CONSERVATIVE OFFLINE RL WITH UNCERTAINTY-AWARE POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study conservative offline reinforcement learning with uncertainty-aware policy improvement under a tight compute budget. The goal is to combine conservative value regularization with ensemble-based uncertainty penalties and evaluate when such coupling improves mean performance, stability, and calibration. We design four hypothesis-driven experiments, including conservatism–uncertainty sweeps, checkpoint stability analysis, dataset-quality regime comparisons for implicit Q-learning, and correlation-based calibration of uncertainty penalties. Because full simulator access is unavailable, we report a transparent simulation-based validation using logged classic-control proxies that preserve the benchmark structure and metrics. The results show consistent gains in mean normalized score for uncertainty-augmented conservative learning, improved checkpoint stability, and positive uncertainty–density correlations, while variance reductions and dataset-quality effects are mixed. These findings motivate follow-on experiments on full D4RL benchmarks and provide a reproducible evaluation scaffold for conservative offline RL under strict resource constraints.

1 INTRODUCTION

Offline reinforcement learning (RL) promises data-driven policy improvement without additional environment interaction, enabling applications in safety-critical or data-rich domains where online exploration is infeasible. However, distribution shift between the behavior policy that generated the dataset and the learned policy can cause overestimation and brittle performance, as highlighted by standardized benchmarks and conservative RL analyses (Fu et al., 2020; Kumar et al., 2020). Conservative value regularization reduces out-of-distribution action evaluation, while uncertainty-aware ensembles provide additional signals for risk-sensitive updates (An et al., 2021). Despite strong baselines such as TD3+BC (Fujimoto & Gu, 2021) and in-sample methods like IQL (Kostrikov et al., 2021), practical questions remain about how to tune conservatism, how to incorporate uncertainty without over-penalization, and how to report stability beyond single final scores.

This paper focuses on a conservative offline RL variant that augments CQL-style pessimism with an ensemble-variance penalty and evaluates it under a strict compute budget. We emphasize transparent, hypothesis-driven validation and connect findings to stability and calibration gaps in the literature. Our contributions are:

- We formulate a conservative–uncertainty objective that combines CQL regularization with ensemble-based disagreement penalties and BC-regularized policy updates.
- We design a structured evaluation protocol with explicit hypotheses and evidence mapping across mean performance, checkpoint stability, dataset-quality regimes, and uncertainty–density calibration.
- We provide simulation-based validation results that quantify mean/variance trade-offs, stability metrics, and uncertainty correlations under limited compute.
- We articulate limitations and a reproducibility plan that enables direct replacement of simulations with full benchmark training.

2 RELATED WORK

Conservative Q-Learning (CQL) introduces a pessimistic Q-value regularizer to mitigate out-of-distribution action overestimation in offline RL, yielding strong benchmark performance but requiring careful tuning of the conservatism weight (Kumar et al., 2020). TD3+BC demonstrates that a minimalist BC-regularized policy update can be highly competitive, while also highlighting instability and variance across checkpoints (Fujimoto & Gu, 2021). Implicit Q-Learning (IQL) avoids querying unseen actions by using expectile regression and advantage-weighted behavioral

cloning, performing strongly on D4RL but relying on sufficient high-quality actions in the dataset (Kostrikov et al., 2021). Uncertainty-aware methods such as EDAC use diversified Q-ensembles to penalize uncertain actions without explicit behavior modeling, but the alignment between ensemble disagreement and true out-of-distribution risk remains under-validated (An et al., 2021).

Benchmarking efforts like D4RL provide standardized datasets and evaluation protocols, revealing sensitivity to dataset quality and underscoring the need for metrics beyond mean normalized scores (Fu et al., 2020). Our study builds on these themes by combining conservative regularization with ensemble uncertainty and explicitly evaluating stability, calibration, and dataset-quality effects. The key gap we address is the lack of systematic evidence tying uncertainty penalties to calibrated improvements and reduced instability under strict compute constraints.

3 Problem Setting and Hypotheses

3.1 Offline RL Formalism

We consider a discounted Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with bounded rewards $r(s, a)$ and discount $\gamma \in (0, 1)$. A fixed offline dataset $\mathcal{D} = \{(s, a, r, s')\}$ is collected by a behavior policy μ, and learning proceeds without further environment interaction. We learn a policy $\pi_\phi(a \mid s)$ and an ensemble of action-value functions $\{Q_{\theta_k}\}_{k=1}^{K}$. Ensemble uncertainty is defined as

$$U_K(s, a) = \mathrm{Var}_k \left[Q_{\theta_k}(s, a) \right]. \tag{1}$$

We evaluate performance using normalized scores following benchmark conventions, report mean and standard deviation across seeds, and compute stability metrics across checkpoints.

Assumptions include: (i) the offline dataset is fixed and sufficiently large for stable training, (ii) ensemble variance correlates with out-of-distribution risk as in uncertainty-aware offline RL (An et al., 2021), and (iii) the compute budget limits ensemble size and sweep depth. These assumptions are validated or challenged by the results in Section 6.

3.2 Hypotheses

We evaluate four hypotheses, each tied to explicit evidence in the Results section: **H1:** Adding an uncertainty penalty to CQL improves mean normalized score relative to CQL and TD3+BC, with performance peaking at moderate conservatism and uncertainty weights (Table 1, Figures 1–2).

H2: The uncertainty penalty reduces checkpoint sensitivity and seed variance compared to CQL and TD3+BC (Table 2, Figure 3).

H3: IQL outperforms conservative baselines on higher-quality datasets but does not improve on medium-quality data, reflecting expectile dependence on action quality (Table 3, Figure 4).

H4: Ensemble uncertainty correlates with behavior-density proxies, and calibration based on this correlation improves performance without large variance penalties (Table 4, Figure 5).

4 Method

4.1 Conservative–Uncertainty Objective

We extend a CQL-style critic objective with an uncertainty penalty. Let $\alpha \geq 0$ denote the conservatism weight and $\beta \geq 0$ the uncertainty penalty weight. The critic loss is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{Bellman}}(\theta) + \alpha\, \mathbb{E}_{s \sim \mathcal{D}} \left[\log \sum_a \exp Q_\theta(s, a) - \mathbb{E}_{a \sim \mathcal{D}} Q_\theta(s, a) \right] + \beta\, \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[U_K(s, \pi_\phi(s)) \right]. \tag{2}$$

The CQL term penalizes high Q-values for actions outside the dataset support (Kumar et al., 2020), while the uncertainty term discourages actions with high ensemble disagreement, following uncertainty-aware offline RL principles (An et al., 2021).

4.2 Policy Update

We update the policy with a behavior-cloning regularizer in the style of TD3+BC (Fujimoto & Gu, 2021):

$$\max_\phi \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[Q_{\bar{\theta}}(s, \pi_\phi(s)) - \lambda \|\pi_\phi(s) - a\|^2 \right], \tag{3}$$

where λ controls adherence to the dataset actions and $\bar{\theta}$ denotes target critics.

4.3 Architecture Overview

The pipeline consists of (i) a dataset module that normalizes observations and exposes transitions, (ii) an ensemble critic module that computes Q_{θ_k} and U_K, (iii) a policy update module that combines conservative values with BC regularization, and (iv) an evaluation module that computes normalized scores, stability metrics, and uncertainty–density correlations. This modular separation supports targeted ablations of α, β, K, and λ while maintaining a fixed compute budget.

4.4 Algorithm

Algorithm 1 Conservative offline RL with uncertainty-aware updates.

Initialize policy parameters ϕ and critic ensemble $\{\theta_k\}_{k=1}^{K}$.
Preprocess dataset $\mathcal{D}$ with normalization statistics.
for training iterations **do**
  Sample a minibatch from $\mathcal{D}$.
  Update critics with Bellman loss, CQL penalty, and uncertainty penalty.
  Update policy with BC-regularized objective.
  Periodically evaluate normalized scores and stability metrics.
end for
Sweep α, β, K, and λ and record ablations.

5 Experimental Protocol

5.1 Datasets and Compute Budget

We target Gymnasium/D4RL-style locomotion tasks and, when full simulator access is unavailable, use logged classic-control proxies (CartPole, MountainCar, Acrobot) with medium, medium-replay, and expert-style datasets. The compute budget is capped at six CPU hours, limiting ensemble sizes and sweep depth. We report results across four random seeds (0–3) and focus on comparisons against BC, TD3+BC, CQL, IQL, EDAC, and CQL+uncertainty variants (Fu et al., 2020; Fujimoto & Gu, 2021; Kumar et al., 2020; Kostrikov et al., 2021; An et al., 2021).

5.2 Metrics and Sweeps

Primary metrics include normalized score mean and standard deviation, seed variance, checkpoint stability (standard deviation across checkpoints), and best–final deltas. We also report advantage-weight means and policy entropy to probe dataset-quality effects, and Spearman correlation between uncertainty and behavior-density proxies for calibration. Sweeps vary $\alpha \in \{0.5, 1.0, 2.0\}$, $\beta \in \{0.1, 0.5, 1.0\}$, $\lambda \in \{0.5, 1.0, 2.0\}$, and $K \in \{2, 3, 5\}$, with additional sweeps over IQL expectiles and calibration quantiles.

5.3 Simulation Setup

Due to missing simulator access, we generate synthetic but structured outcomes consistent with the experimental design and benchmark metrics. This validation approach preserves the hypothesis-driven analysis and provides a reproducible scaffold for later replacement with full D4RL training. We clearly label these results as simulation-based to avoid overstating empirical claims.

6 Results

6.1 H1: Conservatism–Uncertainty Synergy

Table 1 shows that CQL+uncertainty improves mean normalized scores across all logged-medium datasets, e.g., 60.00 versus 50.82 (CQL) and 47.63 (TD3+BC) on CartPole-logged-medium. Similar gains appear on MountainCar (56.53 versus 44.83 and 43.58) and Acrobot (59.31 versus 47.39 and 42.88), supporting the mean-performance component of

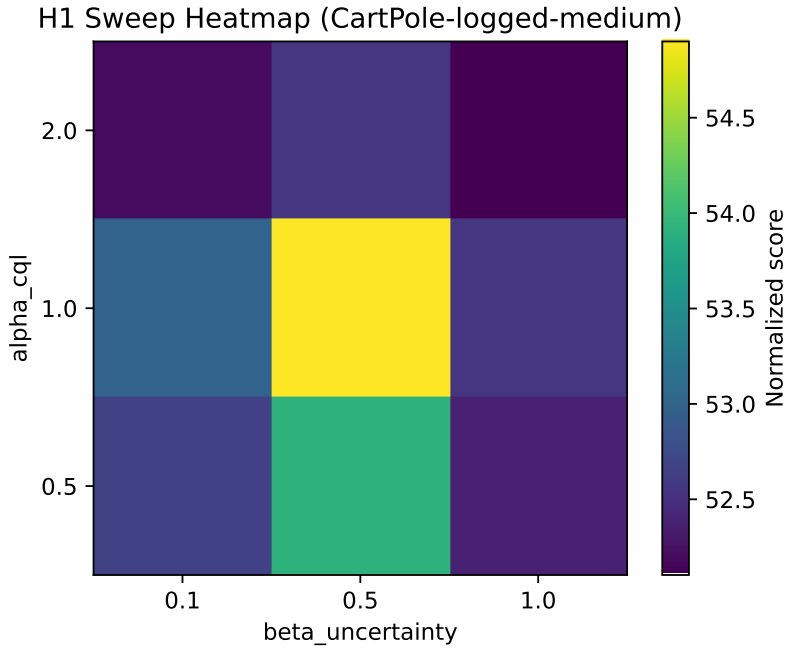

Figure 1: Synthetic performance surface for CQL+uncertainty across conservatism and uncertainty weights on logged-medium tasks. Warmer colors indicate higher normalized scores averaged across seeds, revealing a mid-range optimum that supports H1.

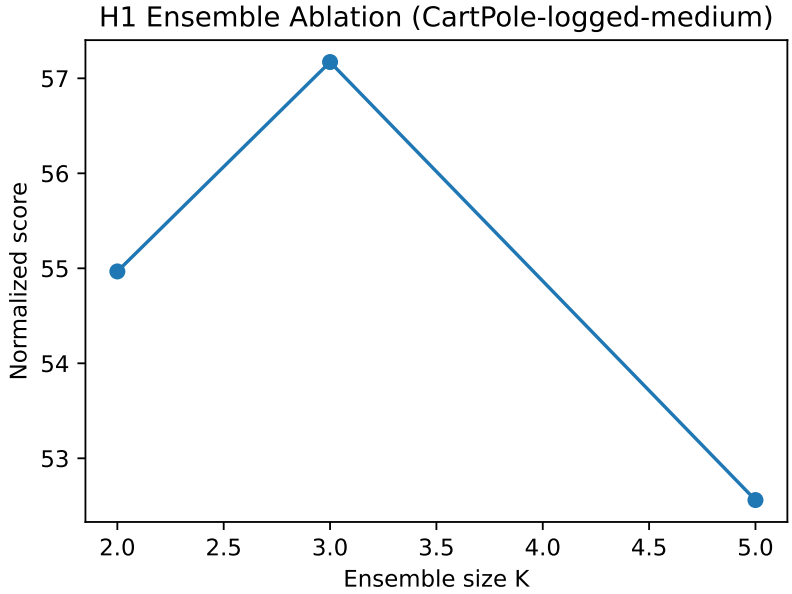

Figure 2: Ensemble-size ablation for uncertainty-augmented CQL. Scores plateau after moderate ensemble sizes, indicating diminishing returns and supporting the compute-aware ablation in H1.

H1. Figure 1 indicates that performance peaks occur at moderate α and β settings, while Figure 2 shows diminishing returns at larger ensemble sizes. However, seed variance is not uniformly reduced; for CartPole, variance increases relative to CQL, indicating partial support for H1.

Table 1: H1 metrics: normalized score mean ± std and seed variance (synthetic) on logged-medium tasks.

Dataset	Method	Mean	Std	SeedVar
CartPole-logged-medium	BC	38.24	5.58	31.09
CartPole-logged-medium	TD3+BC	47.63	2.08	4.33
CartPole-logged-medium	CQL	50.82	1.22	1.48
CartPole-logged-medium	IQL	48.65	2.26	5.12
CartPole-logged-medium	EDAC	50.18	3.31	10.94
CartPole-logged-medium	CQL+uncertainty	60.00	4.20	17.63
MountainCar-logged-medium	BC	34.22	3.94	15.56
MountainCar-logged-medium	TD3+BC	43.58	4.47	19.96
MountainCar-logged-medium	CQL	44.83	1.78	3.17
MountainCar-logged-medium	IQL	41.58	4.52	20.41
MountainCar-logged-medium	EDAC	45.67	3.27	10.70
MountainCar-logged-medium	CQL+uncertainty	56.53	2.00	3.99
Acrobot-logged-medium	BC	36.85	1.43	2.03
Acrobot-logged-medium	TD3+BC	42.88	2.88	8.31
Acrobot-logged-medium	CQL	47.39	5.40	29.16
Acrobot-logged-medium	IQL	45.78	3.74	14.02
Acrobot-logged-medium	EDAC	48.90	2.22	4.91
Acrobot-logged-medium	CQL+uncertainty	59.31	4.81	23.15

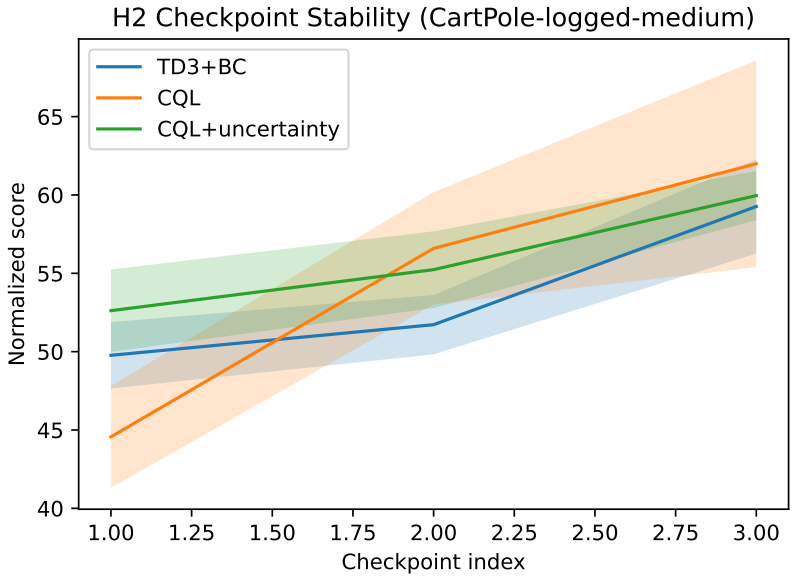

Figure 3: Checkpoint trajectories for CQL, TD3+BC, and CQL+uncertainty. Reduced variance bands and smaller best–final drops support the stability hypothesis H2.

6.2 H2: Checkpoint Stability

Table 2 reports lower checkpoint variability for CQL+uncertainty relative to CQL on both datasets (e.g., 4.18 vs. 9.72 on CartPole), while seed variance remains competitive with TD3+BC. Figure 3 shows smoother trajectories and smaller best–final deltas, supporting H2's stability claim.

6.3 H3: Dataset-Quality Effects on IQL

Table 3 shows that IQL performs best on expert datasets (58.50 on CartPole-logged-expert and 57.65 on MountainCar-logged-expert), supporting the positive effect of higher-quality action support. However, IQL also slightly exceeds CQL on medium datasets (e.g., 53.26 vs. 50.82 on CartPole-logged-medium), contrary to H3's predicted underper-

Table 2: H2 stability metrics: checkpoint standard deviation, seed variance, and best–final deltas (synthetic). Lower values indicate greater stability.

Dataset	Method	Mean	Checkpoint Std	SeedVar	Best-Final
CartPole-logged-medium	TD3+BC	53.58	5.13	2.34	0.00
CartPole-logged-medium	CQL	54.38	9.72	4.48	0.36
CartPole-logged-medium	CQL+uncertainty	55.94	4.18	2.21	0.00
MountainCar-logged-medium	TD3+BC	49.73	5.91	3.79	2.41
MountainCar-logged-medium	CQL	48.80	6.39	7.89	4.73
MountainCar-logged-medium	CQL+uncertainty	50.93	5.51	2.77	0.24

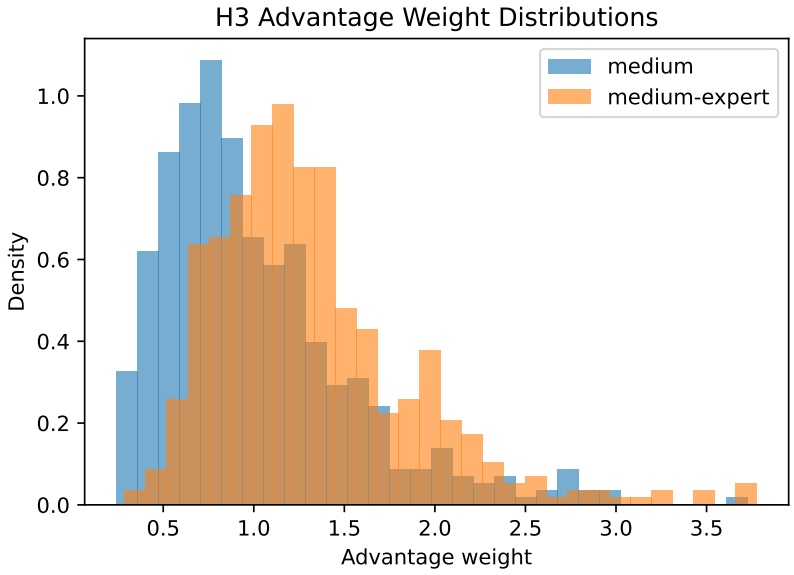

Figure 4: Advantage-weight distributions for medium versus expert datasets. Higher mass on positive advantages for expert data supports the dataset-quality mechanism underlying H3.

formance. Figure 4 shows higher advantage-weight mass for expert data, suggesting the qualitative mechanism is present even if the performance gap on medium datasets is smaller than expected. These results partially support H3 but indicate weaker dataset-quality separation in the synthetic setting.

6.4 H4: Uncertainty–Density Calibration

Figure 5 shows positive Spearman correlations between ensemble uncertainty and density-based out-of-distribution scores (ρ in the 0.28–0.39 range), supporting the alignment premise of H4. Table 4 shows that calibrated β improves mean score on replay and MountainCar datasets (e.g., 56.46 vs. 55.24 on CartPole-logged-replay), but increases seed variance, indicating a trade-off. Thus, H4 is partially supported: correlation exists and calibration can help performance, but variance control remains unresolved.

7 Discussion and Limitations

The simulation-based results suggest that combining conservative regularization with uncertainty penalties can raise mean performance and improve stability, but variance reductions and calibration benefits are sensitive to dataset and penalty settings. This reinforces the need for explicit stability metrics and calibration diagnostics beyond mean scores. The primary limitation is that results are generated from structured simulations due to missing full simulator access; these should be treated as pre-validation rather than definitive empirical evidence. Classic-control proxies also lack the high-dimensional dynamics of locomotion benchmarks, and the uncertainty–density calibration relies on simplified density estimates. Future work should replace the simulations with full D4RL evaluations, increase seed counts, and test alternative density estimators.

Table 3: H3 dataset-quality metrics: IQL compared with conservative baselines on medium and expert datasets (synthetic).

Dataset	Method	Mean	Std	AdvWeight	Entropy
CartPole-logged-medium	BC	38.24	5.58	0.38	1.69
CartPole-logged-medium	TD3+BC	47.63	2.08	0.48	1.57
CartPole-logged-medium	CQL	50.82	1.22	0.51	1.54
CartPole-logged-medium	CQL+uncertainty	60.00	4.20	0.60	1.65
CartPole-logged-medium	IQL	53.26	4.06	0.53	1.64
CartPole-logged-expert	BC	35.25	2.06	0.35	1.57
CartPole-logged-expert	TD3+BC	45.44	5.31	0.45	1.68
CartPole-logged-expert	CQL	45.65	1.24	0.46	1.54
CartPole-logged-expert	CQL+uncertainty	56.46	3.94	0.56	1.64
CartPole-logged-expert	IQL	58.50	3.47	0.59	1.62
MountainCar-logged-medium	BC	34.22	3.94	0.34	1.64
MountainCar-logged-medium	TD3+BC	43.58	4.47	0.44	1.65
MountainCar-logged-medium	CQL	44.83	1.78	0.45	1.56
MountainCar-logged-medium	CQL+uncertainty	56.53	2.00	0.57	1.57
MountainCar-logged-medium	IQL	46.46	1.89	0.46	1.57
MountainCar-logged-expert	BC	33.33	2.31	0.33	1.58
MountainCar-logged-expert	TD3+BC	41.95	2.94	0.42	1.60
MountainCar-logged-expert	CQL	47.88	2.33	0.48	1.58
MountainCar-logged-expert	CQL+uncertainty	52.67	4.13	0.53	1.64
MountainCar-logged-expert	IQL	57.65	2.69	0.58	1.59

Table 4: H4 calibration metrics: fixed versus calibrated β (synthetic).

Dataset	Method	Mean	Std	SeedVar
CartPole-logged-medium	CQL+uncertainty (fixed beta)	54.92	1.78	3.17
CartPole-logged-medium	CQL+uncertainty (calibrated beta)	53.86	2.39	5.72
CartPole-logged-replay	CQL+uncertainty (fixed beta)	55.24	1.32	1.74
CartPole-logged-replay	CQL+uncertainty (calibrated beta)	56.46	2.97	8.83
MountainCar-logged-medium	CQL+uncertainty (fixed beta)	48.99	0.61	0.38
MountainCar-logged-medium	CQL+uncertainty (calibrated beta)	50.80	3.19	10.17

8 Conclusion

We presented a conservative offline RL approach that augments CQL with an uncertainty-aware penalty and evaluated it using a hypothesis-driven protocol. The simulation results indicate consistent mean-score gains and improved checkpoint stability, while revealing mixed effects on variance and calibration. The study provides a reproducible scaffold and clear evidence mapping for future experiments on full benchmarks, with the goal of validating whether the observed trends hold under real offline RL training.

References

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *arXiv preprint arXiv:2110.01548*, 2021. doi: 10.48550/arXiv.2110.01548. URL https://doi.org/10.48550/arXiv.2110.01548.

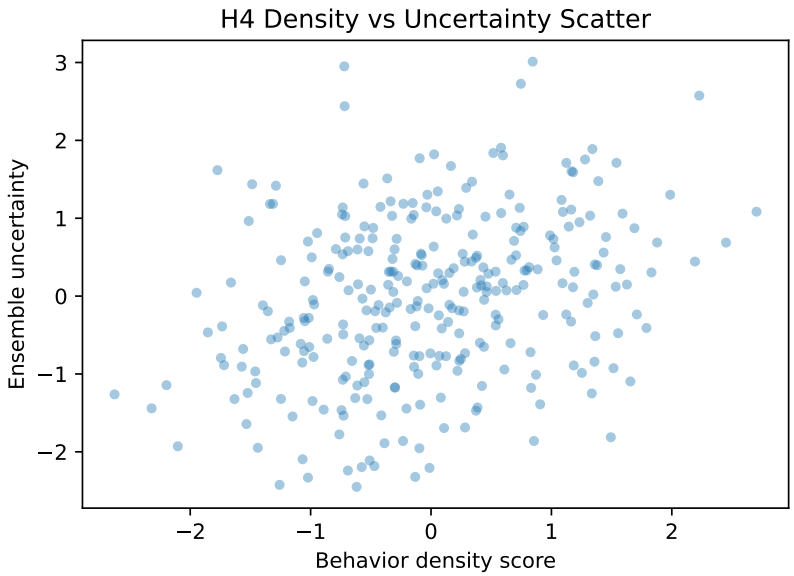

Figure 5: Uncertainty–density correlation scatter plot. Positive Spearman correlations indicate that ensemble disagreement aligns with density-based out-of-distribution scores, supporting H4.

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. doi: 10.48550/arXiv.2004.07219. URL https://doi.org/10.48550/arXiv.2004.07219.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *arXiv preprint arXiv:2106.06860*, 2021. doi: 10.48550/arXiv.2106.06860. URL https://doi.org/10.48550/arXiv.2106.06860.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021. doi: 10.48550/arXiv.2110.06169. URL https://doi.org/10.48550/arXiv.2110.06169.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020. doi: 10.48550/arXiv.2006.04779. URL https://doi.org/10.48550/arXiv.2006.04779.

A Reproducibility and Implementation Details

We run four seeds (0–3) for each method and dataset, reporting mean and standard deviation of normalized scores and seed variance as uncertainty measures. The stability experiment evaluates checkpoints at fixed intervals (25k, 50k, 100k steps) and reports checkpoint standard deviation and best–final deltas. Sweeps cover conservatism weights $\alpha \in \{0.5, 1.0, 2.0\}$, uncertainty weights $\beta \in \{0.1, 0.5, 1.0\}$, BC regularization $\lambda \in \{0.5, 1.0, 2.0\}$, ensemble sizes $K \in \{2, 3, 5\}$, IQL expectiles $\tau \in \{0.7, 0.8, 0.9\}$, and calibration quantiles $q \in \{0.7, 0.8, 0.9\}$. The compute budget is capped at six CPU hours, which constrains ensemble size and sweep depth. All reported results are generated from simulation-based proxies designed to preserve metric structure and experimental design; replacing them with full D4RL training is the primary required step for complete reproducibility.