pii/S0167668713001698.

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 767–777. PMLR, 2020.

Michał Bortkiewicz, Władek Pałucki, Vivek Myers, Tadeusz Dziarmaga, Tomasz Arczewski, Łukasz Kuciński, and Benjamin Eysenbach. Accelerating Goal-Conditioned RL Algorithms and Research. In *International Conference on Learning Representations*, 2025.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. *ArXiv*, abs/1810.12894, 2018. URL https://arxiv.org/abs/1810.12894.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2017.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Jesse Farebrother and Pablo Samuel Castro. CALE: Continuous arcade learning environment. *Advances in Neural Information Processing Systems*, 2024.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax: A differentiable physics engine for large scale rigid body simulation. http://github.com/google/brax, 2021. Version 0.14.1.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, 2018.

Abdelghani Ghanem and Mounir Ghogho. Upper-expectile multi-step q-learning for off-policy reinforcement learning, 2026. URL https://arxiv.org/abs/2608.02034. arXiv:2608.02034.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *ArXiv*, abs/2304.10573, 2023. URL https://arxiv.org/abs/2304.10573.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Robert Tjarko Lange. gymnax: A JAX-based reinforcement learning environment library. http://github.com/RobertTLange/gymnax, 2022. Version 0.0.4.

Seungjae Lee, Jigang Kim, Inkyu Jang, and H. Jin Kim. Dhrl: A graph-based approach for long-horizon and sparse hierarchical reinforcement learning. *ArXiv*, abs/2210.05150, 2022. URL https://arxiv.org/abs/2210.05150.

Yizhuo Li, Jianhao Yan, Yun Luo, Zhi Wang, Futing Wang, Rong-Xi Tan, Kanghui Tian, Ganqu Cui, Ning Ding, Peilin Zhao, Yafu Li, and Yu Cheng. Rethinking critic learning in PPO: Understanding and mitigating value flattening, 2026. URL https://arxiv.org/abs/2609.18708. arXiv:2609.18708.

Jarek Liesen, Chris Lu, and Robert Lange. rejax, 2024. URL https://github.com/keraJLi/rejax.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3878–3887, 2018.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes RL scalable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jing Peng and Ronald J Williams. Incremental multi-step q-learning. In *Machine Learning Proceedings 1994*, pp. 226–232. Elsevier, 1994.

Eduardo Pignatelli, Jarek Liesen, Robert Tjarko Lange, Chris Lu, Pablo Samuel Castro, and Laura Toni. Navix: Scaling minigrid environments with jax. *arXiv preprint arXiv:2407.19396*, 2024.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000.

Waris Radji. Pointax: Jax-native pointmaze environment, 2025. URL https://github.com/riiswa/pointax.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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

Wonchul Shin and Yusung Kim. Guide to control: Offline hierarchical reinforcement learning using subgoal generation for long-horizon and sparse-reward tasks. In *International Joint Conference on Artificial Intelligence*, 2023.

Ruosong Wang, Simon S. Du, Lin F. Yang, and Sham M. Kakakde. Is long horizon rl more difficult than short horizon rl? In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Yuhui Wang, Miroslav Strupl, Francesco Faccio, Qingyuan Wu, Haozhe Liu, Michal Grudzie'n, Xiaoyang Tan, and Jürgen Schmidhuber. Highway reinforcement learning. *ArXiv*, abs/2405.18289, 2024. URL https://arxiv.org/abs/2405.18289.

Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, 1989.

Jiayi Weng, Min Lin, Shengyi Huang, Bo Liu, Denys Makoviichuk, Viktor Makoviychuk, Zichen Liu, Yufan Song, Ting Luo, Yukun Jiang, Zhongwen Xu, and Shuicheng Yan. EnvPool: A highly parallel reinforcement learning environment execution engine. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22409–22421. Curran Associates, Inc., 2022.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Neural Information Processing Systems*, 2019.

Haoran Xu, Li Jiang, Jianxiong Li, Zhuoran Yang, Zhaoran Wang, Victor Wai Kin Chan, and Xianyuan Zhan. Offline RL with no OOD actions: In-sample learning via implicit value regularization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ueYYgo2pSSU.

Haoran Xu, Liyuan Mao, Hui Jin, Weinan Zhang, Xianyuan Zhan, and Amy Zhang. Uni-RL: Unifying online and offline RL via implicit value regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=mjZi8cV18P.

A Population Operators and Proofs

We use bounded measurable value functions and bounded rewards, with $0 \leq \gamma < 1$, a fixed Markov policy π, and $0 < \tau < 1$. Conditional expectations are assumed to be well defined and measurable. Finite state and action spaces suffice for all operator results. The explicit policy formula in Equation (6) also applies to continuous actions by reweighting the behavior-policy measure. Policy iteration in Appendix A.4 imposes the additional finite-space and action-access assumptions stated there.

A.1 Expectile Properties

Lemma 12 (Expectile properties (Bellini et al., 2014, Proposition 5(a), (c), and (f))). *Let $0 < \tau < 1$, and let X, Y be bounded random variables on the same probability space. The expectile is monotone and translation equivariant:*

$$X \leq Y \text{ almost surely} \implies \mathrm{E}_\tau(X) \leq \mathrm{E}_\tau(Y), \tag{19}$$

$$\mathrm{E}_\tau(X + b) = \mathrm{E}_\tau(X) + b, \qquad b \in \mathbb{R}. \tag{20}$$

Moreover, for $0 < \tau_1 \leq \tau_2 < 1$,

$$\mathrm{E}_{\tau_1}(X) \leq \mathrm{E}_{\tau_2}(X).$$

Lemma 13 (Nonexpansiveness). *For $0 < \tau < 1$ and bounded random variables X, Y on the same probability space,*

$$|\,\mathrm{E}_\tau(X) - \mathrm{E}_\tau(Y)| \leq \|X - Y\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the essential supremum norm.

Proof. Let $\epsilon = \|X - Y\|_\infty$. Then $Y - \epsilon \leq X \leq Y + \epsilon$ almost surely. By monotonicity and translation equivariance (Lemma 12),

$$\mathrm{E}_\tau(Y) - \epsilon = \mathrm{E}_\tau(Y - \epsilon) \leq \mathrm{E}_\tau(X) \leq \mathrm{E}_\tau(Y + \epsilon) = \mathrm{E}_\tau(Y) + \epsilon.$$

The claimed bound follows. □

A.2 Existence and Exact Iteration: Proof of Theorem 3

Theorem (3). The operator $\mathcal{T}_\pi^\tau$ is a γ-contraction. Consequently, Equation (4) has a unique bounded solution. For every $Q_0 \in \mathcal{Q}$, the iteration $Q_{k+1} = \mathcal{T}_\pi^\tau Q_k$ satisfies, for all $k \geq 0$,

$$\|Q_k - Q_\pi^\tau\|_\infty \leq \gamma^k \|Q_0 - Q_\pi^\tau\|_\infty . \tag{21}$$

Proof. For bounded Q_1, Q_2, Lemma 12 and Lemma 13 implies

$$\|\mathcal{T}_\pi^\tau Q_1 - \mathcal{T}_\pi^\tau Q_2\|_\infty \leq \gamma \|\mathcal{T}_V^\tau Q_1 - \mathcal{T}_V^\tau Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty . \tag{22}$$

where the first inequality comes from non-expansiveness and the second comes from monotonicity. The contraction mapping theorem gives a unique fixed point Q_π^τ and the rate in Equation (5). Set $V_\pi^\tau = \mathcal{T}_V^\tau Q_\pi^\tau$; then $Q_\pi^\tau = \mathcal{T}_Q V_\pi^\tau$. Conversely, any pair solving the two IQL equations has an action-value component fixed by $\mathcal{T}_\pi^\tau$, so it must equal this pair. □

A.3 Policy Interpretation: Proof of Theorem 5

Theorem (5). The recursive pair equals the ordinary expected-return value pair of π_τ:

$$(Q_\pi^\tau, V_\pi^\tau) = (Q^{\pi_\tau}, V^{\pi_\tau}). \tag{23}$$

For fixed π and $0 < \tau_1 \leq \tau_2 < 1$, $V_\pi^{\tau_1} \leq V_\pi^{\tau_2}$ pointwise, and $V_\pi^{1/2} = V_\pi$. Thus $V^{\pi_\tau} \geq V_\pi$ whenever $1/2 \leq \tau < 1$.

Proof. Let $A_\tau = Q_\pi^\tau - V_\pi^\tau$ for the pair established in Theorem 3. The expectile root condition gives

$$0 = \mathbb{E}_{a\sim\pi}[\Delta_\tau(A_\tau(s,a))] = \mathbb{E}_{a\sim\pi}[w_\tau(s,a)(Q_\pi^\tau(s,a) - V_\pi^\tau(s))]. \tag{24}$$

Since $Z_\tau(s) \geq \min\{\tau, 1-\tau\} > 0$, Equation (6) defines a probability kernel, and division by $Z_\tau(s)$ yields

$$V_\pi^\tau(s) = \mathbb{E}_{a\sim\pi_\tau(\cdot|s)} Q_\pi^\tau(s,a). \tag{25}$$

Together with $Q_\pi^\tau = \mathcal{T}_Q V_\pi^\tau$, this is the ordinary Bellman system for π_τ in the original MDP. Its solution is unique, so $V_\pi^\tau = V^{\pi_\tau}$ and $Q_\pi^\tau = Q^{\pi_\tau}$. The construction uses the expectile action weights characterized in IDQL (Hansen-Estruch et al., 2023); Bellman consistency at the recursive fixed point gives the policy-value identity.

For monotonicity, write $F_\tau = \mathcal{T}_V^\tau \mathcal{T}_Q$, a monotone γ-contraction on state values. For $\tau_1 \leq \tau_2$, Lemma 13 gives $F_{\tau_1} V \leq F_{\tau_2} V$. Iterating F_{τ_2} from the fixed point of F_{τ_1} gives a nondecreasing sequence converging to $V_\pi^{\tau_2}$, so $V_\pi^{\tau_1} \leq V_\pi^{\tau_2}$. At $\tau = 1/2$, the expectile is the action mean and the equations are those for π. This recovers the monotonicity property studied in IQL (Kostrikov et al., 2022). □

Remark 7 (A value pair is unique; a representing policy need not be). Theorem 3 establishes uniqueness of (Q_π^τ, V_π^τ), not uniqueness of every policy with that value. Equation (6) gives one explicit representation with support contained in that of π. No separate policy extraction is needed to define or iterate toward the value pair.

A.4 Policy Improvement: roof of Theorem 7

Theorem (7). Let $1/2 \leq \tau < 1$. Every stationary Markov policy π^+ satisfying Equation (8) obeys

$$V^{\pi^+} \geq V^{\pi_\tau} = V_\pi^\tau \geq V^\pi. \tag{26}$$

Proof. Write $T^\mu V = \mathbb{E}_{a\sim\mu(\cdot|s)} \mathcal{T}_Q V$. Since $Q_\pi^\tau = \mathcal{T}_Q V_\pi^\tau$, Equation (8) gives

$$T^{\pi^+} V_\pi^\tau - V_\pi^\tau = \mathbb{E}_{a\sim\pi^+(\cdot|s)} A_\pi^\tau(s,a) \geq 0.$$

Monotone iteration of the γ-contraction T^{π^+} therefore yields $V^{\pi^+} \geq V_\pi^\tau$. Theorem 5 gives $V_\pi^\tau = V^{\pi_\tau}$. Finally, since $\tau \geq 1/2$, the expectile dominates the mean, so $V_\pi^\tau \geq T^\pi V_\pi^\tau$; monotone iteration of T^π yields $V_\pi^\tau \geq V^\pi$. □

A.5 Coupled Partial Updates: Proof of Theorem 9

Theorem 14 (9). *Let $m = \min\{\tau, 1-\tau\}$ and $M = \max\{\tau, 1-\tau\}$. If $0 < \lambda_q \leq 1$ and $0 < \lambda_v \leq 1/M$, then $\mathcal{J}_\lambda^\tau$ is a contraction in the joint norm, with factor at most*

$$\kappa = \max\{1 - \lambda_q(1-\gamma),\ 1 - \lambda_v m \lambda_q (1-\gamma)\} < 1. \tag{27}$$

For every admissible λ, its unique fixed point is the same pair $z_* = (Q_\pi^\tau, V_\pi^\tau)$ as in Equation (4). For every $z_0 \in \mathcal{Z}$, $z_{k+1} = \mathcal{J}_\lambda^\tau z_k$ satisfies $\|z_k - z_*\|_\infty \leq \kappa^k \|z_0 - z_*\|_\infty$ for all $k \geq 0$.

Proof. Let $m = \min\{\tau, 1-\tau\}$ and $M = \max\{\tau, 1-\tau\}$.

Since $\lambda_q, \lambda_v > 0$, the fixed-point equations in Equations (10)–(11) are equivalent to

$$Q = \mathcal{T}_Q V, \qquad \mathbb{E}_{a\sim\pi(\cdot|s)}[\Delta_\tau(Q(s,a) - V(s))] = 0 \quad \text{for every } s.$$

By expectile stationarity, the second equation is equivalent to $V = \mathcal{T}_V^\tau Q$. Theorem 3 therefore identifies the unique fixed point as $z_* = (Q_\pi^\tau, V_\pi^\tau)$.

To prove contraction, let $z_i = (Q_i, V_i)$ and $(Q_i^+, V_i^+) = \mathcal{J}_\lambda^\tau z_i$, and write

$$D = \|z_1 - z_2\|_\infty = \max\{\|Q_1 - Q_2\|_\infty, \|V_1 - V_2\|_\infty\}, \qquad \kappa_q = 1 - \lambda_q(1-\gamma).$$

Since $0 < \lambda_q \leq 1$ and $\mathcal{T}_Q$ is γ-Lipschitz, Equation (10) gives

$$\begin{aligned} \|Q_1^+ - Q_2^+\|_\infty &\leq (1-\lambda_q)\|Q_1 - Q_2\|_\infty + \lambda_q \gamma \|V_1 - V_2\|_\infty \\ &\leq \kappa_q D. \end{aligned} \tag{28}$$

The function Δ_τ is continuous and piecewise linear with slopes τ and $1-\tau$. Every secant slope therefore lies in $[m, M]$, including when its endpoints straddle zero. Consequently, for each (s, a) there is $\alpha(s, a) \in [m, M]$ such that

$$\begin{aligned}&\Delta_\tau(Q_1^+(s,a) - V_1(s)) - \Delta_\tau(Q_2^+(s,a) - V_2(s)) \\ &\quad = \alpha(s,a)\big[Q_1^+(s,a) - Q_2^+(s,a) - (V_1(s) - V_2(s))\big].\end{aligned}$$

All action expectations below are under $\pi(\cdot \mid s)$. Define $\bar{\alpha}(s) = \mathbb{E}_{a\sim\pi}[\alpha(s,a)] \in [m, M]$. Substituting into Equation (11) yields

$$\begin{aligned}V_1^+(s) - V_2^+(s) = (1 - \lambda_v\bar{\alpha}(s))(V_1(s) - V_2(s)) \\ + \lambda_v\mathbb{E}_{a\sim\pi}\big[\alpha(s,a)(Q_1^+(s,a) - Q_2^+(s,a))\big].\end{aligned}$$

The restriction $\lambda_v \le 1/M$ ensures $1 - \lambda_v\bar{\alpha}(s) \ge 0$. Hence

$$\begin{aligned}|V_1^+(s) - V_2^+(s)| &\le (1 - \lambda_v\bar{\alpha}(s))\,\|V_1 - V_2\|_\infty + \lambda_v\bar{\alpha}(s)\,\big\|Q_1^+ - Q_2^+\big\|_\infty \\ &\le \big[1 - \lambda_v\bar{\alpha}(s) + \lambda_v\bar{\alpha}(s)\kappa_q\big]D \\ &= \big[1 - \lambda_v\bar{\alpha}(s)(1-\kappa_q)\big]D \\ &\le \big[1 - \lambda_v m\lambda_q(1-\gamma)\big]D.\end{aligned}$$

Taking the supremum over s and combining with Equation (28), we obtain

$$\begin{aligned}\|\mathcal{J}_\lambda^\tau z_1 - \mathcal{J}_\lambda^\tau z_2\|_\infty &= \max\{\big\|Q_1^+ - Q_2^+\big\|_\infty, \big\|V_1^+ - V_2^+\big\|_\infty\} \\ &\le \max\{\kappa_q,\, 1 - \lambda_v m\lambda_q(1-\gamma)\}D \\ &= \kappa\,\|z_1 - z_2\|_\infty,\end{aligned}$$

with κ as in Equation (12). Because $\lambda_q, \lambda_v, m > 0$ and $\gamma < 1$, both coefficients are strictly less than one. Applying this bound repeatedly with $z_2 = z_*$ gives

$$\|z_k - z_*\|_\infty \le \kappa^k\,\|z_0 - z_*\|_\infty,$$

which proves geometric convergence to the unique fixed point. □

A.6 Exact Reduction: Proof of Proposition 10

Proposition (10). IQ(τ, λ) at $\lambda_q = 1$ gives exactly the IV(τ, λ) state targets and actor scores. For IV(τ, λ) with $\tau = 1/2$ and $\lambda_v = 2\lambda_{\text{GAE}}$, and $\lambda_v\Delta_\tau(\delta_h^v) = \lambda_{\text{GAE}}\delta_h^v$.

$$\widehat{A}_h = \delta_h = A_h^{\text{GAE}}, \qquad \bar{V}_h = V_h + \lambda_{\text{GAE}}A_h^{\text{GAE}}. \tag{29}$$

Thus the IV recovers GAE exactly.

Proof. At $\lambda_q = 1$, the IQ(τ, λ) action-value update becomes

$$\bar{Q}_h = Q_h + (r_h + c_h\bar{V}_{h+1} - Q_h) = r_h + c_h\bar{V}_{h+1}. \tag{30}$$

Consequently, its state-value residual is

$$\delta_h^v = \bar{Q}_h - V_h = r_h + c_h\bar{V}_{h+1} - V_h, \tag{31}$$

which is precisely the IV(τ, λ) residual δ_h whenever the next-step propagated values agree. Both methods then apply the same state update and actor-score transformation,

$$\bar{V}_h = V_h + \lambda_v\Delta_\tau(\delta_h), \qquad \widehat{A}_h = \tau^{-1}\Delta_\tau(\delta_h). \tag{32}$$

Starting from their common boundary condition $\bar{V}_H = V_H$, backward induction therefore establishes equality of all state targets and actor scores. The action-value head no longer enters this recursion.

For the second reduction, $\Delta_{1/2}(u) = u/2$ and $\lambda_v = 2\lambda_{\text{GAE}}$ give

$$\bar{V}_h = V_h + \lambda_{\text{GAE}}\delta_h, \qquad \widehat{A}_h = \delta_h. \tag{33}$$

Set $\delta_H = 0$, consistently with $\bar{V}_H = V_H$, and define the frozen one-step TD residual

$$e_h = r_h + c_h V_{h+1} - V_h. \tag{34}$$

Substitution into the $\mathrm{IV}(\tau, \lambda)$ residual yields, for $h = H-1, \ldots, 0$,

$$\delta_h = r_h + c_h \bar{V}_{h+1} - V_h \tag{35}$$

$$= e_h + c_h \lambda_{\mathrm{GAE}} \delta_{h+1}. \tag{36}$$

GAE satisfies the same backward recursion,

$$A_h^{\mathrm{GAE}} = e_h + c_h \lambda_{\mathrm{GAE}} A_{h+1}^{\mathrm{GAE}}, \qquad A_H^{\mathrm{GAE}} = 0. \tag{37}$$

Backward induction therefore gives $\delta_h = A_h^{\mathrm{GAE}}$ for every h. Substituting this identity into the actor score and propagated value proves Equation (17). □

A.7 Proof of Corollary 11

Proof. Write $\mathcal{J} = \mathcal{J}_\lambda^\tau$. Applying Theorem 9 successively gives

$$\|\mathcal{J}^{\circ H} z - \mathcal{J}^{\circ H} \widetilde{z}\|_\infty \le \kappa^H \|z - \widetilde{z}\|_\infty.$$

Since $\mathcal{J} z_* = z_*$, its composition also fixes z_*. Strict contraction gives uniqueness, and iteration gives $\|z_k - z_*\|_\infty \le \kappa^{Hk} \|z_0 - z_*\|_\infty$. □

B FAQ: Interpreting Implicit Values and Expected Multi-Step Updates

This FAQ records the expectation convention used in the theory and distinguishes it from common misreadings of nonlinear value updates.

Q1. Does "learning from successes" restrict the method to sparse rewards? No. "Successes" refers to relatively high-return experience, not necessarily a binary terminal event. Neither the operators nor the practical recursions require a success indicator, a threshold, or sparse rewards. Sparse-reward tasks make the motivation particularly visible, but the experimental evaluation also includes dense-reward tasks. The abstraction of higher-valued actions applies in both cases.

Q2. What does convergence "in expectation" mean here? It means convergence of the explicitly defined expected operators in the infinite-data, exact-evaluation setting. The behavior policy is held fixed. At each depth, the action-value backup uses its conditional expectation, and the state-value update uses the expected action-level expectile score:

$$Q^+(s,a) = (1-\lambda_q) Q(s,a) + \lambda_q \mathbb{E}[r + \gamma V(S') \mid s, a], \tag{38}$$

$$V^+(s) = V(s) + \lambda_v \mathbb{E}_{A \sim \pi(\cdot|s)}[\Delta_\tau(Q^+(s,A) - V(s))]. \tag{39}$$

These are Equations (10)–(11), not a new operator. Corollary 11 composes this expected joint operator.

Q3. Does the proof require $\mathbb{E}[\Delta_\tau(X)] = \Delta_\tau(\mathbb{E}X)$? No. Neither the one-step theorem nor either multi-step result invokes that equality. In fact, this is exactly what has made proving convergence of expectile update trick. Theorem 9 relies on the stationary property of $\mathbb{E}[\Delta_\tau(Q - V)] = 0$

Q4. Why does one-step contraction imply multi-step contraction? If $\|\mathcal{J} z - \mathcal{J} \widetilde{z}\|_\infty \le \kappa \|z - \widetilde{z}\|_\infty$, then applying the same inequality H times gives $\|\mathcal{J}^{\circ H} z - \mathcal{J}^{\circ H} \widetilde{z}\|_\infty \le \kappa^H \|z - \widetilde{z}\|_\infty$. Also, $\mathcal{J} z_* = z_*$ implies $\mathcal{J}^{\circ H} z_* = z_*$. Since $\kappa^H < 1$, this fixed point is unique and repeated H-step updates converge to it. This is precisely Corollary 11. No additional interchange of expectations is involved.

Q5. In what sense are implicit values ordinary, unbiased policy values? The fixed point is exactly $(Q^{\pi_\tau}, V^{\pi_\tau})$ for the implicit policy defined in Equation (6), under the original reward and transition laws. It is not generally (Q^π, V^π) for the behavior policy. Thus the higher value at an upper expectile reflects a different action distribution, not a biased evaluation of the same policy or a favorable replacement of environmental outcomes. Within their admissible ranges, the correction weights and multi-step depth change the evaluation dynamics without changing this fixed point. "Implicit value" names the quantity being evaluated; "expectile" names its construction, not an exploration bonus.

Q6. How should an infinite-data or sampled-update argument be checked? First identify the precise conditional quantities estimated by the update. Infinite data removes estimation error when those quantities are estimated consistently; it does not authorize reordering conditional expectations and nonlinear operations. For example, averaging scores of noisy environmental targets and applying the score to an estimated conditional action value are different constructions. The expected-operator theorems use the ordering displayed in Equations (38)–(39). For an approximate implementation of the stated operator, Proposition bounds the effect of operator-evaluation error and proves convergence when that error vanishes. A claim about a different estimator should be analyzed for its own expected map, rather than treated as a counterexample to these operator theorems.

C FAQ: Theoretical Contribution and Relationship to Prior Work

This FAQ distinguishes existing ingredients from the results needed to justify the proposed update structure. The relevant comparison is between mathematical statements and algorithms, not between shared terms such as "expectile," "implicit policy," or "multi-step."

Q1. Which ingredients already come from IQL and IDQL? IQL defines the recursive expectile value equations and analyzes monotonicity and the dataset-supported optimality limit as $\tau \to 1$ (Kostrikov et al., 2022, Section 4.4). IDQL characterizes the implicit action weights associated with a value-loss minimizer for a given Q (Hansen-Estruch et al., 2023, Theorem 4.1). However, they did not establish the existence of implicit value as what we have for tabular Q learning. We use and build on these ingredients explicitly. Theorems 3–5 establish the recursive target and connect it to ordinary policy evaluation. Theorem 9 then addresses a different question: whether the specified coupled *partial* updates reach that same target. The contribution is not a new definition of an expectile or a claim that implicit actors were previously unknown, but filling a gap in previous work, and build foundation for multi-step extension. We believe both that Theorems 3 and Theorem 9 are nontrivial.

Q2. Why does a fixed-Q implicit-actor result not settle the joint-update theorem? Given Q, a characterization of $\arg\min_V \mathbb{E}[L_2^\tau(Q - V)]$ identifies the destination of one regression problem. In Equations (10)–(11), however, Q is changed using V, and V is only partially corrected using the newly updated Q^+. Generally $V^+ \neq \mathcal{T}_V^\tau Q^+$. Even identifying a common stationary point does not show that the iterates approach it. Theorem 9 proves both fixed-point preservation and contraction of the coupled map from arbitrary bounded initial pairs, with explicit admissible correction weights. This is a comparison of guarantees, not a claim to invent partial gradient updates, which IQL already uses. The coupled convergence statement does not follow merely by rewriting an expectile as a weighted mean. We shall also point out that the convergence in expectation does not rely on λ.

Q3. IQL already discusses multi-step dynamic programming. What is different here? In IQL, that phrase refers to repeated dynamic-programming updates rather than a single policy-evaluation/improvement pass (Kostrikov et al., 2022, Section 2). Our estimator instead constructs depth-H backward targets with coupled λ_q and λ_v corrections.

Q4. Does online implicit-value learning in Uni-RL subsume these results? Uni-RL formulates an f-divergence-regularized control objective and corresponding behavior-regularized Bellman equations (Xu et al., 2025, Section 3.1). However, Uni-RL itself neither states nor proves a contraction theorem for its online value updates. Although the earlier IVR (Xu et al., 2023) theory it builds on includes a regularized Bellman contraction result, this does not directly establish our results for the expectile operator. We prove existence and uniqueness of the recursive implicit value, identify it as an ordinary expected-return value in the original MDP, and establish convergence of the sequential map in Equations (10)–(11) to that same target under explicit parameter conditions. Uni-RL does not establish these operator-specific results. We make no claim to the first online use of implicit values.

Q5. Is the composition bound itself the main theoretical contribution? No. Once Theorem 9 has established contraction and the common fixed point, Corollary 11 follows by induction. Its role is to complete the logical connection to multi-step propagation. Counting the composition corollary as an independent major advance would obscure this structure. However, multi-step propagation is necessary for long horizon tasks that suffer from such value learning interference.

Q6. What does the theory contribute to the algorithmic design? It separates *which value is learned* from *how that value is propagated*. The expectile level τ determines the implicit-value target. Under the stated assumptions, changing λ_q, λ_v, or the expected propagation depth changes the evaluation dynamics without changing that target. Thus partial fitting and longer propagation need not trade the intended fixed point for a different one. The analysis also identifies the role of the separate Q function in environmental averaging and connects the value-only construction to GAE. These are design consequences of the coupled formulation, not guarantees of arbitrary neural-network optimization or actor improvement.

D Further Related Works

Horizon reduction. is another strategy for learning over a long horizon. (Bacon et al., 2016) pioneered the Option-Critic architecture, which compresses long sequential decisions into fewer high-level steps. A series of algorithms have emerged, (Ahn et al., 2025) identifies the high-level policy as the bottleneck and further compresses high-level planning steps, reducing the effective horizon and improving long-term planning. (Xie et al., 2019) addresses the exponential variance explosion under long horizons, which recursively estimates state marginal distributions to achieve optimal horizon dependence in mean-squared error. (Wang et al., 2020) takes a geometric perspective, leveraging the inherent quasimetric structure of optimal value functions. In addition, (Park et al., 2025) empirically validates that a long horizon is a primary scalability barrier in offline RL and actively compresses the horizon to unlock performance gains on large-scale datasets. However, horizon reduction has not yet been successfully implemented in the online reinforcement learning setting to our best knowledge. In any case, it is complementary to our motivation.

Critic-free LLM Post-Training. Related concerns make critic-free updates attractive in outcome-supervised large language model (LLM) reasoning. DeepSeekMath identifies both the difficulty of learning token-level values from end-of-response rewards and the computational cost of a separate critic; its Group Relative Policy Optimization (GRPO) replaces the learned baseline with a baseline computed from grouped sampled rewards (Shao et al., 2024; DeepSeek-AI, 2025). These examples motivate revisiting value estimation rather than assuming that a learned critic is always helpful. However, in MDPs with stochastic transitions that are outside the control of agents, maintaining a critic is still essential.

Algorithm 1 PPO-style training with $\mathrm{IQ}(\tau, \lambda)$ or $\mathrm{IV}(\tau, \lambda)$ targets

Require: Actor π_θ; critic networks; τ, λ_v; λ_q for $\mathrm{IQ}(\tau, \lambda)$

for each online rollout iteration **do**
  Freeze behavior actor π_{old} and rollout critic predictions
  Collect fragments; retain final observations and distinguish termination from truncation
  Split fragments at resets; use zero bootstrap for true terminals and a value bootstrap for truncations
  Select rollout anchors: minimum Q heads for $\mathrm{IQ}(\tau, \lambda)$, or minimum V heads for $\mathrm{IV}(\tau, \lambda)$
  Compute detached $\bar{Q}, \bar{V}, \widehat{A}$ backwards using Equations (13)–(14) or (15)
  for each optimization epoch and minibatch **do**
    Form $\rho_h(\theta) = \pi_\theta(a_h \mid s_h)/\pi_{\mathrm{old}}(a_h \mid s_h)$
    Apply the baseline advantage normalization to the stored actor scores
    Maximize $\mathbb{E}_h[\min\{\rho_h \widehat{A}_h, \mathrm{clip}(\rho_h, 1-\epsilon, 1+\epsilon)\widehat{A}_h\}]$ plus entropy regularization
    Fit each critic head to the corresponding detached targets using Equation (16)
  end for
end for

E Experimental Setup

Online training and baseline configurations. The experiments use the JAX/Rejax implementation (Liesen et al., 2024). For Gymnax and Navix, the recorded protocol uses Rejax's per-environment settings for parallel environments, rollout length, learning rate, training budget, and discount. JaxGCRL tasks follow their original configurations. Pointax uses the same set of per-environment

parameters tuned with the Rejax tuning scripts. The global values documented for the primary experiments are in Table 1. Environment-specific numeric settings are not reproduced in this table; exact replication additionally requires the experiment configurations. The one-step and hybrid comparisons refer to the implementations used for the corresponding plotted runs.

Table 1: Documented global settings. Per-environment choices and ablation overrides are separate. Excluding Atari

Setting	Value
Advantage normalization	Enabled
PPO clipping threshold	0.2
GAE reference parameter	0.95
Value-loss coefficient	0.5
Optimization epochs	16
Default expectile	$\tau = 0.6$
State correction weight	$\lambda_v = 0.95/\tau$
IQ action-value correction weight	$\lambda_q = 0.9$
Default IV and IQ value-head ensemble	$N = 2$
Actor–critic component ablation	$N = 1, \tau = 0.6$

For Atari games, we follow the CleanRL xla jax PPO implementation Huang et al. (2022), including network sharing and we use ensemble $n = 2$, $\lambda_q = 0.99$ and $\tau = 0.55$. We suspect that for much longer steps, setting $\lambda_q = 0.9$ might slow down the value propagation too much, and $\tau = 0.6$ might introduce too much maximization bias.

Reported statistics. Primary learning curves summarize 32 random seeds; bands denote one standard deviation. These are descriptions of between-seed variability, not confidence intervals or significance tests. The plots retain the training horizons of their respective experiments. No aggregate score is formed by averaging raw returns across tasks with different reward scales. Comparing final performance, learning speed, and area under the curve would require explicitly choosing the estimand and reporting uncertainty for it.

Compute. The reported setup uses an NVIDIA A800 80GB PCIe GPU with CUDA 12.2 and vectorized JAX environments. The recorded approximate run durations are 2–5 minutes per Gymnax environment, 3–5 minutes per Navix environment, 5–10 minutes per Pointax environment, and 30–60 minutes per JaxGCRL environment. These ranges are descriptive, not controlled throughput comparisons or a complete accounting of tuning cost.

F Complete Experimental Curves

The following figures provide the complete primary comparison groups and the reported diagnostic studies. The primary count of 22 refers to ten sparse-reward, four randomized, and eight dense-reward settings. DeepSea, parameter sweeps, and the additional navigation tasks are not included in that count.

F.1 Sensitivity of Expectile τ

To study the effect of the expectile parameter τ in IV, we vary $\tau \in \{0.6, 0.7, 0.8, 0.9\}$ on four environments (Figure 7). In general $\tau = 0.6$.

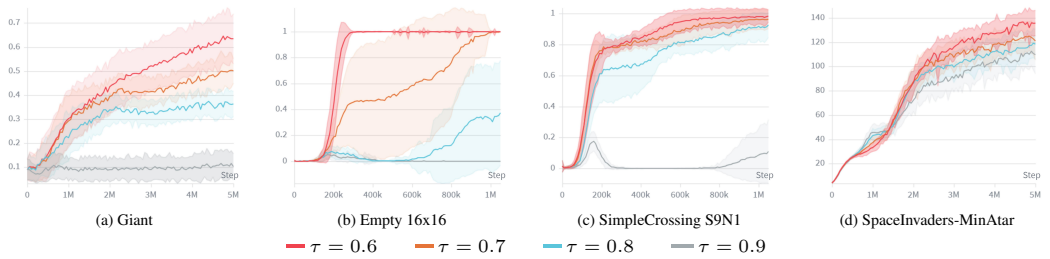


Figure 7: Sensitivity of IV to the expectile parameter τ on four representative environments. Curves show the mean episodic return over 32 random seeds with shaded regions indicating ± 1 standard deviation, plotted against environment steps.

F.2 Sensitivity Analysis of Target Entropy Tuning

We evaluated the sensitivity of the IV algorithm to the entropy coefficient across the Giant and Empty 16x16 environments. As illustrated in Figure 8, the performance remains remarkably consistent across the tested range of coefficients, indicating that our optimistic value estimation framework is robust to the specific choice of entropy regularization.

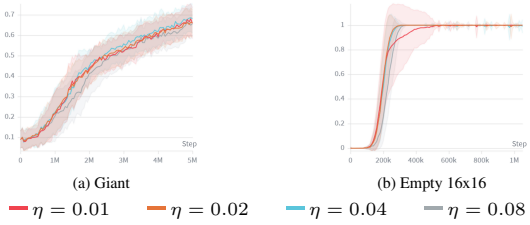


Figure 8: Sensitivity Analysis of Target Entropy . The results show that performance is largely robust to the entropy coefficient η in both environments.

F.3 Sensitivity Analysis of λ_q

In Figure 9, we can see that the Implicit Q is relatively robust w.r.t. the λ_q, with 0.9 being the better value.

F.4 Additional Experimental Results

Figure 10 presents results for 10 additional **Navix** and **Pointax** environments. We additionally report MinAtar results in Figure 11 as supplementary evidence for the stochastic MDP experiments in the main text.

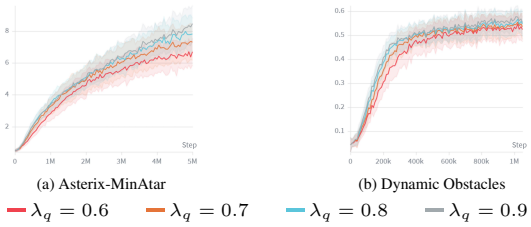


Figure 9: Sensitivity analysis of the multi-step parameter λ_q. The plots illustrate how the weighting of multi-step returns affects the stability and convergence speed of the IQ algorithm in stochastic and complex environments.

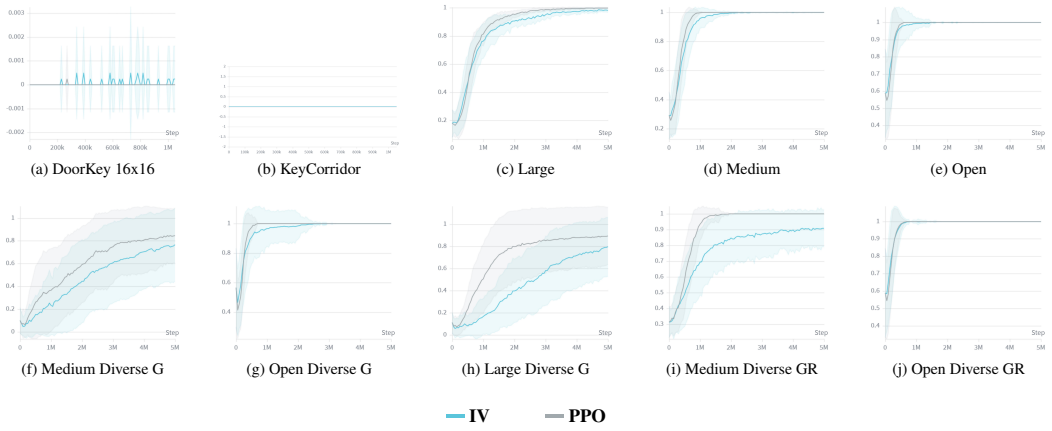


Figure 10: Experimental results for the remaining environments. Subfigures (a)–(b) represent tasks from the **Navix** suite, while (c)–(j) are from **Pointax**. These curves showcase performance in environments where both algorithms achieve comparable outcomes due to task simplicity or environmental constraints.

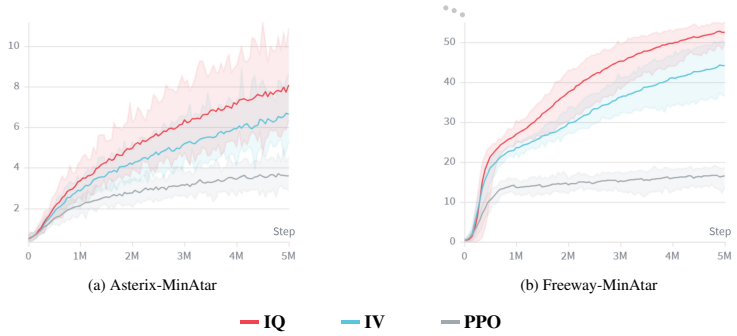


Figure 11: Additional stochastic MDP results on two MinAtar environments. These experiments complement the stochastic-transition results reported in the main text. Curves show the mean episodic return over 32 random seeds, with shaded regions indicating ± 1 standard deviation.

Additional Stochastic MDP Results. To complement the stochastic MDP evaluation in the main text, we report additional results on two MinAtar environments. These experiments provide further evidence for comparing **IQ**, **IV**, and **PPO** under stochastic environment dynamics.