Beyond One-Size-Fits-All: Diagnosis-Driven Online Reinforcement Learning with Offline Priors

Guozheng Ma
NTU
guozheng001@e.ntu.edu.sg

Lu Li
Mila & UdeM
lu.li@mila.quebec

Zilin Wang
Oxford
zilin.wang@lmh.ox.ac.uk

Pierre-Luc Bacon
Mila & UdeM
pierre-luc.bacon@mila.quebec

Dacheng Tao
NTU
dacheng.tao@ntu.edu.sg

Abstract

Online reinforcement learning (RL) agents increasingly depend on knowledge acquired offline to achieve practical efficiency. Originally studied in offline-to-online RL, this paradigm now spans foundation model post-training and embodied intelligence, with prior types expanding from offline datasets and pre-trained policies to increasingly diverse knowledge sources such as multimodal foundation models and generative world models. Offline priors have become central to how deep RL is developed and deployed. However, this reliance introduces a challenge that the prevailing benchmark-driven paradigm cannot resolve: because prior validity varies across deployments and shifts during training, no single approach to managing it is universally optimal, and benchmark rankings offer limited guidance for real-world deployments. Rather than pursuing universal solutions, we argue that the field should shift to diagnosis-driven tension management, in which deployment-specific evidence guides how the learner relates to its priors throughout training, enabling both flexible and adaptive deployment. We support this position with a framework characterizing how priors reshape online optimization through three functional roles, controlled experiments demonstrating help-or-hurt reversals, cross-domain evidence from foundation model post-training to embodied intelligence, and engagement with five substantive counterarguments.

1 Introduction

Online reinforcement learning (RL) agents increasingly rely on knowledge acquired offline to achieve practical performance. In foundation model post-training, large-scale pre-trained models provide the base capability that online RL refines for alignment, reasoning, and agentic applications [1, 2]. In embodied intelligence, simulators, demonstration datasets, world models, and pre-trained policies supply the prior knowledge that makes online learning on physical platforms feasible [3, 4, 5, 6]. Despite differences in domain, objective, and prior type, offline priors now provide the foundation from which agents learn through online experience [7, 8, 9, 10].

Despite these benefits, how to use offline priors remains a persistent challenge: the same reliance decision that helps in one setting often hurts in another. In offline-to-online RL, whether to preserve the pre-trained policy, retain the offline dataset, or maintain conservative value estimates each produces different outcomes depending on the quality of the offline sources and their relationship to the deployment task [11, 8, 7, 12]. In foundation model post-training, the role of the reference constraint varies with the fidelity of the reward signal: when rewards are verifiable, strong regularization

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

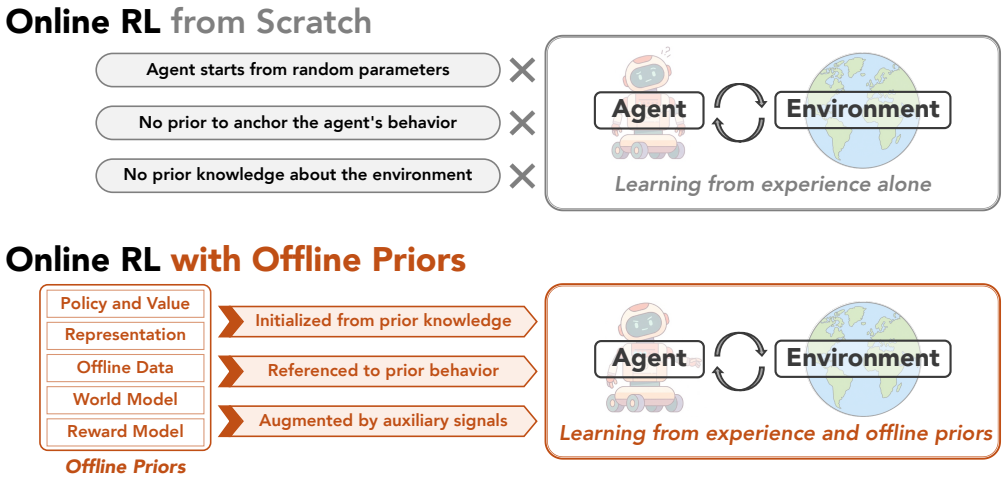

Figure 1: Online RL from scratch (top) versus with offline priors (bottom). The agent-environment interaction loop is identical in both cases. Offline priors add knowledge sources that accelerate learning but also introduce structural tensions analyzed in Sec. 3.

restricts the discovery of novel strategies [13], while when rewards come from learned preference models, the same regularization is essential for preventing overoptimization of proxy scores [1, 14]. In robotics, whether to train in simulation and transfer or to learn directly on physical hardware produces different outcomes depending on the fidelity of the simulator and the complexity of the contact dynamics [15, 6, 16]. These inconsistencies recur across communities, domains, and prior types, suggesting a shared structural property rather than isolated engineering issues.

We argue that these inconsistencies reflect an inherent conflict: offline priors carry knowledge that is inevitably bounded, while online RL exists precisely to push beyond those bounds. • Specifically, whether the prior takes the form of a pre-trained policy, a simulator, or an offline dataset, it encodes knowledge acquired under conditions that may differ from those of deployment. As a result, the extent to which this knowledge remains useful is uncertain and can only be revealed through online interaction. We call this the *bounded commitment* of offline priors: valuable knowledge with an inherently uncertain scope of validity. • On the other hand, online RL drives the agent toward optimal performance through interaction with the deployment environment. This requires the agent to eventually surpass what the prior covers, while relying on its knowledge to get there efficiently. As long as the agent operates within the scope of the prior's validity, reliance is purely beneficial. Once learning pushes beyond that scope, a genuine tension emerges: stronger reliance constrains adaptation, while freer adaptation risks discarding knowledge that still helps elsewhere. Crucially, the agent cannot know where this boundary lies, and the boundary itself shifts as online experience accumulates. Unlike a static trade-off that can be settled at design time, this tension persists and evolves throughout learning. Hence, tension management is a central challenge of this paradigm.

One consequence is that tension management has no universal optimum: the right reliance on each prior depends on the deployment and shifts as learning progresses. Despite this, the field continues to evaluate progress by comparing methods on fixed benchmarks, implicitly assuming that the resulting rankings reflect universal truths rather than deployment-specific matches. The field accumulates condition-specific performance rankings rather than transferable understanding. The way forward requires a shift in research perspective. Rather than asking which method is best, the field should ask what determines when each design choice helps or hurts.

> We argue that the field of online RL with offline priors should move beyond one-size-fits-all methods toward diagnosis-driven tension management. Since no universal optimum exists for how agents should rely on their priors, effective deployment requires diagnostic infrastructure that can assess the prior-deployment match and monitor how tensions evolve during learning, enabling both flexible and adaptive deployment.

We develop this position as follows. Sec. 2 defines the paradigm and introduces a taxonomy of offline prior types. Sec. 3 analyzes why offline priors sharpen the core tensions of online RL and introduces the concept of bounded commitments. Sec. 4 presents evidence that tension management

has no universal optimum and argues for a shift from benchmark-driven to diagnosis-driven tension management. Sec. 5 engages with several potential objections and counterarguments to our core position. Sec. 6 concludes with research opportunities and a broader perspective.

2 The Paradigm of Online RL with Offline Priors

Instead of ***learning from scratch***[1], online RL agents increasingly learn not only from their own online interaction with the environment but also from knowledge acquired offline: setting initial parameters, constraining how far the learner may deviate from prior behavior, or supplying supplementary data or predictions alongside real interaction. Though studied under different names across offline-to-online RL [18, 19], LLM post-training [1, 2], sim-to-real transfer [20, 15], model-based RL [4, 21], and vision-language-action model fine-tuning [22, 23], these mechanisms define a common paradigm that we call ***online RL with offline priors***. As in any RL system, the interaction loop has two sides: an agent that selects actions and an environment that produces states and rewards. Offline knowledge can concern either side, and we organize the resulting priors accordingly.

Table 1: Taxonomy of offline priors in online RL, classified by *what knowledge they encode* and *how they function* during online learning. Agent-side priors encode knowledge about how to act or evaluate; environment-side priors encode knowledge about how the world behaves.

Knowledge Side	**Functional Role**	**Representative Prior Type**		
Agent	Initialization	Policy π_0	Value Q_0 or V_0	Representation ϕ_0
	Reference	Policy π_{ref}	Value Q_{ref} or V_{ref}	† Offline Data $\mathcal{D}$
Environment	Auxiliary	* World Model $\hat{M}$	Reward Model $\hat{R}$	† Offline Data $\mathcal{D}$

† Offline data $\mathcal{D}$ records how the environment transitions and how the collecting policy acts; its functional role depends on which aspect the online algorithm extracts [7, 24].
* World model $\hat{M}$ spans a broad spectrum: compact latent dynamics models [4, 21], engineered simulators used to supplement real-world interaction [20, 15], and world foundation models [25, 26, 27, 28].

Offline priors can shape online optimization through exactly three channels: by determining where optimization begins, by constraining what objective it pursues, or by providing experience beyond direct interaction. • Initialization priors set the starting point for online learning: an offline-trained policy in offline-to-online RL [19], a supervised fine-tuned model in LLM post-training [1], or a simulation-trained controller in sim-to-real transfer [3] each play this role, providing initial competence and reducing the exploration burden that dominates learning from scratch. • Reference priors modify the learning objective by anchoring updates to prior behavior: the KL penalty to a reference policy in RLHF [1], the conservative value penalty in Cal-QL [19], and behavioral cloning regularization [24, 29] are different mechanisms serving the same structural function. In each case, the prior defines a trust region that the online learner is penalized from leaving. • Auxiliary priors provide additional information outside the online loop, whether through retained offline data [7], model-based rollouts [4], or learned reward signals [1]. Unlike reference priors, these do not alter what the optimizer aims to achieve, but expand the evidence it can draw on.

Offline data occupies a unique position in this taxonomy. It may originate from behavior-policy rollouts, expert demonstrations, human preference comparisons, or in-the-wild recordings [30, 5], and it inherently records both how the environment transitions and how the collecting policy acts. Which aspect the online algorithm extracts determines the functional role: replay for value estimation treats data as auxiliary information [7], while behavioral regularization treats the same data as a reference [24, 31]. This dependence on algorithmic use extends beyond data: a pessimistic value function always initializes the learner [19, 8], but additionally serves as an ongoing reference when its conservative penalty is maintained during fine-tuning [32]. In general, functional role is not an inherent property of any prior; it is determined by the algorithm that deploys it.

These priors fundamentally reshape the online learning process. The next section examines why these changes, despite their well-documented benefits, introduce structural tensions into online learning.

[1] Also referred to as *tabula rasa* RL [17].

3 Why Offline Priors Sharpen Tensions

In principle, two fundamental tensions govern the online RL learning process: exploration must be balanced against exploitation, and the plasticity to incorporate new experience must be balanced against the stability of what has already been learned. However, in the challenging tasks that organize mainstream deep RL research the balance tilts sharply to one side. Similarly, sparse rewards, high-dimensional action spaces, and long horizons make useful discoveries so rare that the need for exploration overwhelms any concern about premature exploitation [33, 34]. Bootstrapping from the agent's own shifting value estimates induces optimization pathologies that progressively degrade the network's ability to incorporate new experience [35, 36, 37, 38]. In these settings, failure does not stems from unbalanced tensions, but typically from one-side insufficiency: agents explore too little to discover useful behavior and remain too rigid to learn effectively from new evidence. The research priorities of the past decade reflect this asymmetry: exploration methods overwhelmingly aim to increase coverage [39, 34], and plasticity interventions overwhelmingly aim to restore adaptability [40, 41, 42].

How Priors Sharpen the Tensions. Over the past several years, diverse research communities have effectively addressed these bottlenecks by equipping agents with offline priors before online interaction begins. Pre-trained policies and value functions reduce the exploration burden by providing informed starting behavior rather than random search [18, 19, 1]. Offline data and world models ground the learning process in prior experience, mitigating the cold-start pathologies that degrade network capacity from the earliest updates [7, 4]. These gains are substantial and well documented. However, as priors grow stronger, the previously negligible side of each tension becomes increasingly consequential, and the full two-sided character of both oppositions re-emerges.

This shift is visible along both axes. • On the *stability-plasticity* axis, aggressive online updates risk catastrophic forgetting at the offline-to-online transition [8, 43, 44], while distributional mismatch between offline data and online rollouts can destabilize value estimation [7]. Meanwhile, plasticity failure is no longer only a matter of network capacity, such as dormant neurons or rank collapse [41, 37], but also one of optimization bias, where strong initialization shapes the loss landscape in ways that bias subsequent training toward the prior [11, 45]. • On the *exploration-exploitation* axis, strong priors can narrow the agent's policy toward pre-trained behavioral modes [46, 47], while optimization against learned reward models can drive the agent to exploit proxy scores rather than explore genuinely better behavior [14]. Meanwhile, exploration failure shifts from an inability to reach informative states [33] to a difficulty in moving beyond the signals carried by the prior [46]. In both cases, priors do not simply solve the bottlenecks; they restore and sharpen the full two-sided character of a tension that task difficulty had compressed into a one-sided bottleneck.

Table 2: Structural changes that offline priors introduce to online RL optimization. The reliance parameters μ, λ, β each govern how strongly the learner relies on the prior; setting any to zero recovers the from-scratch case. $\mathcal{M}_{\text{prior}}$ denotes the effective (possibly approximate) environment implied by auxiliary sources ($\mathcal{D}$, $\hat{M}$, $\hat{R}$), and $J(\theta)$ is shorthand for $J(\theta;\, \mathcal{M}_{\text{deploy}})$.

	Online RL from Scratch	Online RL with Offline Priors	Reliance
(a) Initialization	$\theta_0 \sim \text{random}$	$\theta_0 = \mu\, \theta_{\text{prior}} + (1-\mu)\, \theta_{\text{rand}}$	$\mu \in [0, 1]$
(b) Reference	$\max_{\theta}\ J(\theta)$	$\max_{\theta}\ J(\theta) - \lambda\, L_{\text{ref}}(\theta)$	$\lambda \geq 0$
(c) Auxiliary	$\max_{\theta}\ J(\theta;\, \mathcal{M}_{\text{deploy}})$	$\max_{\theta}\ (1-\beta)\, J(\theta;\, \mathcal{M}_{\text{deploy}}) + \beta\, J(\theta;\, \mathcal{M}_{\text{prior}})$	$\beta \in [0, 1]$

Priors as Bounded Commitments. The sharpened tensions described above share a common structural root. Every offline prior encodes knowledge from a source setting that may differ from the deployment environment. The agent cannot fully determine where this knowledge remains valid and where it does not, yet it must *rely on* the prior to learn efficiently and *adapt beyond* it where the prior falls short. We call this epistemic status a ***bounded commitment***: the prior is valuable but its scope of validity is bounded, and the agent must commit to using it without knowing those bounds precisely. The insight that source knowledge has limited validity in new contexts is well established across Bayesian RL [48, 49], transfer learning [50], and adaptive offline RL [51]. The concept of 'bounded commitments' is precisely a name given to this fundamental insight, as it applies to all previous types

of online RL. Table 2 makes this concrete: each functional role from Sec. 2 introduces a reliance parameter (μ, λ, β) that governs how strongly the learner commits to its priors.

Because the validity boundary of each prior is never fully knowable, every reliance configuration is necessarily a bet. Although Table 2 expresses reliance through continuous parameters for analytical clarity, μ, λ, and β abstract over method-level design decisions such as whether to use pre-trained weights, whether to constrain the objective, and whether to supplement the replay buffer. Too much reliance risks trapping the agent in knowledge that does not hold; too little wastes knowledge that could have accelerated learning. This difficulty is structural, not algorithmic, and is compounded by coupling: the KL penalty λ in RLHF, for instance, simultaneously controls policy stability and exploration freedom [46], so that adjusting one tension inevitably affects the other. Because the right reliance depends on how well each prior matches the deployment environment, Sec. 4 examines how this challenge manifests empirically and what it implies for how the field evaluates progress.

Takeaway: Offline priors are fundamentally bounded commitments: valuable knowledge whose scope of validity the agent can never fully determine. This is why offline priors, despite their substantial benefits, sharpen rather than resolve the core tensions of online RL.

4 From Pursuing Universality to Diagnosis-Driven Flexibility and Adaptivity

Despite the structural tensions identified in Sec. 3, the current dominant research paradigm continues to evaluate progress by comparing methods on fixed benchmarks and seeking algorithms that perform well across the board. This approach implicitly assumes that a single reliance configuration can be universally optimal. This section presents evidence against that assumption (Sec. 4.1) and argues that reliable deployment requires diagnostic infrastructure rather than universal methods (Sec. 4.2).

4.1 Tension Management Has No Universal Optimum

How well each prior matches the deployment environment determines the optimal reliance configuration: how closely the offline data covers the online distribution, how accurately a world model reflects real dynamics, and how faithfully a pre-trained policy captures the behavior the task requires. Since this match varies across tasks and conditions, the optimum varies with it.

Illustrative Experiments. We demonstrate this non-universality through controlled experiments in the offline-to-online RL setting. We choose this setting because it abstracts away domain-specific engineering details present in LLM post-training, sim-to-real transfer, and VLA fine-tuning, allowing each reliance parameter from Table 2 to be toggled independently. • For initialization (μ), we compare starting from the offline-trained policy weights *versus* resetting to a randomly initialized network. • For reference (λ), we compare maintaining conservative value penalties throughout fine-tuning *versus* dropping all conservatism and using unconstrained optimization [19, 52]. • For auxiliary information (β), we compare retaining the offline dataset in the replay buffer *versus* discarding it and learning from online data only [7, 8]. Fig. 2 presents representative task pairs from our experiments (full results across all tasks in Appendix A). In each column, the same binary choice produces opposite outcomes across tasks. These reversals are not random variation: they arise because the optimal reliance depends on a complex interaction between the properties of the prior, the structure of the deployment task, and the degree to which they match. Since all three factors vary across settings, no single configuration is reliably beneficial. Recent work has made this pattern precise in offline-to-online RL by identifying distinct regimes in which the optimal strategy qualitatively flips [11]. Our experiments extend this observation by showing that non-universality spans all three reliance dimensions independently, not only the interaction between initialization and data retention.

A Cross-Domain Pattern. The same non-universality appears far beyond offline-to-online RL. • In LLM post-training, the debate over reference regularization illustrates the point directly: removing the KL penalty improves reasoning performance on verifiable tasks [13], but the penalty remains essential for preventing reward hacking when rewards come from learned models [1, 14]. Recent work shows that gradient regularization can outperform KL penalties entirely in some regimes while failing in others [53], and that static length penalties help efficiency on easy tasks but hurt accuracy on hard ones [54]. • In sim-to-real robotics, a large-scale study across three robot platforms finds that widely used algorithmic defaults can be harmful on physical hardware [15], and that end-to-end policy fine-tuning collapses in real-world deployment even when it succeeds in simulation [16].

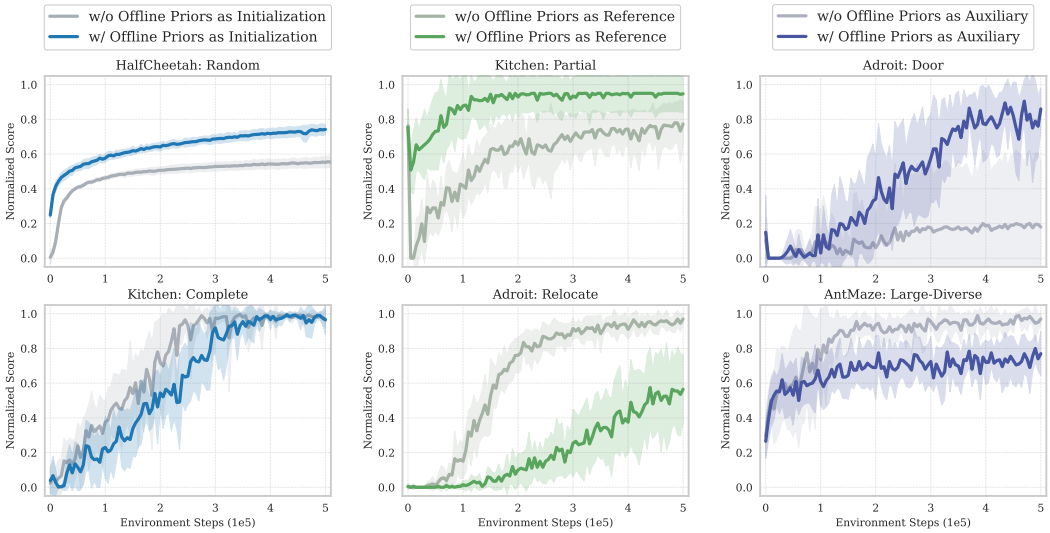

Figure 2: No single reliance configuration is universally optimal. Each column isolates one reliance parameter from Table 2: initialization (μ, left), reference (λ, middle), and auxiliary (β, right). The top row shows a task where stronger reliance on the prior helps; the bottom row shows a task where the same choice hurts. **Left**: RLPD with versus without offline-trained initialization. **Middle**: Cal-QL's conservative penalty versus unconstrained updates, with both initialization and offline data present. **Right**: retaining versus discarding offline data, with both initialization and conservative penalty present. Full results and additional controlled comparisons are provided in Appendix A.

• In vision-language-action model fine-tuning, sequential adaptation with LoRA works remarkably well for large pretrained VLAs but collapses when any single ingredient is removed [55]. • In model-based RL, explicit conservatism helps on high-coverage datasets but fails on low-quality data, where Bayesian approaches without conservatism perform better [56]. Across all these settings, the underlying pattern is the same: the optimal reliance level depends not on the method alone but on the properties of the prior, the demands of the deployment task, and how well the two align. That this appears independently across communities with different methods, vocabularies, and evaluation practices suggests it is structural rather than incidental.

The Inherent Limits of Benchmark-Driven Evaluation. The evidence above implies that each method implicitly encodes a particular reliance configuration. A benchmark ranking therefore reflects how well that configuration matches the evaluation conditions, not a universal ordering of methods. When the deployment setting changes, the ranking can change with it. The standard response to this fragility has been to broaden evaluation by testing on more tasks, more environments, and more data conditions. However, expanding the evaluation suite tends to multiply contradictory findings rather than eliminate them, and aggregate metrics compress these disagreements into a single ranking that explains none of them. As the empirical record grows, what accumulates is not transferable understanding but an expanding catalogue of condition-specific performance rankings.

Recent position papers have raised compatible concerns: that rigorous RL benchmarking is prohibitively expensive [57], that aggregate scores obscure fragile generalization [58], and that standard protocols hide the true cost of hyperparameter selection [59] and mask deployment non-stationarity [60]. The non-universality we identify raises a more fundamental concern: even methodologically sound benchmarking cannot produce transferable conclusions when the optimal configuration is itself deployment-dependent.

The fundamental limitation is not the size or quality of any benchmark but the kind of question that benchmark comparisons can answer. Rankings order methods within a fixed setting; they do not reveal which properties of the prior and the deployment task govern whether a design choice helps or hurts. Progress requires a different kind of question: not which method is best, but what each deployment needs. Without infrastructure to answer that question, the field risks an indefinite cycle of benchmark expansion without convergent insight.

Takeaway: No single reliance configuration is universally optimal: the same design choice that helps in one setting can hurt in another. Benchmark comparisons cannot resolve this because they answer which method wins, not what determines when each choice helps or hurts.

4.2 From Benchmark-Driven to Diagnosis-Driven Tension Management

More or better benchmarks cannot overcome a limitation inherent in ranking-based evaluation itself. The field needs to change not the evaluation tools but the question those tools are designed to answer. Concretely, we advocate a shift from *benchmark-driven* to *diagnosis-driven* tension management: from letting fixed rankings guide method selection to letting deployment-specific evidence guide how the learner relates to its priors throughout training.

The Core Distinction. In the benchmark-driven paradigm, reliance on each prior is configured before deployment based on aggregate evaluation results, and remains fixed or follows a pre-determined schedule once training begins. Online experience serves only to update the policy, even though every transition also carries evidence about the prior itself, including whether value estimates still align with the deployment environment, whether offline data still overlaps with online rollouts, and whether reference behavior still serves the task. This evidence goes unused. The diagnosis-driven paradigm treats it as a first-class signal. The same transitions that update the policy also reveal whether value estimates remain calibrated, whether offline data still provides useful grounding, and whether reference behavior continues to guide learning productively. This evidence informs reliance decisions throughout training, determining when to trust the prior and when to move beyond it. In short, diagnosis-driven tension management is the practice of systematically extracting evidence about prior validity from online interaction and using it to guide reliance decisions.

This is not merely a methodological preference. However carefully a prior is constructed, its validity boundary in any specific deployment remains uncertain until interaction begins. Offline evaluation can estimate prior quality in general, but cannot determine which specific aspects will hold or fail under new conditions. Designing better priors or more robust algorithms can reduce the frequency of severe mismatches, but cannot eliminate the underlying uncertainty: the agent is always using knowledge acquired elsewhere to act in an environment it has not yet fully observed. Online evidence is therefore indispensable, because no other evidence about deployment-specific prior validity exists.

The Diagnostic Dimensions. The information available for making reliance decisions changes fundamentally over the course of training. Before online interaction begins, the practitioner knows only the prior and the task. Once online learning begins, each interaction reveals where the prior holds and where it does not. These two stages require two complementary forms of diagnosis.

The first form is ***prior-deployment match assessment***. Before online training begins, the practitioner should estimate how well each prior fits the task at hand, based on properties that are observable without interaction, such as data coverage, policy quality, or model accuracy [11, 61]. Recent work demonstrates that even coarse estimates carry actionable information: comparing offline policy quality with data quality can already determine which component the practitioner should anchor on [11], and lightweight metrics can predict whether a dataset will support effective fine-tuning [61]. By informing the initial configuration, match assessment enables ***flexibility*** across deployments.

The second form is ***tension dynamics monitoring***. Training is not static: aspects that were initially valuable may become outdated as the agent's own experience grows, turning helpful guidance into a binding constraint. The field already recognizes this implicitly. Methods that anneal conservative penalties [62], schedule warmup phases [8], or periodically reset the reference model [63] all assume that reliance should change during training, but make these adjustments on a fixed schedule rather than in response to observed learning dynamics. Emerging work shows that measurement-driven adjustment is feasible: plasticity metrics can detect capacity loss during training [41, 38, 64], adaptive replay buffers can rebalance data sources based on relevance signals [65], and reward-model monitors can flag proxy divergence [66]. However, these tools remain fragmented across domains, each targeting a specific failure mode rather than assessing prior validity as a whole. By tracking how prior validity evolves during training, dynamics monitoring enables ***adaptivity*** within each deployment.

In practice, these two dimensions interact. Online evidence gathered through dynamics monitoring can retrospectively validate or revise the initial match assessment, and a better initial assessment reduces the burden on runtime monitoring. This loop is almost entirely absent today. Most methods

fix their reliance configuration at the start of training or adjust it on a predetermined schedule, without using online evidence to revise the initial assessment. Closing this loop through principled diagnostic infrastructure is, in our view, necessary for making online RL with offline priors reliably deployable.

Takeaway: The field should shift from benchmark-driven to diagnosis-driven tension management. Online experience carries evidence not only about the task but also about the ongoing validity of each prior. Diagnosis-driven tension management is the practice of extracting this evidence and using it to guide reliance decisions, enabling both flexible and adaptive deployment.

5 Objections and Counterarguments

We consider five potential objections to our call for diagnosis-driven online RL with offline priors.

This is an engineering challenge, not a scientific insight. It is fair to ask whether this paper merely names a phenomenon that practitioners already navigate daily: different deployments need different configurations, and finding good ones is routine engineering. Each community already handles its own regime dependence, but in isolation: the offline-to-online RL community studies initialization versus data retention [11], the RLHF community debates KL penalty strength [13, 53], and the sim-to-real community weighs fine-tuning against freezing [15]. Our contribution is recognizing that these are superficially different expressions of the same structural phenomenon and that the same diagnostic principles apply across all of them. Cross-community unification of this kind has consistently been treated as scientific contribution in ML, from the formalization of transfer learning [67] to the systematization of catastrophic forgetting [68]. A reader may also interpret our formalization as a hyperparameter tuning problem, but the decisions that μ, λ, and β abstract over are not points on a continuous search grid. They abstract over architectural and algorithmic choices such as using pretrained versus random initialization, imposing versus dropping a conservative penalty, and retaining versus discarding offline data. Furthermore, prior validity shifts as online experience accumulates, so the right reliance level at the start of training may not remain right later.

Better offline priors solve the problem at the source. Rather than diagnosing bounded commitment during deployment, one could try to prevent it upstream. Recent work pursues this through larger and more diverse datasets [5, 69], adaptive offline objectives that preserve revision capacity [51, 48], world foundation models that encode broad physical priors [26, 27], and Bayesian formulations that build uncertainty directly into the learned policy [49, 56]. These efforts are valuable and often dramatically improve transfer and generalization. However, in every major deployment paradigm, online adaptation remains a necessary stage: foundation-model priors are designed as efficient initializations to be fine-tuned, not as finished policies for arbitrary deployment [70], and adaptive offline objectives explicitly aim to preserve the capacity for later revision, not to remove the need for it [51, 48]. The reason is structural: online RL exists in the pipeline precisely because the prior does not fully solve the deployment task, and improving beyond the prior necessarily means entering territory where its guidance is no longer reliable. Better priors extend the region where reliance is safe; diagnosis addresses what happens at and beyond its boundary. The two are complementary, and advances in either make the other more effective.

Scaling and algorithmic progress will resolve this. A longer-term version of the previous argument holds that continued progress in model scale, data scale, and algorithm design will eventually make bounded commitment negligible, rendering diagnostic infrastructure a premature investment. There is real evidence for this view: at sufficient scale, some classical pathologies weaken. Large pretrained VLAs show little forgetting during continual adaptation [55, 71], and larger language models exhibit more efficient RL post-training [72]. However, the strongest scaling results are themselves regime-dependent. The VLA recipe that eliminates forgetting requires a specific combination of large model, parameter-efficient tuning, and on-policy RL; removing any ingredient causes collapse [55]. Fine-tuning capacity does not transfer uniformly across tasks and embodiments [73, 74], and platform-dependent defaults persist even with state-of-the-art algorithms [15]. What scale changes is not whether deployment-specific choices matter but which ones matter most. Meanwhile, scaling expands the range of deployments the field attempts to address, introducing new embodiments, task types, and deployment conditions faster than any single advance can uniformly cover. This makes principled diagnosis more necessary as the field scales, not less.

The field only needs better benchmarks, not diagnostic infrastructure. Recent position papers have proposed valuable reforms to RL evaluation: accounting for tuning costs [59], restricting lifetime

access [60], and testing for fragile generalization [58, 57]. This concern is well founded, and we agree that evaluation methodology needs reform. However, even perfect benchmarks answer a different question than diagnostics do. Benchmarks tell us which method tends to work under which conditions; diagnostics tell us whether a specific deployment meets those conditions, and whether the answer is changing as training proceeds. The second question requires online evidence that only deployment interaction can generate, which is why benchmarks and diagnostics are complementary: one narrows the space of candidate methods, the other guides their configuration during training.

Deep RL is too opaque for reliable diagnosis. Measuring prior validity is genuinely harder than measuring network capacity, and no general-purpose diagnostic toolkit exists today. However, the relevant question is not why the network behaves as it does but whether the prior is still helping. The former requires interpretability, which remains hard. The latter requires only observable quantities: performance trends, distribution overlap, and prediction accuracy [11, 61]. Practical building blocks already exist across domains: discrete regime distinctions guide method selection in offline-to-online RL [11], reward-quality monitors flag proxy divergence in RLHF [66], and uncertainty estimates weight synthetic data by reliability in model-based RL [75]. Each measures observable signals and translates them into actionable decisions. The plasticity literature shows this trajectory is viable: from informal recognition to systematic measurement to a productive diagnostic subfield with reusable tools, all within a few years [41, 36]. Prior validity diagnosis is at an earlier stage of the same progression. Furthermore, diagnosis need not be perfect to be productive. Even coarse measurements improve on decisions that would otherwise be made without evidence, and each deployment that uses diagnostic signals generates insights that sharpen future tools. This self-reinforcing cycle between diagnostics and deployment is a promising path toward reliable online RL with offline priors.

6 Conclusion

This paper has argued that offline priors fundamentally reshape the structure of online RL. By introducing knowledge whose scope of validity the agent cannot fully determine, priors transform the one-sided bottlenecks of from-scratch learning into genuine two-sided tensions. We have formalized this through the concept of bounded commitment, shown empirically that no single reliance configuration is universally optimal, and argued that the field should shift from benchmark-driven to diagnosis-driven tension management, in which online experience is used not only to learn the task but also to assess prior validity and guide reliance decisions toward flexible and adaptive deployment.

Research Opportunities. If the field adopts diagnosis-driven tension management, several opportunities open up that the current paradigm does not naturally support.

- *A new class of research contributions.* Under the benchmark-driven paradigm, contributions are measured primarily by performance gains. Diagnosis-driven research values a different kind of output: not a method that wins on a benchmark, but a signal that predicts when a design choice helps or hurts, a metric that assesses prior-deployment match, or a monitor that tracks prior validity during training [11, 61]. Work of this kind already exists but is typically framed as supporting analysis rather than a primary contribution [76, 77, 78].
- *Cross-community knowledge transfer.* Currently, each community rediscovers similar failure modes in isolation: catastrophic forgetting in offline-to-online RL, reward hacking in RLHF, reality-gap collapse in sim-to-real transfer. If these are recognized as manifestations of the same structural phenomenon, diagnostic tools developed in one community can inform practice in others. Plasticity metrics illustrate this potential: dormant neuron ratios originated in a specific experimental setting [41] but now serve as reusable diagnostics across tasks and algorithms.
- *Deployment as a source of scientific knowledge.* Under the current paradigm, deployment is the endpoint of research: methods are developed, evaluated, and then applied. Diagnosis-driven deployment inverts this relationship. Every deployment that uses diagnostic evidence generates insights about the conditions under which each prior holds or fails, feeding back into the design of better tools and more informed future deployments.
- *Methods aware of their own bounded commitments.* The diagnosis-driven paradigm also changes how methods themselves are designed. Rather than seeking algorithms that perform well across the board, researchers can design methods that are explicitly aware of the boundaries of their prior knowledge and capable of adjusting their own reliance as those boundaries are revealed during training. This represents a shift from optimizing for average-case performance to building in the capacity for deployment-specific adaptation.

References

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[3] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[4] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[5] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[6] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16961–16969. IEEE, 2024.

[7] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

[8] Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 1:11, 2025.

[10] Alex Lewandowski, Aditya A. Ramesh, Edan Meyer, Dale Schuurmans, and Marlos C. Machado. The world is bigger! a computationally-embedded perspective on the big world hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[11] Lu Li, Tianwei Ni, Yihao Sun, and Pierre-Luc Bacon. The three regimes of offline-to-online reinforcement learning. *arXiv preprint arXiv:2510.01460*, 2025.

[12] Shenzhi Wang, Qisen Yang, Jiawei Gao, Matthieu Lin, Hao Chen, Liwei Wu, Ning Jia, Shiji Song, and Gao Huang. Train once, get a family: State-adaptive balances for offline-to-online reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47081–47104, 2023.

[13] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[14] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[15] Yarden As, Dhruva Tirumala, René Zurbrügg, Chenhao Li, Stelian Coros, Andreas Krause, and Markus Wulfmeier. What matters for simulation to online reinforcement learning on real robots. *arXiv preprint arXiv:2602.20220*, 2026.

[16] Jacob Levy, Tyler Westenbroek, Kevin Huang, Fernando Palafox, Patrick Yin, Shayegan Omidshafiei, Dong-Ki Kim, Abhishek Gupta, and David Fridovich-Keil. Simulation distillation: Pretraining world models in simulation for rapid real-world adaptation. *arXiv preprint arXiv:2603.15759*, 2026.

[17] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G Bellemare. Reincarnating reinforcement learning: Reusing prior computation to accelerate progress. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[18] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[19] Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36:62244–62269, 2023.

[20] Andrew Wagenmaker, Kevin Huang, Liyiming Ke, Kevin Jamieson, and Abhishek Gupta. Overcoming the sim-to-real gap: Leveraging simulation to learn to explore for real-world RL. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[23] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15665–15672. IEEE, 2025.

[24] Scott Fujimoto and Shixiang (Shane) Gu. A minimalist approach to offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145. Curran Associates, Inc., 2021.

[25] Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026.

[26] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[27] GigaBrain Team, Boyuan Wang, Bohan Li, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigabrain-0.5 m*: a vla that learns from world model-based reinforcement learning. *arXiv preprint arXiv:2602.12099*, 2026.

[28] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026.

[29] Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE, 2023.

[30] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *6th Annual Conference on Robot Learning*, 2022.

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

[32] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

[33] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[34] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

[35] Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pages 16828–16847. PMLR, 2022.

[36] Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, pages 23190–23211. PMLR, 2023.

[37] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

[38] Guozheng Ma, Lu Li, Sen Zhang, Zixuan Liu, Zhen Wang, Yixin Chen, Li Shen, Xueqian Wang, and Dacheng Tao. Revisiting plasticity in visual reinforcement learning: Data, modules and training stages. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE transactions on neural networks and learning systems*, 35(7):8762–8782, 2023.

[40] Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and Andre Barreto. Deep reinforcement learning with plasticity injection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pages 32145–32168. PMLR, 2023.

[42] Timo Klein, Christoph Luther, Manus McAuliffe, Lukas Miklautz, Claudia Plant, and Sebastian Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832*, 2024.

[43] Yicheng Luo, Jackie Kay, Edward Grefenstette, and Marc Peter Deisenroth. Finetuning from offline reinforcement learning: Challenges, trade-offs and practical solutions. *arXiv preprint arXiv:2303.17396*, 2023.

[44] Maciej Wolczyk, Bartłomiej Cupiał, Mateusz Ostaszewski, Michał Bortkiewicz, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Fine-tuning reinforcement learning models is secretly a forgetting mitigation problem. In *Forty-first International Conference on Machine Learning*, 2024.

[45] Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado Van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. *arXiv preprint arXiv:2402.18762*, 2024.

[46] Rosie Zhao, Alexandru Meterez, Sham M. Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. In *Second Conference on Language Modeling*, 2025.

[47] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[48] Dibya Ghosh, Anurag Ajay, Pulkit Agrawal, and Sergey Levine. Offline rl policies should be trained to be adaptive. In *International Conference on Machine Learning*, pages 7513–7530. PMLR, 2022.

[49] Hao Hu, Yiqin Yang, Jianing Ye, Chengjie Wu, Ziqing Mai, Yujing Hu, Tangjie Lv, Changjie Fan, Qianchuan Zhao, and Chongjie Zhang. Bayesian design principles for offline-to-online reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024.

[50] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

[51] Tianwei Ni, Vineet Jain, Akash Karthikeyan, and Pierre-Luc Bacon. From static policies to adaptive priors in offline reinforcement learning. *Preprint*, 2026.

[52] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[53] Johannes Ackermann, Michael Noukhovitch, Takashi Ishida, and Masashi Sugiyama. Gradient regularization prevents reward hacking in reinforcement learning from human feedback and verifiable rewards. *arXiv preprint arXiv:2602.18037*, 2026.

[54] Keqin Peng, Yuanxin Ouyang, Xuebo Liu, Zhiliang Tian, Ruijian Han, Yancheng Yuan, and Liang Ding. Think dense, not long: Dynamic decoupled conditional advantage for efficient reasoning. *arXiv preprint arXiv:2602.02099*, 2026.

[55] Jiaheng Hu, Jay Shim, Chen Tang, Yoonchang Sung, Bo Liu, Peter Stone, and Roberto Martin-Martin. Simple recipe works: Vision-language-action models are natural continual learners with reinforcement learning. *arXiv preprint arXiv:2603.11653*, 2026.

[56] Tianwei Ni, Esther Derman, Vineet Jain, Vincent Taboga, Siamak Ravanbakhsh, and Pierre-Luc Bacon. Long-horizon model-based offline reinforcement learning without conservatism. *arXiv preprint arXiv:2512.04341*, 2025.

[57] Emma Jordan, Adam White, Bruno Castro da Silva, Martha White, and Philip S. Thomas. Position: Benchmarking is limited in reinforcement learning research. In *Forty-first International Conference on Machine Learning*, 2024.

[58] Zihan Chen, Yiming Zhang, Hengguang Zhou, Zenghui Ding, Yining Sun, and Cho-Jui Hsieh. Rethinking rl evaluation: Can benchmarks truly reveal failures of rl methods? *arXiv preprint arXiv:2510.10541*, 2025.

[59] Ziqi Tang and Xuezhou Zhang. Position: Ignoring hyperparameter tuning costs misleads the development of efficient rl algorithms. *preprint*, 2025.

[60] Golnaz Mesbahi, Parham Mohammad Panahi, Olya Mastikhina, Steven Tang, Martha White, and Adam White. Position: Lifetime tuning is incompatible with continual reinforcement learning. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[61] Arip Asadulaev, Fakhri Karray, and Martin Takac. Expert or not? assessing data quality in offline reinforcement learning. *arXiv preprint arXiv:2510.12638*, 2025.

[62] Geonwoo Cho, Jaegyun Im, Doyoon Kim, and Lexin Li. Annealing bridges offline and online RL, 2025.

[63] Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[64] Guowei Xu, Ruijie Zheng, Yongyuan Liang, Xiyao Wang, Zhecheng Yuan, Tianying Ji, Yu Luo, Xiaoyu Liu, Jiaxin Yuan, Pu Hua, Shuzhen Li, Yanjie Ze, Hal Daumé III, Furong Huang, and Huazhe Xu. Drm: Mastering visual reinforcement learning through dormant ratio minimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[65] Chihyeon Song, Jaewoo Lee, and Jinkyoo Park. Adaptive replay buffer for offline-to-online reinforcement learning. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[66] Yuchun Miao, Liang Ding, Sen Zhang, Rong Bao, Lefei Zhang, and Dacheng Tao. Information-theoretic reward modeling for stable rlhf: Detecting and mitigating reward hacking. *arXiv preprint arXiv:2510.13694*, 2025.

[67] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[68] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[69] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[70] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

[71] Huihan Liu, Changyeon Kim, Bo Liu, Minghuan Liu, and Yuke Zhu. Pretrained vision-language-action models are surprisingly resistant to forgetting in continual learning. *arXiv preprint arXiv:2603.03818*, 2026.

[72] Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, et al. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*, 2025.

[73] Donghoon Kim, Minji Bae, Unghui Nam, Gyeonghun Kim, Suyun Lee, Kyuhong Shim, and Byonghyo Shim. Adaptive capacity allocation for vision language action fine-tuning. *arXiv preprint arXiv:2603.07404*, 2026.

[74] Xinghang Li, Peiyan Li, Long Qian, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Xinlong Wang, Di Guo, et al. What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, pages 1–15, 2026.

[75] Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. WIMLE: Uncertainty-aware world models with IMLE for sample-efficient continuous control. In *The Fourteenth International Conference on Learning Representations*, 2026.

[76] Guozheng Ma, Lu Li, Haoyu Wang, Zixuan Liu, Pierre-Luc Bacon, and Dacheng Tao. What makes value learning efficient in residual reinforcement learning? *arXiv preprint arXiv:2602.10539*, 2026.

[77] Johan Obando-Ceron, Walter Mayor, Samuel Lavoie, Scott Fujimoto, Aaron Courville, and Pablo Samuel Castro. Simplicial embeddings improve sample efficiency in actor–critic agents. In *The Fourteenth International Conference on Learning Representations*, 2026.

[78] Isaac Han, Sangyeon Park, Seungwon Oh, Donghu Kim, Hojoon Lee, and KyungJoong Kim. FIRE: Frobenius-isometry reinitialization for balancing the stability–plasticity tradeoff. In *The Fourteenth International Conference on Learning Representations*, 2026.

[79] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

A Illustrative Experiments: Full Results

Section 4.1 presents representative task pairs to illustrate the non-universality of tension management. This appendix provides the complete experimental setup and full results across all tasks and reliance dimensions.

Setup. All experiments are conducted in the offline-to-online RL setting on D4RL benchmarks [79], covering Adroit manipulation (Pen, Relocate, Door), Kitchen (Complete, Partial, Mixed), AntMaze navigation (Large-Diverse, Large-Play, Ultra-Diverse), and MuJoCo locomotion (HalfCheetah, Hopper, Walker2D, each with Random, Medium-Replay, Medium, and Medium-Expert datasets). Each experiment isolates one reliance parameter from Table 2 by toggling it while holding the others fixed. We organize the experiments into three groups corresponding to the three functional roles, with two groups further split to reveal interactions between parameters. Table 3 summarizes the experimental conditions. In all cases, the "with" and "without" conditions differ in exactly one reliance dimension, enabling controlled comparison.

Table 3: Summary of experimental conditions. Each row isolates one reliance parameter while holding the others fixed. Checkmarks indicate active components.

Experiment	**Comparison**	**Init** (μ)	**Ref** (λ)	**Aux** (β)
Initialization	w/ init vs w/o init	**varies**	$\times$	$\checkmark$
Reference (no aux)	w/ ref vs w/o ref	$\checkmark$	**varies**	$\times$
Reference (with aux)	w/ ref vs w/o ref	$\checkmark$	**varies**	$\checkmark$
Auxiliary (no ref)	w/ aux vs w/o aux	$\checkmark$	$\times$	**varies**
Auxiliary (with ref)	w/ aux vs w/o aux	$\checkmark$	$\checkmark$	**varies**

Initialization (μ). This experiment tests whether initializing the online learner from the offline-trained policy and value function improves over random initialization. Both conditions retain the offline dataset in the replay buffer following the RLPD protocol [7] and use standard SAC updates without conservative penalties. The comparison thus isolates the effect of initialization reliance: starting from prior parameters ($\mu = 1$) versus starting from random parameters ($\mu = 0$), with auxiliary support held constant.

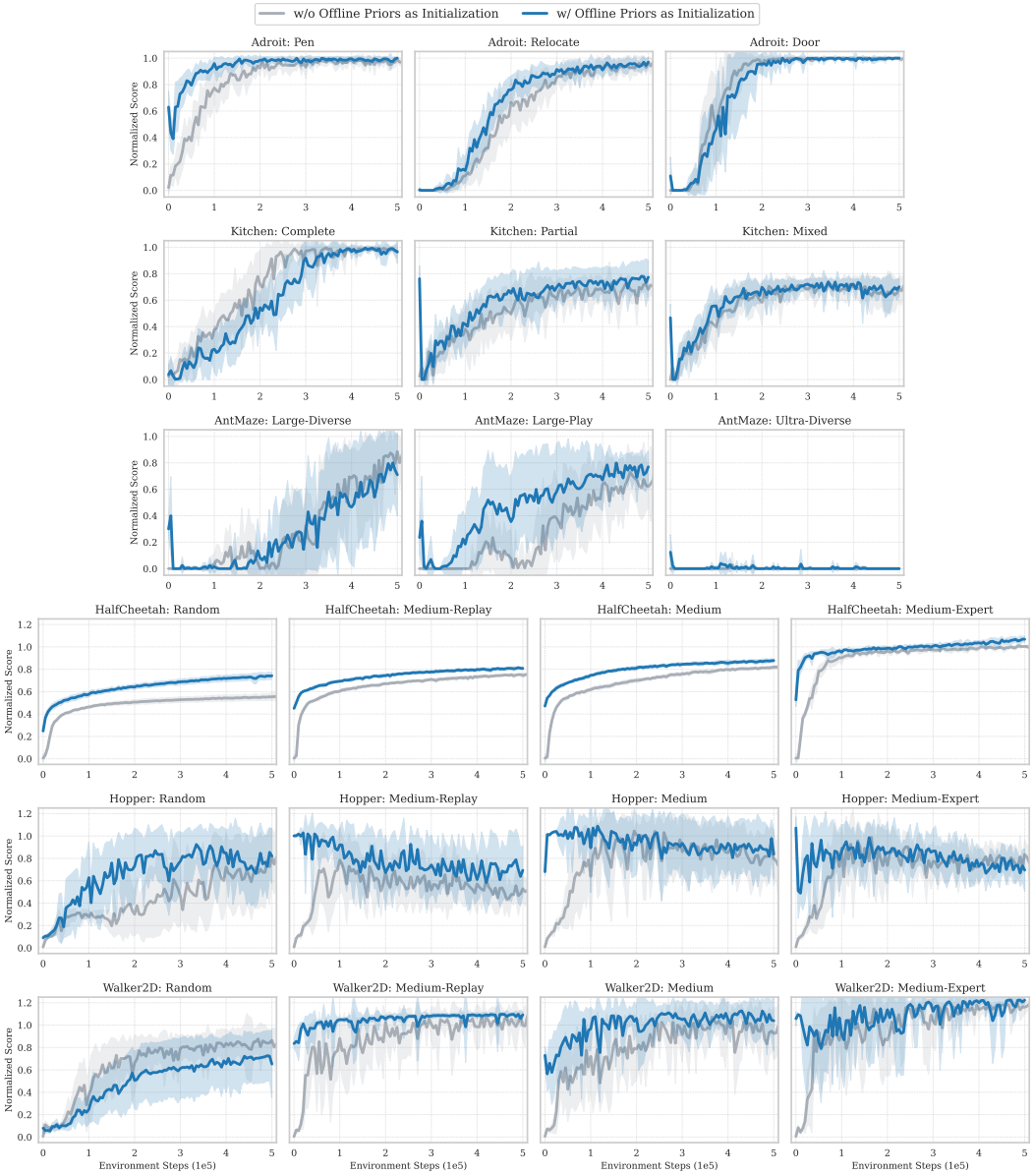

Figure 3: Effect of initialization reliance (μ), with auxiliary data retained in both conditions. Using offline-trained initialization helps on some tasks but hurts on others, illustrating that the optimal μ is task-dependent.

Reference (λ), without auxiliary data. This experiment tests whether maintaining a conservative value penalty during online learning (Cal-QL [19]) improves over unconstrained SAC [52]. Both conditions start from the offline-trained initialization and do not retain offline data in the replay buffer. The comparison isolates the effect of reference reliance: conservative penalty active ($\lambda > 0$) versus no penalty ($\lambda = 0$), without auxiliary support.

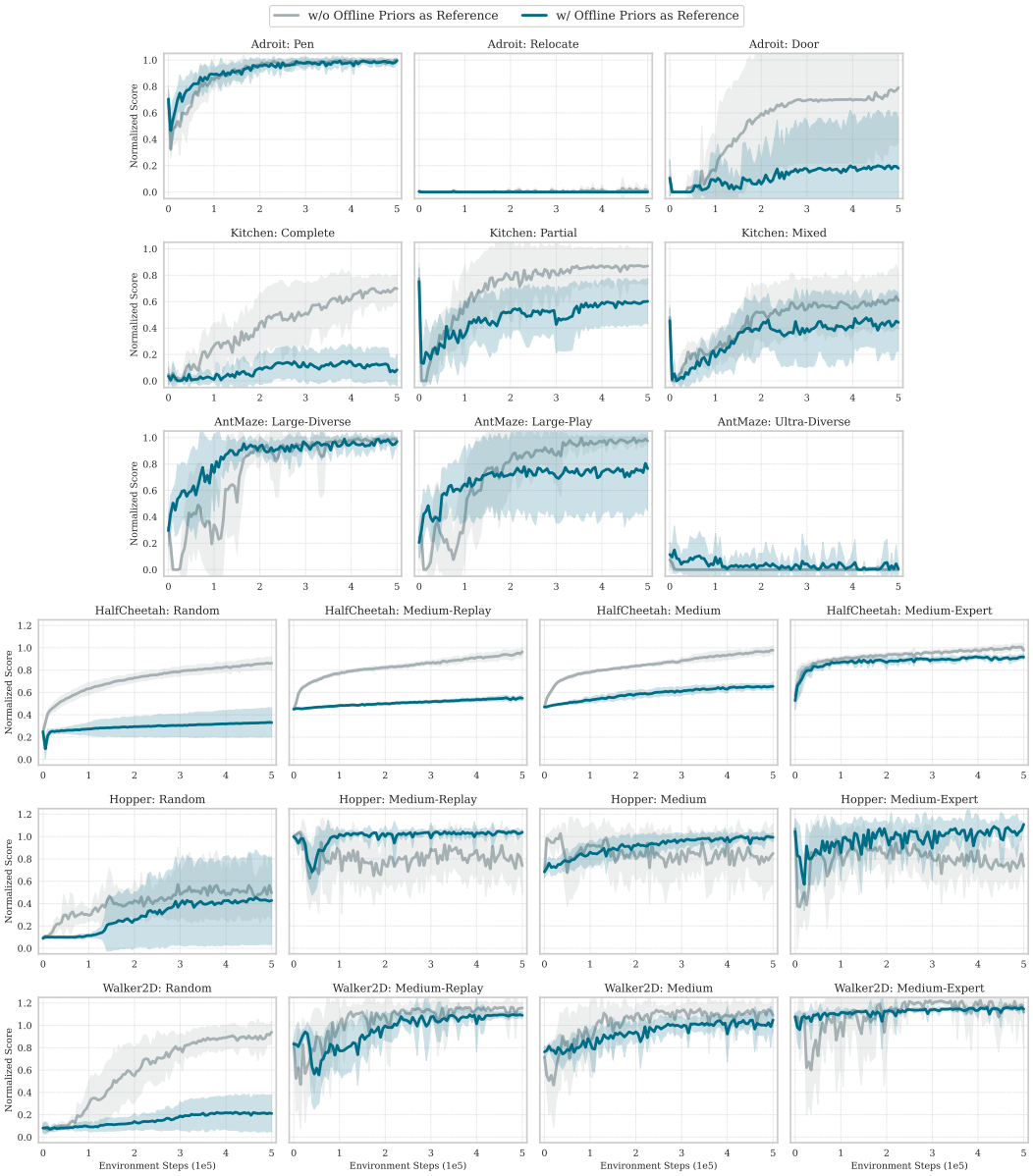

Figure 4: Effect of reference reliance (λ) without auxiliary data. Both conditions use offline-trained initialization. Maintaining conservative value penalties helps on some tasks but hurts on others.

Reference (λ), with auxiliary data. This experiment repeats the reference comparison with auxiliary data present: the offline dataset is retained in the replay buffer in both conditions. The comparison isolates whether using Cal-QL's conservative penalty (as opposed to standard SAC updates) remains beneficial when offline data is also available as auxiliary support. Comparing Figures 4 and 5 reveals how the effect of a reference constraint can itself depend on whether auxiliary data is available, illustrating the coupling between reliance parameters discussed in Section 3.

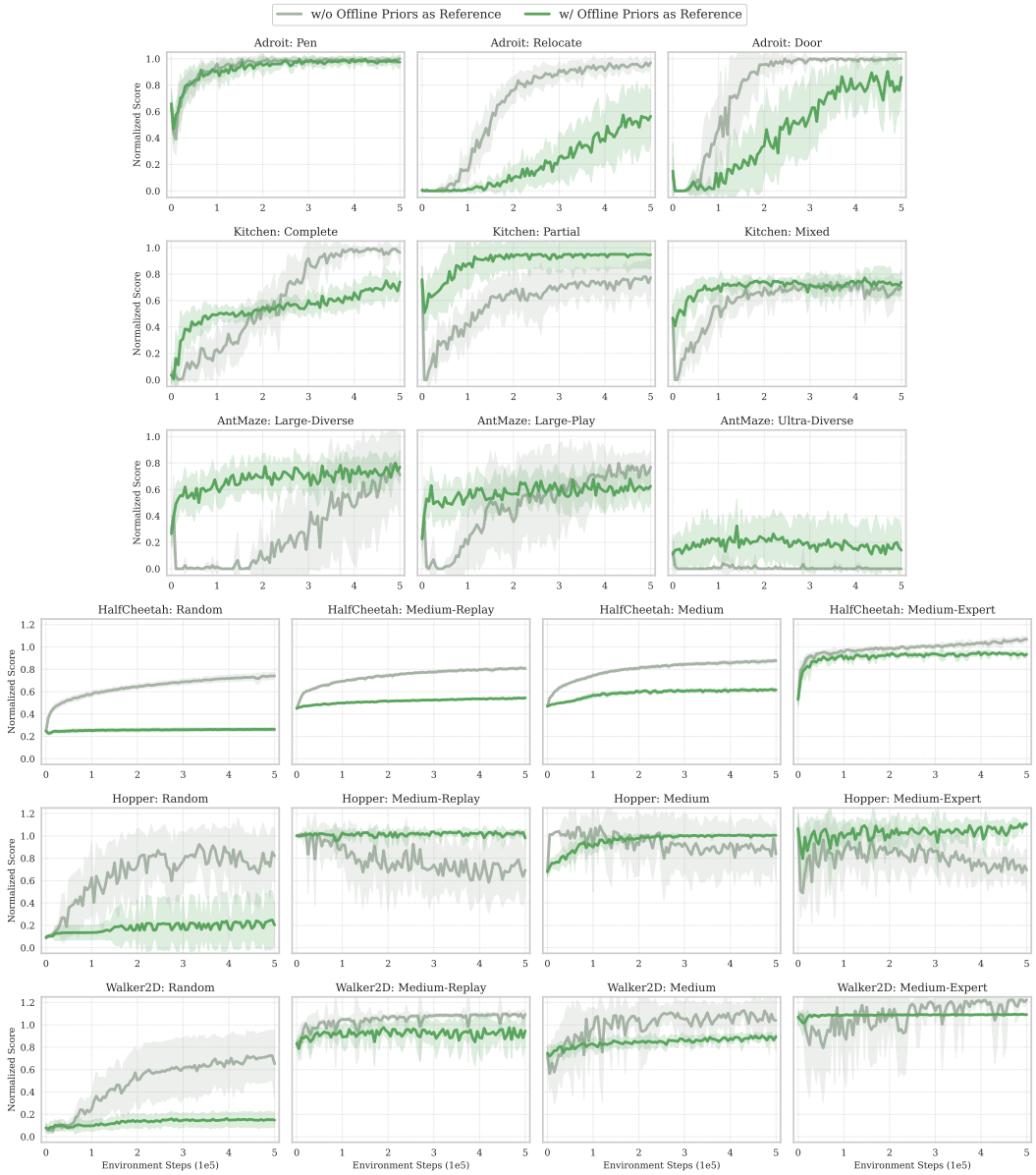

Figure 5: Effect of reference reliance (λ) with auxiliary data ($\beta > 0$). Both conditions use offline-trained initialization and retain offline data. The same conservative penalty that helps or hurts in Figure 4 can behave differently when offline data is also retained, demonstrating interaction between reliance parameters.

Auxiliary (β), without reference constraint. This experiment tests whether retaining the offline dataset in the replay buffer improves over discarding it, when the online learner uses standard SAC without conservative penalties. Both conditions start from the offline-trained initialization. The comparison isolates the effect of auxiliary reliance: offline data retained ($\beta > 0$) versus online data only ($\beta = 0$), without reference constraint.

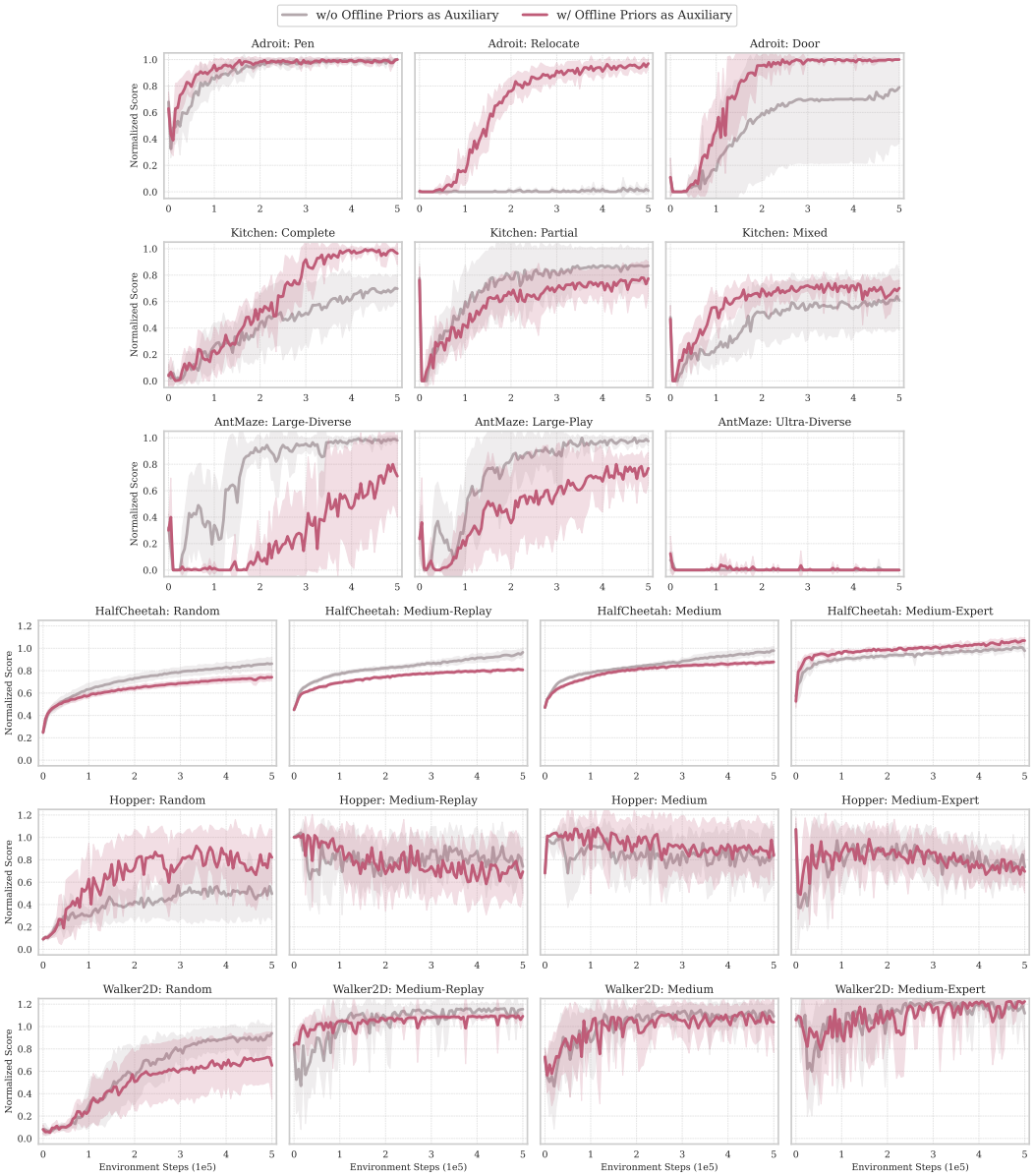

Figure 6: Effect of auxiliary reliance (β) without reference constraint. Both conditions use offline-trained initialization and standard SAC. Retaining offline data helps on some tasks but hurts on others.

Auxiliary (β), with reference constraint. This experiment repeats the auxiliary comparison with a reference constraint present: both conditions use Cal-QL's conservative penalty during online learning. The comparison isolates whether retaining offline data remains beneficial when conservative value estimation is also active. Comparing Figures 6 and 7 reveals how the effect of auxiliary data depends on whether a reference constraint is in place, further confirming that the reliance parameters interact.

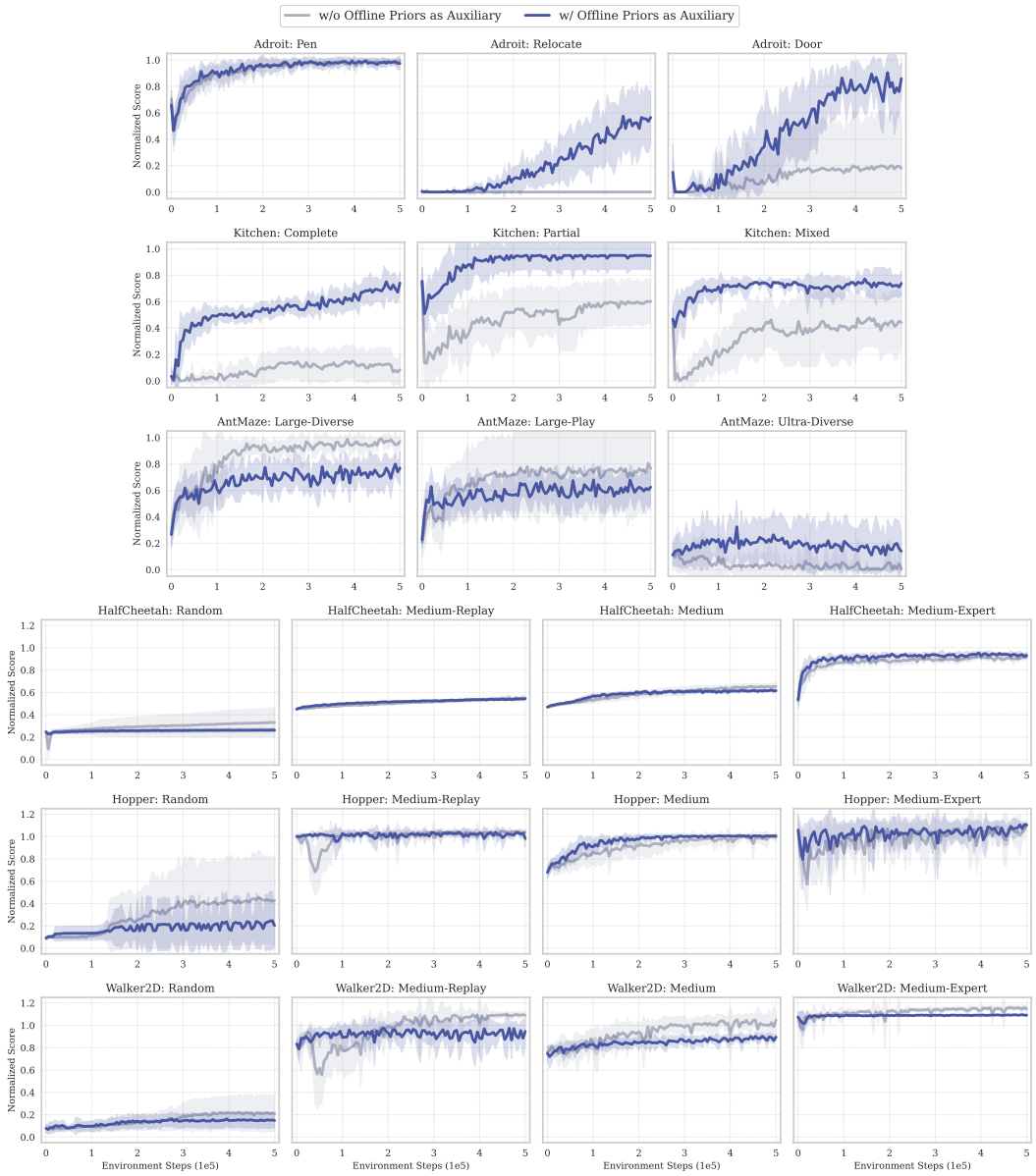

Figure 7: Effect of auxiliary reliance (β) with reference constraint ($\lambda > 0$). Both conditions use offline-trained initialization and Cal-QL's conservative penalty. The effect of retaining offline data can differ from Figure 6, demonstrating interaction between reliance parameters.

Summary. Across all five experiments and all task domains, the same pattern emerges: toggling any single reliance parameter produces help-or-hurt reversals across tasks. No setting of μ, λ, or β is uniformly beneficial. The paired experiments further reveal that the effect of one parameter depends on the configuration of the others: the impact of a reference constraint differs depending on whether auxiliary data is present (Figures 4 vs 5), and the impact of auxiliary data differs depending on whether a reference constraint is active (Figures 6 vs 7). This confirms that the reliance parameters are not independently tunable and supports the argument in Section 4.1 that tension management has no universal optimum.