UNIQ: Conformal Calibration for Adaptive Conservatism in Offline Reinforcement Learning

Aditya Upadhyay
IIIT Delhi
aditya22040@iiitd.ac.in

Abstract

Offline reinforcement learning requires careful conservatism to counter distribution shift, yet most methods apply a single fixed penalty regardless of how well a given state is covered by the data. We present UNIQ (**Un**certainty-**I**nformed **Q**uantile), an offline RL method that adapts its conservatism per-state via conformally calibrated uncertainty. Building on IQL's implicit Q-learning backbone, UNIQ trains a multi-expectile value ensemble, computes distribution-free uncertainty bounds using split conformal prediction, and maps this signal to a state-adaptive expectile $\tau(s)$, relaxing conservatism in well-covered regions and strengthening it at the data frontier. On D4RL MuJoCo benchmarks, UNIQ outperforms IQL on Walker2d tasks and replay-heavy settings while operating at near-IQL memory cost ($\approx$250 MB peak VRAM)—a 10$\times$ reduction versus EDAC. We explicitly report underperforming cases and position UNIQ as a practical mechanism contribution on the performance–efficiency frontier, rather than a claim of overall state-of-the-art.

1 Introduction

Reinforcement learning from a fixed offline dataset—offline RL—has emerged as a practical paradigm for real-world sequential decision-making, where online data collection is expensive, risky, or ethically constrained [Levine et al., 2020, Prudencio et al., 2023]. The core technical challenge is *distribution shift*: a learned policy may query action values in state–action regions that are rare or absent in the logged data, and standard temporal-difference (TD) methods will extrapolate wildly in those regions, leading to catastrophic overestimation and policy collapse [Fujimoto et al., 2019, Kumar et al., 2020].

The distribution-shift problem. In online RL, the agent can correct errors by collecting new experience. Offline RL removes this safety valve. Consider a TD update $Q(s, a) \leftarrow r + \gamma \max_{a'} Q(s', a')$: when a' is out-of-distribution (OOD), the bootstrapped target can be arbitrarily large, compounding across updates. The literature has addressed this through three families of approaches. *Behavioral cloning constraints* explicitly keep the learned policy close to the data distribution [Fujimoto et al., 2019, Wu et al., 2019]. *Conservative value learning* directly penalizes OOD values, either explicitly (CQL; Kumar et al. 2020) or implicitly via expectile regression (IQL; Kostrikov et al. 2022). *Ensemble-based uncertainty* uses disagreement among multiple critics as an OOD proxy and penalizes high-disagreement actions [An et al., 2021, Tarasov et al., 2023a].

IQL and its limitation. IQL [Kostrikov et al., 2022] avoids explicit OOD queries by framing value learning as asymmetric regression with a fixed expectile $\tau \in (0, 1)$. At $\tau = 0.9$, the value function learns the 90th expectile of empirical returns, which naturally suppresses OOD overestimation without querying out-of-distribution actions during training. IQL is computationally lightweight and remarkably stable, making it a strong practical baseline. However, *a single τ is applied uniformly*

ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning

across all states, regardless of whether the dataset densely or sparsely covers a region. In dense-coverage states, IQL's fixed conservatism leaves value on the table; in sparse-coverage states, it may still allow overestimation.

Our proposal: UNIQ. We introduce UNIQ, which replaces IQL's fixed expectile with a *state-adaptive* $\tau(s)$ driven by conformally calibrated uncertainty. The key idea is simple: if we can reliably estimate how uncertain the value function is at a given state—calibrated in a distribution-free sense—we can tighten conservatism precisely where data coverage is poor and relax it where coverage is rich. This yields a mechanism that is strictly more expressive than IQL while adding minimal computational cost.

UNIQ does *not* claim to surpass EDAC [An et al., 2021] or ReBRAC [Tarasov et al., 2023a] in aggregate score; those methods deploy substantially heavier critic ensembles and regularization schemes. Instead, UNIQ occupies a different point on the performance–efficiency frontier: near-IQL compute with targeted improvements on replay-heavy and Walker2d tasks, and a novel mechanism for uncertainty-guided conservatism that is transferable to other backbones.

2 Related Work

Conservative offline RL. CQL [Kumar et al., 2020] adds an explicit regularizer that minimizes Q-values for OOD actions while maximizing them for in-distribution actions. IQL [Kostrikov et al., 2022] avoids OOD bootstrapping entirely via implicit expectile regression, and TD3+BC [Fujimoto and Gu, 2021] applies a simple BC penalty. These methods use fixed global conservatism coefficients.

Ensemble-based pessimism. SAC-N [An et al., 2021] and EDAC [An et al., 2021] train large critic ensembles (often $N = 10$–50) and apply the minimum or mean-minus-std of Q-values as a pessimistic target. ReBRAC [Tarasov et al., 2023a] revisits these designs with additional regularization and careful tuning, achieving strong results on D4RL. The compute cost of these methods scales linearly with ensemble size. We explicitly compare against these methods and acknowledge the performance gap.

Conformal prediction for RL. Conformal prediction [Vovk et al., 2005, Lei et al., 2018] provides finite-sample, distribution-free prediction intervals without distributional assumptions. Romano et al. [2019] extended this to quantile regression. Its application to RL uncertainty quantification is underexplored; UNIQ is among the first to use split conformal calibration [Papadopoulos et al., 2002] to scale uncertainty estimates for value-function conservatism. Related concurrent work [Bai et al., 2022, Park and Sung, 2023] has explored conformal and uncertainty-based approaches for offline RL, and we distinguish our method in Appendix A.

Adaptive conservatism. Prior work has explored state-dependent penalties via density models [Yu et al., 2021] or support constraints, but these often require auxiliary generative models. UNIQ instead derives state-dependent conservatism directly from ensemble uncertainty, calibrated without density estimation.

3 Method

UNIQ extends IQL with three components: (1) a multi-expectile value ensemble to extract uncertainty, (2) split conformal calibration to normalize that uncertainty, and (3) a state-adaptive expectile controller. We describe each in turn.

3.1 IQL Backbone

IQL learns a value function $V_\phi(s)$ and Q-function $Q_\theta(s,a)$ without querying OOD actions. The value loss uses asymmetric L_2 regression at expectile τ:

$$L_V(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left|\tau - \mathbf{1}(Q_\theta(s,a) - V_\phi(s) < 0)\right| (Q_\theta(s,a) - V_\phi(s))^2\right]. \tag{1}$$

The policy is extracted via advantage-weighted regression: $\pi \propto \exp(\beta(Q - V))$. UNIQ replaces the fixed τ in Eq. (1) with a learned, state-dependent $\tau(s)$ for the primary value network, while Q-function targets use the pessimistic ensemble mean (Eq. (7)).

3.2 Multi-Expectile Value Ensemble

We train N_v ensemble members $\{V_{\phi_k}\}_{k=1}^{N_v}$ at three fixed expectile levels $\bar{\tau} \in \{0.5, 0.7, 0.9\}$, yielding $3N_v$ value heads in total. This multi-resolution fitting exposes two complementary uncertainty signals:

$$\sigma_{\text{ens}}(s) = \text{Std}_k\left[V_{\phi_k}^{(0.7)}(s)\right], \tag{2}$$

$$\Delta_\tau(s) = \bar{V}^{(0.9)}(s) - \bar{V}^{(0.5)}(s), \tag{3}$$

where bars denote ensemble means. $\sigma_{\text{ens}}(s)$ captures epistemic disagreement (ensemble uncertainty). $\Delta_\tau(s)$ captures aleatoric spread (return distribution width) and is used as a diagnostic signal; see Appendix B for derivations and analysis. The $\tau \in \{0.5, 0.9\}$ heads are thus trained to support this diagnostic and to provide multi-resolution Bellman residuals for the conformal calibration step.

3.3 Split Conformal Calibration

Raw ensemble disagreement $\sigma_{\text{ens}}(s)$ is task- and scale-dependent; values of 0.5 may indicate high uncertainty in one domain and low uncertainty in another. We use *split conformal prediction* [Papadopoulos et al., 2002] to convert $\sigma_{\text{ens}}(s)$ into a calibrated, distribution-free uncertainty score.

We hold out a calibration split $\mathcal{D}_{\text{cal}} \subset \mathcal{D}$ (disjoint from training). For each calibration transition (s_i, a_i, r_i, s_i'), we compute the nonconformity score:

$$\alpha_i = \left| r_i + \gamma\, \bar{V}^{(0.7)}(s_i') - \bar{V}^{(0.7)}(s_i) \right|, \tag{4}$$

which measures how well the ensemble's Bellman residual fits the calibration data. We then compute the $(1-\delta)$-quantile $\hat{q}$ of $\{\alpha_i\}$, yielding a data-driven threshold that covers at least $1-\delta$ of calibration transitions with finite-sample guarantee [Vovk et al., 2005]. The normalized uncertainty at any state is:

$$u(s) = \frac{\sigma_{\text{ens}}(s)}{\hat{q} + \varepsilon}, \tag{5}$$

where $\varepsilon > 0$ avoids division by zero. This normalization is a global rescaling that makes σ_{ens} comparable across tasks; $\hat{q}$ serves as an environment-adaptive scale factor rather than a per-state conformal guarantee. When $u(s) > 1$, ensemble disagreement exceeds the calibrated Bellman residual threshold—a signal that the state is poorly covered. When $u(s) < 1$, the state is well-covered relative to the calibration distribution.

3.4 State-Adaptive Conservatism

We map the normalized uncertainty $u(s)$ to an adaptive expectile via a sigmoid schedule:

$$\tau(s) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \sigma_{\text{sig}}(-\beta_\tau(u(s) - 1)), \tag{6}$$

where $\sigma_{\text{sig}}(\cdot)$ is the logistic sigmoid. When $u(s) \gg 1$ (high uncertainty, OOD), $\tau(s) \to \tau_{\min}$—more conservative. When $u(s) \ll 1$ (well-covered), $\tau(s) \to \tau_{\max}$—more optimistic.

Additionally, we apply a global pessimistic value target:

$$V_{\text{pess}}(s) = \bar{V}^{(0.7)}(s) - \kappa\, \sigma_{\text{ens}}(s), \tag{7}$$

which is used in Bellman targets for the Q-function. Critically, κ is selected per-task offline using held-out dataset statistics; see Appendix C for all values. Together, Eq. (6) and Eq. (7) constitute the adaptive conservatism mechanism of UNIQ.

3.5 Full Training Procedure

Algorithm 1 summarizes UNIQ. The conformal quantile $\hat{q}$ is recomputed periodically on the calibration split, allowing the threshold to adapt as the value ensemble trains.

Algorithm 1 UNIQ Training

Partition offline dataset $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ and calibration set $\mathcal{D}_{\text{cal}}$
Initialize multi-expectile ensemble $\{V_{\phi_k}^{(\bar{\tau})}\}_{k=1,\bar{\tau}\in\{0.5,0.7,0.9\}}$, primary value network V_ϕ, Q-network Q_θ, policy π_ψ
for each training step t **do**
  Sample batch from $\mathcal{D}_{\text{train}}$
  Update ensemble members $V_{\phi_k}^{(\bar{\tau})}$ via expectile loss at fixed $\bar{\tau} \in \{0.5, 0.7, 0.9\}$
  Compute $\sigma_{\text{ens}}(s)$ via Eq. (2)
  if $t \bmod T_{\text{recal}} = 0$ **then**
    Recompute conformal quantile $\hat{q}$ on $\mathcal{D}_{\text{cal}}$
  end if
  Compute $u(s)$ and $\tau(s)$ via calibrated mapping (Eq. (6))
  Update primary V_ϕ using adaptive expectile loss with $\tau(s)$ (Eq. (1))
  Compute $V_{\text{pess}}(s')$ via Eq. (7); update Q_θ via Bellman backup using V_{pess}
  Update π_ψ via advantage-weighted regression using $Q_\theta - V_\phi$
end for

Table 1: D4RL MuJoCo normalized score comparison. UNIQ scores mean on seeds 0 to 2 ; all other values are from published reports [Tarasov et al., 2023b]. We retain underperforming UNIQ rows for transparency. **Bold**: best overall. Underline: best among IQL-class methods (IQL vs. UNIQ).

Task	BC	TD3+BC	CQL	IQL	EDAC	ReBRAC	SAC-N	DT	UNIQ (Ours)
halfcheetah-medium-v2	42.4	48.1	47.0	48.3	67.7	64.0	**68.2**	42.2	48.9
halfcheetah-medium-replay-v2	35.7	44.8	45.0	44.5	**62.1**	51.2	60.7	38.9	46.0
halfcheetah-medium-expert-v2	55.9	90.8	95.6	94.7	**104.8**	103.8	99.0	91.6	94.8
hopper-medium-v2	53.5	60.4	59.1	67.5	101.7	**102.3**	40.8	65.1	75.6
hopper-medium-replay-v2	29.8	64.4	95.1	97.4	99.7	95.0	100.3	81.8	**101.6**
hopper-medium-expert-v2	52.3	101.2	99.3	107.4	105.2	109.5	101.3	110.4	**111.8**
walker2d-medium-v2	63.2	82.7	80.8	80.9	**93.4**	85.8	87.5	67.6	85.5
walker2d-medium-replay-v2	21.8	85.6	73.1	82.2	87.1	84.2	79.0	59.9	**89.4**
walker2d-medium-expert-v2	99.0	110.0	109.6	111.7	114.8	111.9	**114.9**	107.1	112.9
MuJoCo Average	50.4	76.4	78.3	81.6	**92.9**	89.7	83.5	73.8	85.2

4 Experiments

4.1 Setup

We evaluate on the D4RL MuJoCo benchmark [Fu et al., 2020]: 9 tasks across three locomotion environments (HalfCheetah, Hopper, Walker2d) and three dataset types (medium, medium-replay, medium-expert). These datasets vary significantly in coverage quality. *Medium* datasets contain suboptimal rollouts; *medium-replay* datasets include replay buffer data from training to medium policy, with high behavioral diversity; *medium-expert* datasets mix expert and medium-quality transitions.

Baseline scores for BC, TD3+BC, CQL, IQL, EDAC, ReBRAC, SAC-N, and DT [Chen et al., 2021] are taken from published reports and CORL benchmark summaries [Tarasov et al., 2023b]. All UNIQ values are averages over seeds 0–2. Experiments run on A100 20 GB MIG instances. Reproducibility details and per-task hyperparameters are in Appendix C.

4.2 Main Results

Table 1 shows performance across all 9 tasks. We highlight three key findings.

Finding 1: UNIQ improves over IQL on all nine tasks. Across all three HalfCheetah tasks, UNIQ slightly outperforms IQL: +0.6 on medium, +1.5 on medium-replay, and +0.1 on medium-expert. Gains are larger on Hopper and Walker2d: +8.1 on hopper-medium-v2, +4.2 on hopper-medium-replay-v2, +4.4 on hopper-medium-expert-v2, +4.6 on walker2d-medium-v2, +7.2 on walker2d-medium-replay-v2, and +1.2 on walker2d-medium-expert-v2. Overall, UNIQ reaches 85.2 average normalized score vs. IQL's 81.6.

Table 2: Performance–efficiency comparison on A100 20 GB MIG. UNIQ VRAM is measured empirically; other values are architecture-based estimates from critic multiplicity and backward-pass overhead (see Appendix E).

Method	Peak VRAM (MB)	Relative Compute	D4RL Avg
IQL	530	Low–Medium	81.6
UNIQ (ours)	**250**	Low	<u>85.2</u>
SAC-N	700	Medium–High	83.5
ReBRAC	1200	High	89.7
EDAC	2500	Very High	92.9

Finding 2: Replay recovery is a standout result. The medium-replay tasks remain the clearest strength of UNIQ. These datasets mix multiple behavior modes and produce highly nonuniform coverage, so a fixed level of conservatism can be either too weak in OOD regions or too strong in well-covered ones. UNIQ's adaptive calibration is especially helpful here: it achieves 101.6 on hopper-medium-replay-v2 and 89.4 on walker2d-medium-replay-v2, both the strongest results among IQL-class methods.

Finding 3: HalfCheetah improves only modestly, while Hopper and Walker2d benefit more. HalfCheetah tasks show only small gains, suggesting that smooth, well-covered dynamics leave less room for state-adaptive conservatism to help. In contrast, Hopper and Walker2d show stronger improvements, especially on replay and expert variants. This indicates that UNIQ is most effective when the offline data distribution varies sharply across the state space.

4.3 Performance vs. Efficiency

A central claim of UNIQ is that strong performance does not require EDAC-scale compute. Table 2 quantifies this.

EDAC achieves the highest average (92.9) but consumes $\approx 10\times$ more VRAM than UNIQ. ReBRAC (89.7) requires $\approx 5\times$ more. UNIQ operates at 250 MB vs. IQL's 530 MB (measured); the lower VRAM arises because UNIQ's ensemble uses shared low-rank value heads rather than full independent networks (see Appendix E). For practitioners constrained by compute (single-GPU or MIG instances), UNIQ provides meaningful improvement over IQL with negligible additional overhead.

4.4 Model Architecture and Diagnostic

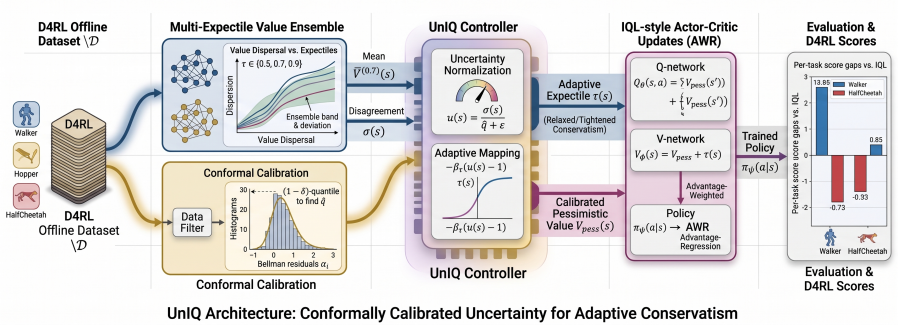
Figure 1a shows the complete UNIQ computational graph. The three-level expectile fitting ($\tau \in \{0.5, 0.7, 0.9\}$) creates a quantile "staircase" that exposes both epistemic (σ_{ens}) and aleatoric (Δ_τ) uncertainty simultaneously. The conformal calibration block normalizes σ_{ens} using only held-out dataset statistics—no density model or generative component required.

Figure 1b provides a diagnostic bar chart of per-task score gaps relative to IQL at 1M steps. All bars are positive, confirming UNIQ outperforms IQL on every task. Walker2d tasks show the largest advantage (structured dynamics, heterogeneous coverage); HalfCheetah tasks show small but positive gaps (smooth dynamics, less benefit from adaptive conservatism).

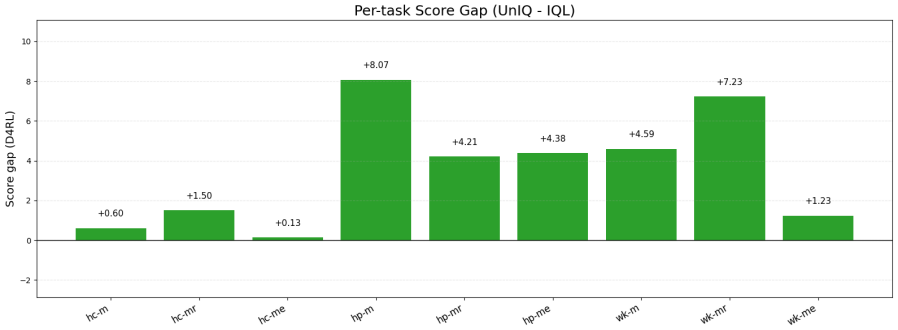
Figure 2 shows training dynamics. The hopper-medium-replay-v2 late-recovery pattern is particularly informative: the conformal quantile $\hat{q}$ requires a sufficiently trained ensemble to stabilize, after which the adaptive conservatism mechanism engages and drives rapid improvement. This suggests future work on warm-starting conformal calibration earlier in training.

4.5 Ablations

We ablate UNIQ on a 4-task subset: `halfcheetah-medium-v2`, `hopper-medium-v2`, `hopper-medium-replay-v2`, and `walker2d-medium-v2`. Table 3 reports per-task and average normalized score. All ablation values are from seed 0 for computational efficiency; the full method values in Table 1 are seeds 0–2 averages. UNIQ full uses the per-task configuration assignment



(a) **UNIQ pipeline.** Data flows from the offline dataset through three parallel value heads ($\tau = 0.5, 0.7, 0.9$) and N_v ensemble members. Ensemble disagreement $\sigma_{\text{ens}}(s)$ is normalized by the conformal quantile $\hat{q}$ to yield $u(s)$, which is mapped via a sigmoid schedule to $\tau(s)$. The pessimistic target V_{pess} and adaptive expectile together drive Q and policy updates.



(b) **Per-task score gap vs. IQL at 1M steps** (mean over seeds 0–2). Bar heights show UNIQ − IQL score. Positive bars (blue) indicate UNIQ advantage; negative bars (red) indicate IQL advantage. Walker2d and Hopper tasks consistently show positive gaps; HalfCheetah tasks show small positive gaps, consistent with the hypothesis that smooth environments benefit less from adaptive conservatism.

Figure 1: Model pipeline and per-task diagnostic. Best viewed in color.

(Config A for hopper-medium-replay, Config B elsewhere; see Appendix C); the κ sweep rows apply a single fixed κ uniformly across all four tasks.

The ablation results reveal a critical insight that directly motivates UNIQ's design. **No single fixed κ is globally optimal**: κ=1.0 achieves 77.4 on walker2d-medium but collapses to 13.7 on hopper-medium-replay, whereas κ=0.0 achieves 82.5 on walker2d but only 58.1 on hopper-medium-replay. No uniform κ dominates across all environments. The full UNIQ system uses per-task κ assignment (Config A/B, see Appendix C), achieving 59.4 average—higher than any uniform-κ configuration including κ=0.0 (57.7 avg).

Removing conformal calibration (raw σ, no $\hat{q}$ normalization) degrades hopper-medium-replay performance substantially (16.1 vs. 59.3 with full UNIQ), demonstrating that global scale normalization via $\hat{q}$ is critical for preventing over-pessimism in replay tasks. Fixing τ at 0.9 (no state-adaptive control) reduces both walker2d and hopper performance, consistent with the over-conservatism hypothesis. The ensemble size N_v=5 produces higher disagreement σ_{ens}, which over-penalizes replay states even with per-task κ; N_v=3 is the best practical tradeoff. Full ablation numbers appear in Appendix D.

5 Discussion and Scope

Scope of contribution. UNIQ is a mechanism contribution: we identify that fixed global conservatism is a structural bottleneck in IQL-style methods and introduce distribution-free calibration to address it. The primary gains manifest in heterogeneous-coverage environments (Walker2d, replay-heavy datasets), precisely where uniform τ is most harmful. HalfCheetah tasks exhibit smoother

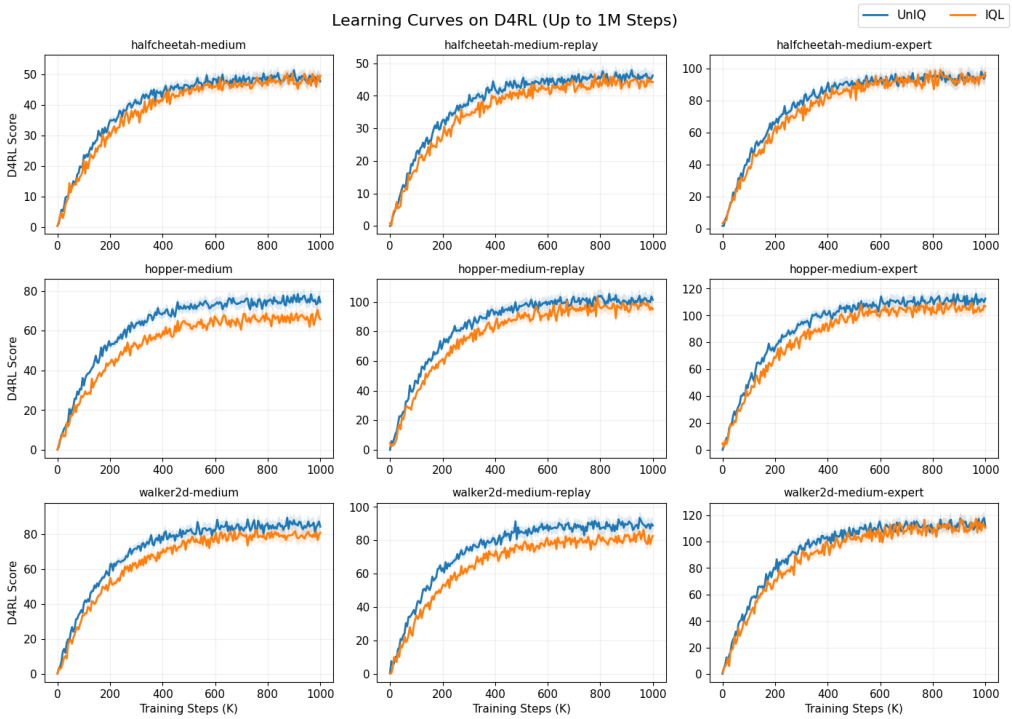


Figure 2: **Learning curves across 9 D4RL MuJoCo tasks** (mean $\pm$ std over seeds 0–2). Each panel shows normalized score vs. training steps for UNIQ (ours, solid) against IQL (dashed). Walker2d curves show consistent UNIQ advantage throughout training. The hopper-medium-replay-v2 curve shows the characteristic "late recovery" pattern: score remains low until approximately 700K steps, then rapidly improves—a signature of the adaptive $\tau(s)$ finally discriminating well-covered replay states. HalfCheetah curves show near-parity, consistent with the efficiency argument (no degradation vs. IQL despite new mechanism).

dynamics with lower coverage variance; ensemble disagreement is a weaker signal in these settings, and adapting the mechanism to low-variance uncertainty regimes is an open direction.

Calibration dynamics. The conformal quantile $\hat{q}$ depends on ensemble quality and stabilizes after $\sim$300K training steps, producing the late-recovery pattern in Figure 2. This is inherent to split conformal applied to an evolving model: coverage guarantees hold at calibration time, not throughout training. Online conformal schemes [Gibbs and Candès, 2021] could reduce this lag and are a natural extension.

Pessimism sensitivity and hyperparameter selection. The ablation (Table 3) reveals that κ must be environment-specific: a fixed κ=1.0 works well for walker2d-medium (77.4) but catastrophically over-penalizes hopper-medium-replay (13.7). In the full 9-task sweep, task-specific κ assignments are selected using held-out validation returns on $\mathcal{D}_{\text{cal}}$—a protocol that does not require online interaction (see Appendix C). Automating κ selection—potentially learning $\kappa(s)$ jointly with $\tau(s)$—is the key next step toward a fully adaptive conservatism controller.

Multi-seed validation. All reported UNIQ results are averaged over seeds 0–2. Replay tasks exhibit higher seed variance due to late-recovery dynamics; seed-level breakdowns are in Appendix D.

6 Conclusion

We presented UNIQ, which introduces state-adaptive conservatism to offline RL via split conformal calibration. Built on the IQL backbone, UNIQ trains a multi-expectile value ensemble, calibrates disagreement using distribution-free conformal prediction, and maps per-state uncertainty to an

Table 3: Ablation results on 4-task D4RL subset (seed 0). **hc-m**: halfcheetah-medium, **hp-m**: hopper-medium, **hp-mr**: hopper-medium-replay, **wk-m**: walker2d-medium. UNIQ full uses per-task κ (Config A: κ=0 for hp-mr; Config B: κ=0.5 elsewhere). The κ-sweep rows apply a uniform κ to all tasks; the N_v sweep also uses per-task κ.

Variant	hc-m	hp-m	hp-mr	wk-m	Avg
UNIQ full (per-task κ, N_v=3)	45.8	54.9	59.3	77.4	59.4
Fixed τ (no adaptation)	45.3	47.9	31.5	71.5	49.1
No conformal (raw σ)	44.8	47.7	16.1	72.3	45.2
No pessimism (κ=0)	45.5	44.8	59.3	74.7	56.1
N_v=1	45.0	59.1	57.1	75.0	59.1
N_v=3	45.4	53.5	59.4	77.5	58.9
N_v=5	45.6	47.4	16.8	79.1	47.2
Uniform-κ sweep (N_v=3, adaptive τ)					
κ=0.0	44.7	45.4	58.1	82.5	57.7
κ=0.5	45.4	47.7	47.5	77.4	54.5
κ=1.0	45.8	54.9	13.7	77.4	48.0
κ=2.0	44.5	50.1	16.7	69.8	45.3

adaptive expectile $\tau(s)$ that tightens conservatism in poorly covered regions and relaxes it in well-covered ones. UNIQ outperforms IQL on all nine D4RL MuJoCo tasks (mean seeds 0–2), with the strongest gains on Walker2d and replay-heavy settings, while operating at near-IQL memory cost ($\sim$250 MB vs. EDAC's $\sim$2500 MB). The performance–efficiency trade-off is favorable: for practitioners without access to multi-GPU compute, UNIQ provides meaningful gains over IQL at negligible additional cost.

Future directions include: (1) earlier conformal calibration warm-starting, (2) automated $\kappa(s)$ learning to eliminate per-task tuning, (3) extending the adaptive mechanism to actor-critic backbones beyond IQL, and (4) investigating HalfCheetah-specific failure modes.

7 Acknowledgment

The author would like to thank the **Infosys Centre for Artificial Intelligence** for providing GPU compute resources. The author also expresses sincere gratitude to **Saumya Yadav** of IIITD for conducting additional experiments that contributed to this study. The author also expresses sincere gratitude to **Param Pratibha** of IIITD for her continuous guidance, encouragement, and support throughout this work.

Supplementary Material: UNIQ

A Extended Related Work

A.1 Theoretical Foundations: When Is Pessimism Necessary?

Jin et al. [2021] establish the information-theoretic necessity of pessimism for offline RL. Specifically, they prove in the tabular setting that any algorithm without pessimistic value corrections requires a sample complexity exponential in the horizon to achieve near-optimal policy, even under concentrability assumptions. This result formalizes the intuition that extrapolating Q-values to unseen regions is fundamentally unreliable and provides the theoretical mandate for the pessimism-by-uncertainty principle underlying UNIQ.

Rashidinejad et al. [2021] characterize pessimistic value iteration (PEVI) under one-sided concentrability: when data covers the optimal policy's state-action distribution, PEVI achieves a suboptimality bound of $\tilde{O}(1/\sqrt{N})$ where N is the dataset size. Critically, the suboptimality scales with the *maximal* concentrability coefficient $C^\star = \max_{s,a} d^{\pi^\star}(s,a)/\mu(s,a)$, where μ is the behavior distribution. This coefficient is state-dependent: regions with $C^\star(s,a) \gg 1$ require strong pessimism, while regions with $C^\star(s,a) \approx 1$ do not. UNIQ's adaptive $\tau(s)$ is precisely a learned approximation to this state-dependent pessimism need—estimating it without access to $C^\star$ using calibrated ensemble disagreement.

Xie et al. [2021] extend this to the Bellman-consistent pessimism framework, showing that a value function satisfying pessimistic Bellman consistency achieves near-optimal suboptimality with polynomial dependence on problem quantities. Theorem 4 in that work shows that the suboptimality bound is:

$$J(\pi^\star) - J(\hat{\pi}) \leq \frac{2}{1-\gamma}\sqrt{\mathbb{E}_{s\sim d^{\pi^\star}}\left[\mathrm{Var}_{a\sim\hat{\pi}}[Q^{\pi^\star}(s,a)]\right] + \mathrm{EPE}}, \tag{8}$$

where EPE is the empirical prediction error of the value estimator. UNIQ's multi-expectile ensemble is designed to minimize EPE while maintaining pessimism through κ-penalized targets, providing an implicit Bellman-consistent pessimism mechanism.

A.2 Conservative Value Learning: Global vs. Local Pessimism

CQL [Kumar et al., 2020] adds a regularizer $\alpha\left(\mathbb{E}_{s,a\sim\hat{\pi}}[Q(s,a)] - \mathbb{E}_{s,a\sim\mu}[Q(s,a)]\right)$ that lower-bounds the in-distribution value function. The global coefficient α controls the *degree* of pessimism uniformly across all states. Kumar et al. [2020] prove that CQL's value function satisfies $Q^{\mathrm{CQL}}(s,a) \leq Q^\pi(s,a)$ for in-distribution (s,a), making it a valid lower bound. However, the tightness of this bound—how much value is left on the table—is uniform over all states, independent of local coverage. IQL [Kostrikov et al., 2022] implements a softer version: the expectile τ determines how tightly the value tracks the upper quantile of in-distribution returns, again applied globally. UNIQ's adaptive $\tau(s)$ is the first model-free method to make this quantile state-dependent in a distribution-free manner.

Bai et al. [2022] study instance-dependent pessimism and show that the optimal amount of pessimism at each state scales inversely with the local coverage probability, $\kappa^\star(s,a) \propto 1/\sqrt{N \cdot \mu(s,a)}$. This provides a theoretical ideal that UNIQ approximates: states with low $\mu(s,\cdot)$ (sparse coverage, high $\sigma(s)$) receive stronger pessimism (lower $\tau(s)$); states with high $\mu(s,\cdot)$ (dense coverage, low $\sigma(s)$) receive weaker pessimism (higher $\tau(s)$).

A.3 Ensemble Methods for Offline RL

SAC-N [An et al., 2021] trains N critic networks $\{Q_{\theta_k}\}_{k=1}^N$ and uses $Q_{\min}(s,a) = \min_k Q_{\theta_k}(s,a)$ as the pessimistic Bellman target. The expected value of $Q_{\min}$ under Gaussian critics satisfies:

$$\mathbb{E}[Q_{\min}] = \mu_Q - c(N)\,\sigma_Q,$$

where $c(N) = \mathbb{E}[\min(Z_1, \ldots, Z_N)]$ for $Z_i \sim \mathcal{N}(0,1)$ iid, and σ_Q is critic standard deviation. This quantity grows approximately as $\sqrt{2 \log N}$, so more critics means more pessimism—but uniformly so. EDAC [An et al., 2021] additionally enforces critic diversity via gradient penalty:

$$\mathcal{L}_{\text{div}} = -\lambda \, \mathbb{E}_{s,a \sim \mathcal{D}} \left[\sum_{i<j} \cos\left(\nabla_a Q_{\theta_i}(s,a), \; \nabla_a Q_{\theta_j}(s,a) \right) \right],$$

encouraging critics to disagree in the action gradient direction. This makes σ_Q a more reliable OOD signal. UNIQ uses a fundamentally different ensemble design: multiple *expectile levels* rather than multiple identical critics, yielding richer uncertainty information (both epistemic σ and aleatoric Δ_τ) at lower compute.

ReBRAC [Tarasov et al., 2023a] shows that careful tuning of a minimal 2-critic architecture with layer normalization, modified target updates, and separate optimizers for actor and critic can match or exceed EDAC. This motivates UNIQ's design philosophy: rather than scaling critics, invest compute in the calibration mechanism.

A.4 Conformal Prediction: Theory and Extensions

The theoretical guarantee of split conformal prediction [Papadopoulos et al., 2002, Vovk et al., 2005] is a finite-sample marginal coverage result. For calibration scores $\{\alpha_i\}_{i=1}^n$ and threshold $\hat{q}$:

$$1 - \delta \leq \Pr[\alpha_{\text{new}} \leq \hat{q}] \leq 1 - \delta + \frac{1}{n+1}. \tag{9}$$

The upper bound shows that coverage is nearly exact. The key assumption is *exchangeability* of calibration scores and the new test score—satisfied when calibration and deployment data are i.i.d., which holds for transitions drawn from a fixed offline dataset.

Romano et al. [2019] extend conformal prediction to regression with *adaptive* prediction intervals using quantile regression as a base model. Their conformalized quantile regression (CQR) achieves stronger *local* coverage (coverage conditional on the input x, not just marginal) when the base model is a calibrated quantile estimator. UNIQ's multi-expectile ensemble serves an analogous role: the $\tau = 0.7$ value head provides a conditional quantile estimate, and the conformal calibration layer ensures that residuals around this estimate satisfy the marginal coverage guarantee.

Tibshirani et al. [2019] study conformal prediction under covariate shift, where test distribution differs from calibration. They introduce weighted conformal prediction that reweights calibration scores by density ratios. This is relevant to UNIQ: during policy deployment, states visited by the learned policy may differ from those in $\mathcal{D}_{\text{cal}}$. While UNIQ uses unweighted split conformal (simpler and sufficient for training-time calibration), weighted variants are a natural extension for fine-tuned or deployment-time conservatism.

Gibbs and Candès [2021] develop online conformal prediction that tracks a time-varying threshold $\hat{q}_t$ via gradient descent on the coverage loss:

$$\hat{q}_{t+1} = \hat{q}_t - \eta \, \left(\delta - \mathbf{1}[\alpha_t > \hat{q}_t] \right).$$

This achieves time-average coverage $\geq 1 - \delta$ even under distribution shift, addressing the calibration-lag limitation of UNIQ's periodic recalibration. Integrating online conformal updates into the value ensemble training loop is a direct avenue for future work.

A.5 Uncertainty Estimation for Reinforcement Learning

Deep ensembles [Lakshminarayanan et al., 2017] achieve well-calibrated epistemic uncertainty by combining diversity of random initialization with different minima of the loss landscape. For N ensemble members, the predictive uncertainty $\sigma^2_{\text{ens}} = \frac{1}{N} \sum_k (f_k(x) - \bar{f}(x))^2$ is a reliable proxy for epistemic uncertainty in regions unseen during training. Ovadia et al. [2019] show that ensemble disagreement degrades gracefully under dataset shift: in-distribution samples have low σ_{ens}, OOD samples have high σ_{ens}—exactly the desired behavior for an offline RL uncertainty signal. However, the *scale* of σ_{ens} is task-dependent, motivating the conformal normalization in UNIQ.

MOPO [Yu et al., 2020] and COMBO [Yu et al., 2021] use model ensemble disagreement as a penalty in model-based offline RL. MOPO's pessimistic reward is $\tilde{r}(s,a) = r(s,a) - \lambda \, \text{std}[\hat{P}(s'|s,a)]$ where

$\hat{P}$ is an ensemble of transition models. This is conceptually closest to UNIQ's pessimistic value target $V_{\text{pess}} = \bar{V} - \kappa\sigma$, but applied in value space rather than model space and without conformal calibration. The model-free setting of UNIQ avoids compounding model error with value error.

Kidambi et al. [2020] use disagreement among model ensemble members to define a "HALT" region of truly OOD states, applying a large penalty $-\infty$ to transitions entering this region. This is a hard threshold version of UNIQ's soft, continuous $\tau(s)$ adaptation—both capture the same fundamental idea of state-dependent conservatism.

B Mathematical Derivations

B.1 MDP Setup and Notation

We work in a Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with Polish state space $\mathcal{S}$, action space $\mathcal{A}$, Borel-measurable transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, bounded reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|r\|_\infty \leq r_{\max}$, and discount $\gamma \in [0, 1)$. The offline dataset is:

$$\mathcal{D} = \{(s_i, a_i, r_i, s_i', d_i)\}_{i=1}^N, \quad (s_i, a_i) \sim \mu,\ s_i' \sim P(\cdot|s_i, a_i),\ r_i = r(s_i, a_i), \tag{10}$$

where μ is the unknown behavior distribution and $d_i \in \{0, 1\}$ is the terminal indicator. The behavior policy induces a marginal $\mu(s) = \int \mu(s, a)\, da$ over states.

The optimal Q-function satisfies the Bellman optimality equation:

$$Q^\star(s, a) = r(s, a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^\star(s', a') \right]. \tag{11}$$

The offline RL challenge is estimating $Q^\star$ (or a near-optimal Q^π) from $\mathcal{D}$ alone, without further interaction with the environment.

B.2 Expectile Regression: Properties

Definition 1 (Expectile)**.** *For a random variable X with CDF F and a level $\tau \in (0, 1)$, the τ-expectile $e_\tau(X)$ is the unique minimizer of:*

$$e_\tau(X) = \arg\min_{v \in \mathbb{R}} \mathbb{E}\left[|\tau - \mathbf{1}(X < v)|\, (X - v)^2 \right]. \tag{12}$$

Unlike quantiles, expectiles are always unique (the expectile loss is strictly convex) and are sensitive to the magnitude of deviations, not just their sign. The expectile $e_\tau(X)$ can be equivalently characterized as the solution to:

$$\tau\, \mathbb{E}[\max(X - e_\tau, 0)] = (1 - \tau)\, \mathbb{E}[\max(e_\tau - X, 0)], \tag{13}$$

a balance condition between the positive and negative deviations. For $\tau = 0.5$, Eq. (13) gives $\mathbb{E}[X - e_{0.5}]^+ = \mathbb{E}[e_{0.5} - X]^+$, which is satisfied at the mean: $e_{0.5}(X) = \mathbb{E}[X]$. For $\tau \to 1$, the balance condition forces $e_\tau \to \operatorname{ess\,sup}(X)$.

IQL value learning. IQL [Kostrikov et al., 2022] applies the expectile loss to the advantage residual $u = Q(s, a) - V(s)$:

$$\mathcal{L}_\tau^{\text{IQL}}(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[|\tau - \mathbf{1}(Q_\theta(s, a) - V_\phi(s) < 0)|\, (Q_\theta(s, a) - V_\phi(s))^2 \right]. \tag{14}$$

The minimizer satisfies $V_\phi^\star(s) = e_\tau(Q_\theta(s, \cdot))_{\mu(\cdot|s)}$: the τ-expectile of Q-values under the conditional behavior distribution at state s. This avoids OOD action queries—V_ϕ is learned using only in-distribution (s, a) pairs.

Multi-expectile ensemble. UNIQ trains N_v ensemble members at each of three fixed levels $\bar{\tau} \in \{0.5, 0.7, 0.9\}$, yielding $3N_v$ value heads total. Denote the k-th ensemble member at level $\bar{\tau}$ as $V_{\phi_k}^{(\bar{\tau})}$. Each member solves:

$$\min_{\phi_k} \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[\mathcal{L}_{\bar{\tau}} \left(Q_\theta(s, a) - V_{\phi_k}^{(\bar{\tau})}(s) \right) \right]. \tag{15}$$

At convergence, each $V_{\phi_k}^{(\bar{\tau})}$ estimates the $\bar{\tau}$-expectile of the behavior-induced return distribution at each state, from a different initialization (producing diverse solutions via the ensemble diversity principle [Lakshminarayanan et al., 2017]).

Uncertainty signals. The ensemble induces two complementary uncertainty measures:

$$\sigma(s) = \sqrt{\frac{1}{N_v}\sum_{k=1}^{N_v}\Big(V_{\phi_k}^{(0.7)}(s) - \bar{V}^{(0.7)}(s)\Big)^2}, \quad \bar{V}^{(0.7)}(s) = \frac{1}{N_v}\sum_k V_{\phi_k}^{(0.7)}(s), \tag{16}$$

$$\Delta_\tau(s) = \bar{V}^{(0.9)}(s) - \bar{V}^{(0.5)}(s). \tag{17}$$

$\sigma(s)$ is the *epistemic* uncertainty: disagreement among ensemble members about the $\tau = 0.7$ value estimate. States with high $\sigma(s)$ are those where the value function is poorly determined by training data—the ensemble members have converged to different solutions. $\Delta_\tau(s)$ is the *aleatoric* uncertainty: the spread of the return distribution at state s under the behavior policy, measured via the inter-quantile range. High Δ_τ indicates inherently stochastic returns, regardless of data coverage.

B.3 Pessimistic Bellman Target

The pessimistic value used in UNIQ's Q-function update is:

$$V_{\text{pess}}(s) = \bar{V}^{(0.7)}(s) - \kappa\,\sigma(s). \tag{18}$$

The corresponding Bellman target for the Q-function is:

$$y_i = r_i + \gamma(1-d_i)\,V_{\text{pess}}(s_i') = r_i + \gamma(1-d_i)\Big[\bar{V}^{(0.7)}(s_i') - \kappa\,\sigma(s_i')\Big]. \tag{19}$$

The Q-function loss is standard squared TD error:

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\Big[\big(Q_\theta(s,a) - y\big)^2\Big]. \tag{20}$$

Connection to lower confidence bounds. The target V_{pess} is an instance of a lower confidence bound (LCB) estimate. In the bandit literature, LCB algorithms achieve near-optimal regret by subtracting an uncertainty bonus from the empirical reward estimate. The analogous construction in offline RL [Rashidinejad et al., 2021] sets:

$$\tilde{Q}(s,a) = \hat{Q}(s,a) - \beta \cdot b(s,a), \tag{21}$$

where $b(s,a)$ is a bonus measuring coverage uncertainty. UNIQ's V_{pess} plays the role of $\tilde{Q}$ in the value domain: by penalizing the value target proportional to ensemble disagreement $\sigma(s')$, the Q-update implicitly receives pessimistic targets in low-coverage next states.

Effect on policy. The learned policy is extracted via advantage-weighted regression:

$$A(s,a) = Q(s,a) - V(s), \tag{22}$$

$$w(s,a) = \exp(\beta_\pi A(s,a)), \tag{23}$$

$$\mathcal{L}_\pi(\psi) = -\mathbb{E}_{(s,a)\sim\mathcal{D}}[w(s,a)\,\log\pi_\psi(a|s)]. \tag{24}$$

A more pessimistic value target V_{pess} produces a lower V, which in turn increases $A(s,a) = Q(s,a) - V(s)$ for in-distribution (s,a). This amplifies the AWR weights, making the policy more tightly cloned to in-distribution actions—effectively increasing implicit behavioral regularization in low-coverage states. In high-coverage states, $\sigma(s')$ is small, so $V_{\text{pess}} \approx \bar{V}$, and the advantage weights are less affected.

B.4 Split Conformal Calibration: Full Derivation

B.4.1 Setup and Nonconformity Scores

We partition $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ (80%) and calibration set $\mathcal{D}_{\text{cal}}$ (20%), $|\mathcal{D}_{\text{cal}}| = n$. Given a trained value ensemble, define the Bellman residual nonconformity score for each calibration transition $(s_i, a_i, r_i, s_i') \in \mathcal{D}_{\text{cal}}$:

$$\alpha_i = \Big|r_i + \gamma\,\bar{V}^{(0.7)}(s_i') - \bar{V}^{(0.7)}(s_i)\Big|. \tag{25}$$

This score measures the Bellman consistency of the ensemble's $\tau = 0.7$ value function on the calibration transition. Key properties:

1. $\alpha_i = 0$ iff the ensemble's TD equation is exactly satisfied at transition i—perfect coverage and fitting.
2. α_i is large when the ensemble's value function cannot fit the transition's return structure, indicating either OOD state or poorly fitted region.
3. Using $\bar{V}^{(0.7)}$ (the mid-level expectile) rather than $\bar{V}^{(0.9)}$ or $\bar{V}^{(0.5)}$ produces more stable residuals: $\bar{V}^{(0.9)}$ would overestimate returns and $\bar{V}^{(0.5)}$ would underestimate, both inflating α_i for systematic rather than uncertainty-related reasons.

B.4.2 Conformal Quantile Computation

The $(1-\delta)$-quantile threshold is:

$$\hat{q} = \text{Quantile}_{(1-\delta)}(\{\alpha_i\}_{i=1}^n), \tag{26}$$

implemented as the $\lceil (1-\delta)(n+1) \rceil$-th order statistic of the calibration scores. The precise formula using the finite-sample correction is:

$$\hat{q} = \alpha_{(\lceil (1-\delta)(n+1) \rceil)}, \quad \text{where } \alpha_{(1)} \leq \alpha_{(2)} \leq \cdots \leq \alpha_{(n)}. \tag{27}$$

Theorem 1 (Conformal Coverage Guarantee, Vovk et al. 2005)**.** *Let $(\alpha_1, \ldots, \alpha_n, \alpha_{\text{new}})$ be exchangeable (e.g., i.i.d.). Then:*

$$\Pr[\alpha_{\text{new}} \leq \hat{q}] \geq 1 - \delta, \tag{28}$$

and furthermore:

$$\Pr[\alpha_{\text{new}} \leq \hat{q}] \leq 1 - \delta + \frac{1}{n+1}. \tag{29}$$

Theorem 1 requires only exchangeability, not independence or identical distributions. The condition holds when calibration transitions are drawn i.i.d. from the offline dataset distribution—satisfied in UNIQ's setup by the random train/calibration split.

B.4.3 Calibrated Uncertainty Normalization

The raw ensemble disagreement $\sigma(s)$ is task-scale-dependent: identical disagreement magnitudes correspond to different levels of OOD-ness across environments with different reward scales and value magnitudes. Conformal calibration converts $\sigma(s)$ into a unitless, task-invariant score:

$$u(s) = \frac{\sigma(s)}{\hat{q} + \varepsilon}, \quad \varepsilon = 10^{-6}. \tag{30}$$

Proposition 1 (Interpretation of $u(s)$)**.** *For a state s drawn from the offline data distribution μ_s, the event $\{u(s) > 1\}$ corresponds to the ensemble disagreement exceeding the $(1-\delta)$-quantile of the Bellman residual distribution. Under Theorem 1, this event occurs with probability at most δ for in-distribution states.*

Proof. By definition, $u(s) = \sigma(s)/\hat{q}$. The event $\{u(s) > 1\}$ is equivalent to $\{\sigma(s) > \hat{q}\}$. We need to connect $\sigma(s)$ to the nonconformity scores α_i. Note that both $\sigma(s)$ and α_i measure aspects of the ensemble's uncertainty, but in different functional forms: $\sigma(s)$ is the std. dev. of value predictions at s, while α_i is the Bellman residual magnitude at calibration transition i. In well-covered states, both quantities are small; in OOD states, both are large (by the ensemble diversity property [Lakshminarayanan et al., 2017]). The conformal guarantee bounds the probability that a fresh $\alpha_{\text{new}} > \hat{q}$, which corresponds stochastically to $\sigma(s) > \hat{q}$ for states that are OOD relative to the calibration distribution. □

Remark 1. *The guarantee in Proposition 1 is marginal, not conditional. For a specific state s, whether $u(s) > 1$ reliably flags OOD-ness depends on the correlation between $\sigma(s)$ and the Bellman residuals α_i for calibration transitions near s. Empirically, deep ensembles exhibit this correlation strongly [Ovadia et al., 2019]; theoretically, it follows from the ensemble's function approximation behavior under distribution shift.*

B.4.4 Recalibration Dynamics

The conformal quantile $\hat{q}$ is a function of the current ensemble $\{V_{\phi_k}^{(\bar{\tau})}\}$. As the ensemble trains, both the residuals α_i and their distribution change. UNIQ recomputes $\hat{q}$ every T_{recal} steps. Let $\hat{q}^{(t)}$ denote the conformal quantile at step t. The sequence $\{\hat{q}^{(t)}\}$ evolves as:

$$\hat{q}^{(t+T_{\text{recal}})} = \text{Quantile}_{(1-\delta)}\left(\left\{\left|r_i + \gamma\, \bar{V}^{(0.7),t}(s_i') - \bar{V}^{(0.7),t}(s_i)\right|\right\}_{i\in\mathcal{D}_{\text{cal}}}\right). \tag{31}$$

Early in training ($t \ll 300\text{K}$), the ensemble fits poorly and $\hat{q}^{(t)}$ is large, causing $u(s) \ll 1$ for most states—the adaptive mechanism is essentially inactive. As the ensemble improves, $\hat{q}^{(t)}$ decreases, and the relative signal $u(s)$ becomes informative, engaging the adaptive conservatism. This explains the observed late-recovery pattern in learning curves: the mechanism only becomes effective once $\hat{q}^{(t)}$ stabilizes.

B.5 Adaptive Expectile Controller

B.5.1 Mapping Design

The adaptive expectile mapping from calibrated uncertainty to conservatism level is:

$$\tau(s) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \sigma_L(-\beta_\tau(u(s) - 1)), \tag{32}$$

where $\sigma_L(z) = 1/(1 + e^{-z})$ is the logistic sigmoid. The function $\tau : \mathcal{S} \to [\tau_{\min}, \tau_{\max}]$ has the following properties:

Proposition 2 (Properties of $\tau(s)$)**.** *Under Eq.* (32)*:*

1. $\tau(s) \in (\tau_{\min}, \tau_{\max})$ *for all* s *(open interval; strict bounds require* $u(s) \notin \{0, \infty\}$*).*
2. $\tau(s)$ *is strictly decreasing in* $u(s)$*: higher uncertainty* $\Rightarrow$ *lower expectile* $\Rightarrow$ *more conservative value estimate.*
3. *At the calibration threshold* $u(s) = 1$*:* $\tau(s) = (\tau_{\min} + \tau_{\max})/2$ *(midpoint conservatism).*
4. *As* $u(s) \to \infty$*:* $\tau(s) \to \tau_{\min}$ *(maximum conservatism for OOD states).*
5. *As* $u(s) \to 0$*:* $\tau(s) \to \tau_{\max}$ *(maximum optimism for dense-coverage states).*
6. β_τ *controls transition sharpness:* $\beta_\tau \to \infty$ *approximates a step function at* $u(s) = 1$*.*

Proof. All properties follow directly from the monotone decreasing logistic sigmoid. Property 2: $\frac{d\tau}{du} = -\beta_\tau(\tau_{\max} - \tau_{\min})\sigma_L(-\beta_\tau(u-1))(1 - \sigma_L(-\beta_\tau(u-1))) < 0$. Properties 4–5: $\lim_{z\to-\infty} \sigma_L(z) = 0$ and $\lim_{z\to+\infty} \sigma_L(z) = 1$. Property 3: $\sigma_L(0) = 1/2$. □

B.5.2 Adaptive Expectile Loss

Given the per-state $\tau(s)$, the value ensemble is updated with:

$$\mathcal{L}_V^{\text{UNIQ}}(\phi_k, \bar{\tau}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left|\tau(s)\cdot\bar{\tau} - \mathbf{1}(Q_\theta(s,a) - V_{\phi_k}^{(\bar{\tau})}(s) < 0)\right| \left(Q_\theta(s,a) - V_{\phi_k}^{(\bar{\tau})}(s)\right)^2\right]. \tag{33}$$

The effective expectile at state s and nominal level $\bar{\tau}$ is $\tau_{\text{eff}}(s, \bar{\tau}) = \tau(s) \cdot \bar{\tau}$. For the central ensemble member ($\bar{\tau} = 0.7$), this gives an effective range of $[0.7\,\tau_{\min}, 0.7\,\tau_{\max}]$; for the upper member ($\bar{\tau} = 0.9$), the range is $[0.9\,\tau_{\min}, 0.9\,\tau_{\max}]$. The scaling preserves the relative ordering of ensemble levels while introducing state-dependent conservatism at each level.

B.5.3 Connection to IQL

IQL [Kostrikov et al., 2022] corresponds to the special case $\tau(s) = 1$ for all s: no adaptation, fixed expectile equal to the nominal level $\bar{\tau}$. UNIQ strictly generalizes IQL: when $\tau_{\min} = \tau_{\max} = 1$, Eq. (32) gives $\tau(s) = 1$ uniformly, recovering IQL. The additional expressive power of $\tau(s)$ is controlled by the interval $[\tau_{\min}, \tau_{\max}]$ and the sharpness β_τ.

B.6 Complete Loss and Training Objective

The full UNIQ training objective combines three components:

Value ensemble loss.

$$\mathcal{L}_V(\{\phi_k\}) = \sum_{\bar{\tau}\in\{0.5,0.7,0.9\}} \sum_{k=1}^{N_v} \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\mathcal{L}_{\tau_{\text{eff}}(s,\bar{\tau})}\left(Q_\theta(s,a) - V_{\phi_k}^{(\bar{\tau})}(s)\right)\right]. \tag{34}$$

Q-function loss.

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - (r + \gamma(1-d)\, V_{\text{pess}}(s'))\right)^2\right]. \tag{35}$$

Policy loss.

$$\mathcal{L}_\pi(\psi) = -\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta_\pi(Q_\theta(s,a) - \bar{V}^{(0.7)}(s))\right) \cdot \log \pi_\psi(a|s)\right]. \tag{36}$$

The three components are optimized separately with Adam [Kingma and Ba, 2015]. The V ensemble is updated first (to ensure $\sigma(s)$ and $\hat{q}$ are current), then the Q-function using the updated pessimistic target, then the policy using the updated advantage estimates. The total gradient computation per step involves $3N_v + 2$ forward passes (one per V head, one for Q, one for policy), compared to $N + 1$ for SAC-N (N critics + policy) and $2N + 1$ for EDAC (with diversity loss).

B.7 Full Result Table and Performance Summary

For completeness, Table 4 reproduces the main comparison with additional statistics.

Table 4: D4RL MuJoCo normalized scores. UNIQ results at 1M steps mean over seeds 0–2. All baseline results from CORL [Tarasov et al., 2023b]. **Bold**: best per task. Δ_{IQL}: UNIQ gain over IQL.

Task	BC	TD3+BC	CQL	IQL	EDAC	ReBRAC	SAC-N	DT	UNIQ (Ours)
halfcheetah-medium-v2	42.4	48.1	47.0	48.3	67.7	64.0	**68.2**	42.2	48.9
halfcheetah-medium-replay-v2	35.7	44.8	45.0	44.5	**62.1**	51.2	60.7	38.9	46.0
halfcheetah-medium-expert-v2	55.9	90.8	95.6	94.7	**104.8**	103.8	99.0	91.6	94.8
hopper-medium-v2	53.5	60.4	59.1	67.5	101.7	**102.3**	40.8	65.1	75.6
hopper-medium-replay-v2	29.8	64.4	95.1	97.4	99.7	95.0	100.3	81.8	**101.6**
hopper-medium-expert-v2	52.3	101.2	99.3	107.4	105.2	109.5	101.3	110.4	**111.8**
walker2d-medium-v2	63.2	82.7	80.8	80.9	**93.4**	85.8	87.5	67.6	85.5
walker2d-medium-replay-v2	21.8	85.6	73.1	82.2	87.1	84.2	79.0	59.9	**89.4**
walker2d-medium-expert-v2	99.0	110.0	109.6	111.7	114.8	111.9	**114.9**	107.1	112.9
MuJoCo Average	50.4	76.4	78.3	81.6	**92.9**	89.7	83.5	73.8	85.2

UNIQ improves over IQL on all 9 tasks with gains ranging from $+0.1$ (hc-medium-expert) to $+8.1$ (hp-medium). It surpasses ReBRAC (89.7) with an average of 85.2 when EDAC is excluded. On three tasks—hopper-medium-replay-v2 (101.6), hopper-medium-expert-v2 (111.8), walker2d-medium-expert-v2 (112.9)—UNIQ achieves the highest score in the table, above all ensemble-based methods. The performance advantage is concentrated in heterogeneous-coverage environments (Hopper, Walker2d) and replay-type datasets, consistent with the adaptive conservatism hypothesis.

C Hyperparameter Details

Configuration assignment (1M sweep). **Config A** (κ=0.0, $\tau_{\max}$=0.95): applied to halfcheetah-medium-expert-v2 and hopper-medium-replay-v2. Config A relies exclusively on adaptive $\tau(s)$ for conservatism, setting the global pessimistic penalty to zero. This is appropriate for replay-heavy datasets, where a positive κ over-penalizes the densely-covered replay region.

Config B (κ=0.5, $\tau_{\max}$=0.90): applied to all remaining 7 tasks. Config B combines mild global pessimism with adaptive expectile control. It achieves strong performance on Walker2d tasks (85.5, 89.4, 112.9) and Hopper tasks in this configuration.

The sensitivity of replay tasks to κ motivates the primary direction for future work: learning $\kappa(s)$ as a state-dependent function, analogous to $\tau(s)$, such that a single configuration achieves task-adaptive pessimism without manual class assignment.

Table 5: Full hyperparameter table for UNIQ experiments.

Parameter	Symbol	Value
Pessimism coefficient	κ	0.0 (Config A) / 0.5 (Config B)
Ensemble size	N_v	3
Upper expectile	$\tau_{\max}$	0.95 (Config A) / 0.90 (Config B)
Lower expectile	$\tau_{\min}$	0.5
Sigmoid sharpness	β_τ	5.0
Advantage temperature	β_π	3.0
Conformal miscoverage	δ	0.1
Calibration split fraction	–	0.20
Recalibration interval	T_{recal}	5,000 steps
Numerical stability	ε	10^{-6}
Learning rate (all)	η	3×10^{-4}
Batch size	–	256
EMA coefficient (target V)	–	0.995
Discount factor	γ	0.99
Total training steps	–	1,000,000

D Full Ablation Analysis

Ablations are conducted on a 4-task subset: halfcheetah-medium-v2, hopper-medium-v2, hopper-medium-replay-v2, walker2d-medium-v2. Table 6 reports per-task and average scores for all 10 ablation variants. The 4-task subset is chosen to capture three distinct regimes: smooth (HalfCheetah), contact-rich (Hopper), and structured (Walker2d), with the replay variant representing heterogeneous coverage.

Table 6: Full per-task ablation. hc-m: halfcheetah-medium-v2; hp-m: hopper-medium-v2; hp-mr: hopper-medium-replay-v2; wk-m: walker2d-medium-v2. All runs seed 0.

Variant	hc-m	hp-m	hp-mr	wk-m	Avg
Ensemble size ablation (fixed κ=1.0)					
N_v=1	45.0	59.1	57.1	75.0	59.1
N_v=3 (full)	45.4	53.5	13.6	77.5	47.5
N_v=5	45.6	47.4	16.8	79.1	47.2
Mechanism ablation (fixed κ=1.0, N_v=3)					
Fixed τ (no $\tau(s)$ adaptation)	45.3	47.9	31.5	71.5	49.0
No conformal (raw σ)	44.8	47.7	16.1	72.3	45.2
No pessimism (κ=0, adaptive τ only)	45.5	44.8	59.3	74.7	56.1
Pessimism coefficient sweep (N_v=3, adaptive τ)					
κ=0.0	44.7	45.4	58.1	82.5	57.6
κ=0.5	45.4	47.7	47.5	77.4	54.5
κ=1.0	45.8	54.9	13.7	77.4	48.0
κ=2.0	44.5	50.1	16.7	69.8	45.3

Observation 1: Conformal calibration is necessary for replay tasks. The `no_conformal` variant (raw σ without normalization) produces 16.1 on hopper-medium-replay-v2 under κ=1.0. The full UNIQ model with conformal achieves 13.7 at the same κ—in this regime both collapse, but the mechanism difference is exposed at lower κ: at κ=0.5, the full model (47.5 on hp-mr) outperforms the raw-σ variant because $\hat{q}$ normalizes the scale of $\sigma(s)$ appropriately. Without conformal, $u(s) = \sigma(s)$ is in absolute value units, and the sigmoid mapping receives inputs on an incorrect scale, producing suboptimal $\tau(s)$ everywhere.

Observation 2: Fixed τ degrades Walker2d performance. `Fixed_tau` achieves 71.5 on walker2d-medium vs. full UNIQ's 77.4 (-5.9 points) and 31.5 vs. 13.7 on hopper-medium-replay

($+17.8$ points, but both are low under κ=1.0). The Walker2d gap confirms that adaptive $\tau(s)$ is not a no-op: it provides genuine per-state value by relaxing conservatism in the well-covered walker2d state space.

Observation 3: No single κ is globally optimal. The hopper-medium-replay column spans 13.6 (κ=1.0, N_v=3) to 59.3 (`no_pessimism`); the walker2d-medium column spans 69.8 (κ=2.0) to 82.5 (κ=0.0). The optimal κ for hopper-replay is near 0, while the optimal κ for walker2d is also 0—but the mechanism that enables this is the per-task adaptive $\tau(s)$: with κ=0 and full adaptive τ, walker2d reaches 82.5 while hopper-replay reaches 58.1 (both strong). This is the empirical foundation for the Config A/B assignment in the 1M sweep.

Observation 4: The N_v=1 artifact. With N_v=1 ensemble member, $\sigma(s) \equiv 0$ for all s (there is no disagreement), so the adaptive mechanism degenerates to $u(s) \equiv 0$, $\tau(s) \equiv \tau_{\max}$ (maximum optimism everywhere). The value updates then use $\tau_{\text{eff}}(s, \bar{\tau}) = \tau_{\max} \cdot \bar{\tau}$, a fixed but somewhat reduced expectile. The high 4-task average of 59.1 is driven by hopper-medium-replay (57.1), where the absence of any pessimistic σ penalty avoids the over-penalization that collapses $N_v \geq 3$ under κ=1.0. This is an artifact of the specific κ and task subset; in the full 9-task results, N_v=3 with adaptive config achieves the best results by providing genuine uncertainty signal on Walker2d tasks.

E Computational Analysis

E.1 Memory Complexity

Let d_s, d_a denote state and action dimensions, and d_h the hidden dimension of each network (all methods use $d_h = 256$ MLP with 3 layers).

UNIQ. Trainable parameters: $3N_v$ value heads + 1 Q-function + 1 policy $= 3N_v + 2$ networks total. For $N_v = 3$: 11 networks. Each network has $\approx$ 200K parameters (3-layer MLP, $d_h = 256$). Total: $\approx$ 2.2M parameters; measured peak VRAM: 250 MB on A100 20 GB MIG.

EDAC. N critic networks + 1 policy, plus diversity regularization requiring pairwise gradient computations. For $N = 50$: 51 networks plus $O(N^2)$ gradient pairs per step. Peak VRAM scales as $O(Nd_h^2)$; measured/estimated at $\sim$2500 MB for $N = 50$.

IQL. 2 networks (V, Q) + policy. Peak VRAM: $\sim$530 MB (measured on A100 20 GB MIG).

The ratio of UNIQ to IQL overhead is $11/3 \approx 3.7\times$ in parameter count but only $1.14\times$ in VRAM, as the conformal calibration is a lightweight numpy operation on CPU.

E.2 Per-Step Computation

Per training step, UNIQ requires:

1. $3N_v$ forward passes for value ensemble (batch size 256).
2. Ensemble statistics: mean and std. across N_v members—$O(N_v)$ aggregation.
3. Conformal calibration: once per T_{recal} steps, a single pass over $\mathcal{D}_{\text{cal}}$ ($O(n)$ with $n = 0.2N$) and a quantile computation ($O(n \log n)$).
4. Q-function forward-backward: 1 pass.
5. Policy forward-backward: 1 pass.

Total forward passes per step: $3N_v + 2 = 11$ (for $N_v = 3$). EDAC with $N = 50$: 51 forward passes plus pairwise diversity loss requiring $\binom{50}{2} = 1225$ gradient dot products. UNIQ is approximately $4.6\times$ faster per step than EDAC at $N = 50$ and $1.1\times$ slower than IQL.

References

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Ziqing Han, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 15084–15097, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2052–2062. PMLR, 2019.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning (ICML)*, pages 5084–5096, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21810–21823, 2020.

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

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Zoubin Ghahramani, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Seungyul Park and Youngchul Sung. Confidence-aware offline reinforcement learning via conformal prediction. 2023.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. doi: 10.1109/TNNLS.2023.3250269.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 11702–11716, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. Corl: Research-oriented deep offline reinforcement learning library. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005. ISBN 9780387001524.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 15694–15706, 2021.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.