Transfer-Ready Critics: Auditing Conservatism Footprints for Offline-to-Online RL

Anonymous Authors[1]

Abstract

Offline-to-online reinforcement learning often transfers conservative critic ensembles as intact initialization objects. In REDQ-style fine-tuning, however, online Bellman targets are constructed from randomly sampled critic heads, raising a checkpoint-level question: *which conservatively pretrained heads should be trusted after transfer?* We introduce FOCUS (Footprint-based Offline-to-online Critic Selection), a transition-time audit that scores each critic head by its conservative Bellman footprint and target exposure before online updates begin. Matched seed-0 AntMaze medium- and large-diverse checkpoint audits reveal that critic-head footprints are not static: heterogeneity ranges from nearly homogeneous to severe-outlier regimes, and the highest-footprint head changes across pretraining. This suggests that offline policy success alone can hide risk in the transferred critic target generator. Reduced-warmup experiments show a sharper lesson: forcing high-footprint heads into REDQ targets can substantially destabilize transfer at a high-heterogeneity checkpoint, while aggressive low-footprint half-ensemble selection is mixed. FOCUS therefore exposes a risk-diversity trade-off in how conservative critic audits should be used, rather than merely proposing another fixed ensemble selector.

1. Introduction

Offline-to-online (O2O) reinforcement learning initializes from an offline dataset and then improves through online interaction. This setting is attractive when online exploration is expensive, but it is also fragile: conservative offline critics can be miscalibrated once the policy begins to move beyond the dataset. Existing work addresses this boundary through conservative value initialization, replay mixing, warmup rollouts, uncertainty-aware objectives, or ensemble aggregation (Kumar et al., 2020; Nakamoto et al., 2023; Lee et al., 2022; Ball et al., 2023; Zhou et al., 2025; Guo et al., 2023; Werge et al., 2025).

Modern O2O pipelines often transfer not a single critic, but a REDQ ensemble, where online Bellman targets are formed by subsampling heads and taking a minimum (Chen et al., 2021). WSRL stabilizes this boundary with warmup rollouts, but it still treats the restored REDQ heads as an exchangeable target pool. We ask a more specific checkpoint-internal question: *after conservative offline pretraining, which critic heads should be trusted to construct online Bellman targets?*

We introduce FOCUS (Footprint-based Offline-to-online Critic Selection), a checkpoint-level audit-and-selection procedure for conservatively pretrained REDQ critics. Each head is scored by a conservative Bellman footprint on offline calibration data, together with a min-target exposure statistic. The audit reveals that critic-head structure is checkpoint-dependent: some checkpoints are nearly homogeneous, while others contain severe high-footprint outliers; moreover, the identity of the outlier head changes across pretraining. This makes a fixed head-index rule implausible and motivates a per-checkpoint transfer audit.

As a minimal intervention study, FOCUS constructs alternative online target pools from the audit ranking while leaving the actor, optimizer, replay logic, and warmup protocol unchanged. We evaluate two deliberately simple stress tests: FOCUS-LOW, which keeps the lowest-footprint half of the ensemble, and FOCUS-HIGH, which keeps the highest-footprint half. These are not meant to be final deployment rules; they test whether the audit ranking identifies target pools that can change online transfer behavior.

Contributions. (i) We formulate *critic transfer-readiness* as a checkpoint-level diagnostic problem for O2O RL. (ii) We define per-head conservative footprints and min-target exposure for conservatively pretrained REDQ critics. (iii) We show in matched seed-0 medium- and large-diverse checkpoint audits that footprint heterogeneity varies sub-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

stantially across pretraining and that outlier head identity is transient. (iv) We test low- and high-footprint target-pool interventions under reduced warmup, finding a clean risky-pool failure mode at one high-heterogeneity checkpoint and mixed behavior elsewhere, which exposes a risk-diversity tradeoff for future audit-guided methods.

2. Method: Conservative-Footprint Audit and Selection

Let a frozen offline checkpoint contain policy π^{off} and REDQ critics $\{Q_h^{\text{off}}\}_{h=1}^H$, with head set $[H] = \{1, \dots, H\}$. Let $\mathcal{C} = \{(s_i, a_i, r_i, s_i', m_i)\}_{i=1}^N$ be an offline calibration batch, where $m_i \in \{0, 1\}$ is the continuation mask. FOCUS is applied once, before online fine-tuning.

Checkpoint-internal Bellman reference. For each transition, define the frozen reference

$$z_i^{\text{off}} = r_i + \gamma m_i \min_{j \in [H]} Q_j^{\text{off}}\big(s_i', \pi^{\text{off}}(s_i')\big). \tag{1}$$

This is not an oracle target; it is an internal reference used to compare heads under the same frozen checkpoint.

Per-head conservative footprint. For each head h, define the one-sided pessimistic Bellman gap and Bellman inconsistency

$$p_h = \frac{1}{N} \sum_{i=1}^N \big(z_i^{\text{off}} - Q_h^{\text{off}}(s_i, a_i)\big)_+, \tag{2}$$

$$e_h = \frac{1}{N} \sum_{i=1}^N \big(Q_h^{\text{off}}(s_i, a_i) - z_i^{\text{off}}\big)^2. \tag{3}$$

The footprint score is

$$\rho_h = p_h + \lambda_e e_h, \qquad \lambda_e = 0.1. \tag{4}$$

To summarize checkpoint-level heterogeneity, we use

$$\text{CV}(\rho) = \frac{\text{Std}_{h \in [H]}(\rho_h)}{\text{Mean}_{h \in [H]}(\rho_h) + \varepsilon}. \tag{5}$$

In our study, $\text{CV}(\rho) < 0.05$ is treated as no-signal, while $\text{CV}(\rho) > 0.15$ indicates strong heterogeneity.

Min-target exposure. A head with large ρ_h need not dominate REDQ targets. Let M be the online REDQ subset size ($M = 2$ here), and let $\mathcal{U}_M([H])$ denote the uniform distribution over size-M subsets of $[H]$. For $q_{ij} = Q_j^{\text{off}}(s_i', \pi^{\text{off}}(s_i'))$, define the exposure score

$$d_h = \frac{1}{N} \sum_{i=1}^N \Pr_{\mathcal{J} \sim \mathcal{U}_M([H])} \left(h \in \arg\min_{j \in \mathcal{J}} q_{ij} \right), \tag{6}$$

Algorithm 1 FOCUS audit and transition-time head-pool selection

Input: frozen checkpoint $(\pi^{\text{off}}, \{Q_h^{\text{off}}\})$, calibration batch $\mathcal{C}$, threshold τ, pool size K, subset size M
Compute z_i^{off} for all i using Eq. (1)
for each head $h \in [H]$ **do**
  Compute (p_h, e_h, ρ_h) using Eqs. (2)–(4)
  Estimate d_h by Monte Carlo subset sampling and set $\eta_h = \rho_h \tilde{d}_h$
end for
Compute $\text{CV}(\rho)$ and rank heads by η_h
Construct stress-test pools $\mathcal{H}^{\text{low}}$ and $\mathcal{H}^{\text{high}}$ from the ranking
Compare standard full-head REDQ targets against $\mathcal{H}^{\text{low}}$ and $\mathcal{H}^{\text{high}}$ target pools
Use the results to diagnose whether the checkpoint contains risky or useful footprint structure

where the event is false when $h \notin \mathcal{J}$, and ties are averaged uniformly in implementation. We normalize by the uniform baseline $1/H$ and set

$$\tilde{d}_h = \frac{d_h}{1/H}, \qquad \eta_h = \rho_h \tilde{d}_h. \tag{7}$$

When d_h is nearly uniform, $\eta_h \approx \rho_h$; when exposure is skewed, η_h upweights high-footprint heads that are more frequently exposed to min-target selection.

Transition-time head-pool selection. Given a pool size K, the audit induces two stress-test pools:

$$\mathcal{H}^{\text{low}} = \text{TopK}\big(\{-\eta_h\}_{h=1}^H, K\big), \tag{8}$$

$$\mathcal{H}^{\text{high}} = \text{TopK}\big(\{\eta_h\}_{h=1}^H, K\big). \tag{9}$$

Online REDQ targets then sample $\mathcal{J} \sim \mathcal{U}_M(\mathcal{H})$ from either $\mathcal{H} = [H]$ for standard WSRL, $\mathcal{H} = \mathcal{H}^{\text{low}}$ for FOCUS-LOW, or $\mathcal{H} = \mathcal{H}^{\text{high}}$ for FOCUS-HIGH:

$$y = r + \gamma m \min_{j \in \mathcal{J}} Q_{\bar{\phi}_j}(s', \pi_\theta(s')). \tag{10}$$

Thus, FOCUS does not relearn aggregation weights or alter the actor objective; it only changes which pretrained heads are permitted to generate early online targets. In the present experiments we use $K = 5$ to stress-test the ranking; the results below show that such half-ensemble selection is informative but not always the right practical intervention.

3. Experiments

Experimental setup. We study a seed-0 CAL-QL+REDQ run on AntMaze medium-diverse for online stress tests, and a matched seed-0 AntMaze large-diverse run for diagnostic robustness. Checkpoint audits use 50K calibration transitions. Online runs use

Table 1. Seed-0 checkpoint audit. High-success checkpoints can still contain severe critic-head footprint outliers. The outlier identity is checkpoint-dependent.

Step	Succ.	Mean ρ	CV(ρ)	Corr(ρ, d)	Max ρ	Max head
550K	0.65	3.629	**0.635**	+0.386	**10.527**	6
600K	0.45	2.723	0.050	-0.464	3.031	2
650K	0.40	3.044	0.147	+0.197	3.888	8
700K	0.50	3.202	0.275	+0.492	5.763	9
750K	0.65	2.324	0.130	+0.842	3.177	4
800K	0.60	2.962	0.377	+0.049	6.278	8
850K	0.65	2.545	**0.334**	-0.739	5.075	5
900K	0.55	2.432	0.038	+0.441	2.666	6

Large-diverse diagnostic-only sweep over the matched 550K–900K window: CV(ρ) ranges from 0.054 to 0.824, max ρ reaches 8.77, the max-footprint head changes in 7/7 consecutive pairs, and 6/10 heads become max-footprint at least once.

a shared reduced-warmup protocol: restored checkpoint, 200K online steps, warmup 1250 steps, update-to-data ratio 4, batch size 1024, reward scale/bias 10/-5, no online CQL loss, ten critic heads, target subset size two, and 20 evaluation trajectories every 5K steps. The only online difference is the target-head pool: standard WSRL samples from all heads; FOCUS-LOW and FOCUS-HIGH restrict sampling to the lowest- or highest-η five heads, respectively.

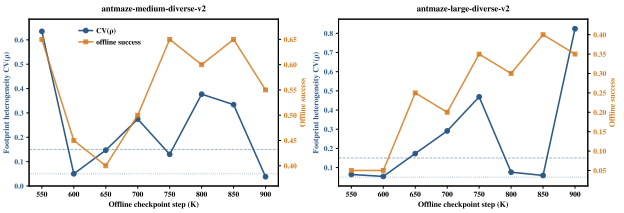
3.1. Checkpoint-level audit

We first sweep the 550K–900K checkpoint window and ask whether similar offline policy success implies similar critic transfer-readiness. Table 1 shows that the answer is no. High-success checkpoints can still differ markedly in critic-head structure: for instance, 550K and 850K both have success ≈ 0.65, but their heterogeneity and outlier anatomy differ substantially. Moreover, the highest-footprint head is not fixed; it changes across pretraining. This turns the central story from "one rogue head" into a stronger phenomenon: *critic-footprint structure is a transient checkpoint property.*

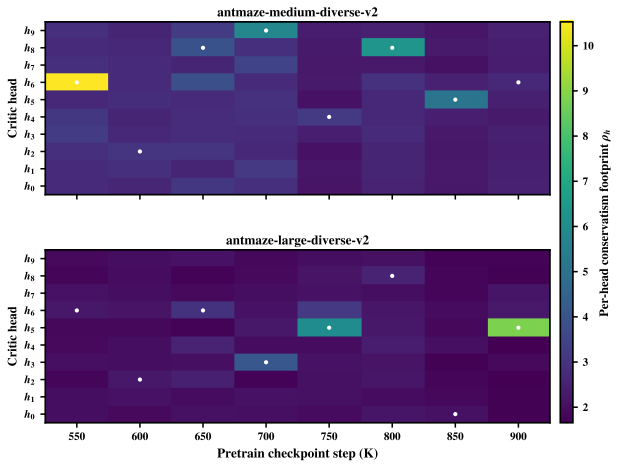
As a diagnostic robustness check, we repeat the audit on AntMaze large-diverse over the same 550K–900K window. The transient-footprint pattern persists: in both environments, the maximum-footprint head changes at every consecutive checkpoint pair (7/7 transitions), 6 of 10 heads become the maximum-footprint head at least once, and half of checkpoints exceed the strong-heterogeneity threshold CV(ρ) > 0.15. Online stress tests remain limited to medium-diverse.

3.2. Reduced-warmup online transfer

Checkpoint selection. Medium online checkpoints are selected from the same 550K–900K window; large-diverse is used only as diagnostic replication. We do not fine-tune all eight medium checkpoints online. Each 200K reduced-warmup run costs nontrivial GPU time, so we select four



(a) Footprint heterogeneity and offline success for medium-diverse (left) and large-diverse (right).



(b) Per-head footprint heatmap across matched medium- and large-diverse windows.

Figure 1. Checkpoint-internal footprint dynamics. Color denotes per-head conservatism footprint ρ_h; white dots mark the maximum-footprint head at each checkpoint. Across the same 550K–900K window, both environments show transient outlier identities: the maximum-footprint head changes at every consecutive checkpoint pair, and 6 of 10 heads become the maximum at least once. This suggests that footprint heterogeneity is not a medium-diverse artifact, while online stress tests are performed only on medium-diverse.

representative checkpoints that span the audit regimes. We choose 550K and 850K as heterogeneous intervention checkpoints (CV(ρ) $= 0.635$ and 0.334), and 600K and 900K as low-CV controls (CV(ρ) $= 0.050$ and 0.038). This choice is deliberate rather than arbitrary: 550K and 850K have similar offline success (≈ 0.65) but different critic structure, while 600K and 900K test whether footprint-based intervention becomes weak when the audit signal is nearly absent.

Metrics. Besides full-trajectory metrics, we report post-burn-in summaries that better match the visual transition pattern. Let $\text{AUC}^{20\text{K}+}$ denote success-rate area-under-curve after the first 20K online steps; $\text{Std}^{20\text{K}+}$ is the standard deviation over the same window; $\text{Frac}^{20\text{K}+}_{\geq 0.9}$ is the fraction of evaluations after 20K with success at least 0.9; and Final is the success rate at the last evaluation. These post-burn-in metrics discount the first few evaluations, which can contain

Table 2. Seed-0 reduced-warmup online transfer on selected checkpoints. FOCUS-LOW and FOCUS-HIGH use $K{=}5$ target pools. Metrics are computed after a 20K-step burn-in except Final.

Ckpt	CV(ρ)	Method	AUC$^{20K+}\uparrow$	Std$^{20K+}\downarrow$	Frac$^{20K+}_{\geq 0.9}\uparrow$	Final $\uparrow$
550K	0.635	WSRL	0.895	**0.110**	0.703	**1.00**
		FOCUS-LOW	0.881	0.131	0.622	0.85
		FOCUS-HIGH	**0.916**	0.118	**0.784**	0.95
600K	0.050	WSRL	0.818	0.186	0.432	0.95
		FOCUS-LOW	**0.846**	**0.181**	0.541	0.85
		FOCUS-HIGH	0.812	0.289	**0.676**	**1.00**
850K	0.334	WSRL	0.938	**0.043**	0.892	0.95
		FOCUS-LOW	**0.961**	0.047	**0.919**	**1.00**
		FOCUS-HIGH	0.853	0.120	0.514	0.95
900K	0.038	WSRL	0.926	0.069	0.811	0.95
		FOCUS-LOW	0.900	0.185	0.784	0.95
		FOCUS-HIGH	**0.954**	**0.059**	**0.865**	**1.00**

a short adaptation shock after changing the target-head pool.

The cleanest mechanism case is 850K. Here FOCUS-LOW remains competitive with standard WSRL and slightly improves post-burn-in occupancy ($\text{Frac}^{20\text{K}+}_{\geq 0.9} = 0.919$ vs. 0.892), whereas FOCUS-HIGH substantially degrades transfer ($\text{AUC}^{20\text{K}+} = 0.853$, $\text{Frac}^{20\text{K}+}_{\geq 0.9} = 0.514$, $\text{Std}^{20\text{K}+} = 0.120$). This supports a narrow but important claim: at a checkpoint with meaningful footprint heterogeneity, the audit can identify a risky head pool whose forced use hurts online adaptation.

The remaining checkpoints refine this claim rather than overturn it. At 600K and 900K, where CV(ρ) is near the no-signal regime, no consistent low-footprint advantage appears; this is consistent with the audit becoming weak when critic heads are nearly homogeneous. At 550K, despite very high heterogeneity, the predicted ordering breaks: FOCUS-HIGH slightly exceeds WSRL, while FOCUS-LOW underperforms. We do not force a stronger conclusion from this counterexample. Instead, it suggests that high footprint is not a universal "bad head" label: at some checkpoints, high-footprint heads may provide useful conservative anchoring or ensemble diversity. The online study therefore supports the audit as a way to expose transfer-relevant structure, while showing that naive half-ensemble pruning is too coarse as a final intervention.

4. Discussion and Limitations

Why this is a different direction. Prior work uses critic ensembles for pessimism, uncertainty, exploration, or adaptive aggregation. FOCUS asks a different question: *what structure exists inside the transferred critic ensemble itself?* Our object is not a new aggregation rule, but a checkpoint-level transfer-readiness audit. This perspective is motivated by the sweep: policy success does not fully characterize critic transfer safety, and high-footprint heads are transient rather than tied to a fixed index.

What the current evidence supports. The strongest current claim is diagnostic: critic-footprint heterogeneity can vary sharply across pretraining, and the outlier head identity can change. The online claim is intentionally narrower: the 850K checkpoint shows a clear risky-pool failure mode for FOCUS-HIGH, whereas aggressive FOCUS-LOW selection is mixed across checkpoints. We therefore view the current evidence as support for the audit and for the broader risk-diversity question, not as evidence that a fixed low-footprint half-ensemble is a finished algorithm.

Scope. This is a focused workshop study. Online stress tests are seed-0 and AntMaze medium-diverse only; multi-seed and multi-task validation of the online claim remain future work. The diagnostic phenomenon, however, appears in matched medium- and large-diverse audits, suggesting that conservatism-footprint heterogeneity is not specific to medium-diverse. We also use global half-ensemble pools as stress tests. The mixed behavior of FOCUS-LOW indicates that practical use of critic audits should likely preserve more ensemble diversity, or become state-conditioned or time-windowed as footprint risk changes during online adaptation.

5. Conclusion

FOCUS reframes O2O transfer as a critic transfer-readiness problem. A checkpoint can have strong offline policy performance while hiding a heterogeneous or outlier-prone REDQ critic ensemble. By auditing per-head conservatism footprints at transfer time, FOCUS turns critic selection into a concrete, testable object rather than an implicit assumption. Our current sweep suggests a general phenomenon—critic-head footprints emerge, disappear, and move across checkpoints during conservative pretraining. The reduced-warmup study further shows both sides of the problem: some high-footprint pools can be harmful, but aggressive low-footprint selection is not universally beneficial. The next challenge is to use these audits without discarding useful ensemble diversity.

Impact Statement

Offline-to-online RL is often used when online interaction is costly, so the transferred critic ensemble directly shapes the online Bellman targets an agent learns from. Hidden critic-head miscalibration can waste interaction budget or destabilize early adaptation before fine-tuning recovers. FOCUS adds a lightweight transfer-readiness diagnostic: before treating a checkpoint as ready, inspect the critic ensemble that will generate online targets. This can help identify when a pretrained critic appears reliable, when it should be treated with caution, and when stronger adaptation mechanisms are needed.

References

Ball, P. J., Smith, L., Kostrikov, I., and Levine, S. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, 2023.

Chen, X., Wang, C., Zhou, Z., and Ross, K. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021.

Guo, S., Sun, Y., Hu, J., Huang, S., Chen, H., Piao, H., Sun, L., and Chang, Y. A simple unified uncertainty-guided framework for offline-to-online reinforcement learning. *arXiv preprint arXiv:2306.07541*, 2023.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Lee, S., Seo, Y., Lee, K., Abbeel, P., and Shin, J. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, 2022.

Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., and Levine, S. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems*, 2023.

Werge, N., Wu, Y.-S., Tasdighi, B., and Kandemir, M. Directional ensemble aggregation for actor-critics. *arXiv preprint arXiv:2507.23501*, 2025.

Zhou, Z., Peng, A., Li, Q., Levine, S., and Kumar, A. Efficient online reinforcement learning fine-tuning need not retain offline data. In *International Conference on Learning Representations*, 2025.