ROAD: Adaptive Data Mixing for Offline-to-Online Reinforcement Learning via Bi-Level Optimization*

Letian Yang[1], **Xu Liu**[1], **Yiqiang Lu**[2], **Jian Liu**[2], **Weiqiang Wang**[2] and **Shuai Li**[1]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Ant Group, Shanghai, China

Abstract

Offline-to-online reinforcement learning harnesses the stability of offline pretraining and the flexibility of online fine-tuning. A key challenge lies in the non-stationary distribution shift between offline datasets and the evolving online policy. Common approaches often rely on static mixing ratios or heuristic-based replay strategies, which lack adaptability to different environments and varying training dynamics, resulting in suboptimal tradeoff between stability and asymptotic performance. In this work, we propose Reinforcement Learning with Optimized Adaptive Data-mixing (ROAD), a dynamic plug-and-play framework that automates the data replay process. We identify a fundamental objective misalignment in existing approaches. To tackle this, we formulate the data selection problem as a bi-level optimization process, interpreting the data mixing strategy as a meta-decision governing the policy performance (outer-level) during online fine-tuning, while the conventional Q-learning updates operate at the inner level. To make it tractable, we propose a practical algorithm using a multi-armed bandit mechanism. This is guided by a surrogate objective approximating the bi-level gradient, which simultaneously maintains offline priors and prevents value overestimation. Our empirical results demonstrate that this approach consistently outperforms existing data replay methods across various datasets, eliminating the need for manual, context-specific adjustments while achieving superior stability and asymptotic performance.

1 Introduction

Deep reinforcement learning (RL) has achieved remarkable access in various domains [Wang *et al.*, 2024; Mazyavkina *et al.*, 2021], yet its deployment in real-world applications is fundamentally restricted by the prohibitive sample complexity of online learning methods [Ladosz *et al.*, 2022; Kumar *et al.*, 2023], especially in safety-critical scenarios (e.g., healthcare) or costly environments (e.g., autonomous driving) [Yu *et al.*, 2021; Zhao *et al.*, 2025]. Offline RL offers a promising alternative by leveraging previously collected datasets to extract a policy without environment interaction [Levine *et al.*, 2020]. However, strictly offline learning is inherently limited by the quality and coverage of the static datasets, failing to reach a performance comparable to agents trained online. To bridge this gap, offline-to-online (O2O) RL has emerged as a new paradigm, aiming to initialize an agent with a safe offline policy (the offline phase) and subsequently refine it via direct online interaction (the online phase) [Nair *et al.*, 2020]. This paradigm seeks to maintain the safety of offline policy while enabling the plasticity to explore and improve beyond the offline dataset.

Under such a twofold objective, effective O2O training is non-trivial. The transition from offline pretraining to online fine-tuning introduces severe distribution shift [Lee *et al.*, 2022; Huang *et al.*, 2025]. Naively concatenating offline and online RL algorithms typically results in a catastrophic "unlearning" phenomenon, characterized by a significant performance dip at the beginning of online fine-tuning [Nakamoto *et al.*, 2023]. A majority of works balance pessimism from offline RL [Kumar *et al.*, 2020; Kostrikov *et al.*, 2022] and optimism required for online RL [Haarnoja *et al.*, 2018] through Q-value calibration [Nair *et al.*, 2020; Nakamoto *et al.*, 2023].

However, there have been limited works on addressing the distribution shift directly through data-level distribution alignment. Existing data-centric solutions largely rely on static strategies [Nakamoto *et al.*, 2023; Zhang *et al.*, 2023] or fixed heuristics [Lee *et al.*, 2022; Liu *et al.*, 2025b] to mix offline and online dataset. However, we argue that there is no universally optimal data mixing strategy. Instead, the optimal data replay pattern is tailored to the quality of the offline data, the environment properties and even the stage of training. For instance, a high proportion of offline data might be necessary in early training to stabilize updates, whereas later stages might benefit from prioritizing online experience to correct the bias from offline data.

In this paper, we investigate the mechanism by which data mixing strategies influence online fine-tuning, characterizing data mixing as a reshaping of the Q-learning loss landscape. From this perspective, we identify an objective misalignment problem in offline-to-online settings, where the conventional

*An extended version of this paper with all appendices is available at https://arxiv.org/abs/2605.14497.

Q-learning objective of minimizing the Bellman error does not suffice to guarantee RL performance.

To address this objective misalignment issue, we propose **r**einforcement learning with **o**ptimized **a**daptive **d**ata-mixing (ROAD), a dynamic, plug-and-play framework that adaptively determines the optimal data mixing strategy. We fundamentally reframe the data selection problem as a bi-level optimization process in the functional space of data mixing strategies, where the inner-level updates follow standard Q-learning, and the outer-level takes data mixing as a meta-decision problem to maximize the expected performance of the online policy, thereby aligning Q-learning with the ultimate goal of maximizing RL performance. A solution is provided by computing the gradient of the outer-level objective.

To further deal with the overestimation bias of the outer-level objective and thus bridge the gap between theory and practice, ROAD constructs a surrogate as an unbiased estimator for the outer-level objective. A practical algorithm utilizing a sliding-window multi-armed bandit (MAB) to optimize the data mixing ratio is proposed to optimize this surrogate objective. Intuitively, ROAD selects an arm that maximizes the performance increment in trusted regions where the uncertainty of the Q-function estimator is negligible and mitigates training instability by suppressing value overestimation.

In summary, our contributions are as follows:

- We identify an objective misalignment problem in data-centric O2O RL and solve this misalignment by formulating it as a bi-level optimization, where the data-mixing strategy is considered a meta-decision variable.
- We propose ROAD, a plug-and-play framework that approximates the bi-level optimization theory with a computable, unbiased surrogate objective and a MAB mechanism optimizing data mixing ratios in real-time.
- We evaluate ROAD across extensive benchmarks and the results show that ROAD outperforms static and heuristic-based data selection strategies, ensuring training stability and asymptotic performance.

2 Related Works

The primary challenge of O2O RL lies in the mismatch between the principles underlying the two phases. Specifically, offline algorithms feature restrictive policy constraints and pessimistic Q-value estimation [Kumar *et al.*, 2020; Kostrikov *et al.*, 2022] to ensure conservatism, which inherently limits the exploration capabilities during online fine-tuning [Xie *et al.*, 2021; Song *et al.*, 2022; Wagenmaker and Pacchiano, 2023], resulting in slow performance gains. To mitigate this, transitional frameworks have been proposed. Some algorithms have adopted implicit or dynamic constraints [Nair *et al.*, 2020; Zhao *et al.*, 2022; Lee *et al.*, 2022] which naturally loosen in the online phase. Similarly, Cal-QL [Nakamoto *et al.*, 2023] conducts offline Q-value estimation in a calibrated rather than pessimistic manner; WSRL [Zhou *et al.*, 2025] introduces a separate warm-start phase to calibrate Q-value and then discards the offline dataset. Another line of work embraces the discrepancy between offline and online settings [Liu *et al.*, 2025a]. APL [Zheng *et al.*, 2023] employs pessimistic updates on offline data and optimistic updates on online data; PEX [Zhang *et al.*, 2023] maintains separate policies for offline and online learning; MOORL [Chaudhary *et al.*, 2025] treats the O2O RL problem as a meta-learning problem with two distinct tasks, offline and online learning.

2.1 Offline Data Replay for Offline-to-Online RL

While algorithmic consistency is crucial, the strategy to address data-level distribution shift plays a decisive role in ensuring stability and plasticity in the online fine-tuning phase. Although some argue that offline data can be completely discarded in online learning [Zhou *et al.*, 2025], its adaptiveness to various environments is questioned [Li *et al.*, 2025a]. The predominant solution to offline data utilization during online interaction is to replay offline data in the online replay buffer.

Static Mixing Strategies Early attempts initialize the replay buffer with offline dataset [Hester *et al.*, 2018; Vecerik *et al.*, 2017], which proves unstable [Hansen *et al.*, 2023]. Instead, the current mainstream is to sample independently from two data sources with a mixing ratio. QT-Opt gradually increase the online data ratio from 1% to 50% [Kalashnikov *et al.*, 2018]. Cal-QL, RLPD and PROTO implement symmetric sampling (a fixed 50%-50% split between offline and online data) [Nakamoto *et al.*, 2023; Ball *et al.*, 2023; Li *et al.*, 2023], which is also adopted in concurrent researches [Wu *et al.*, 2025; Li *et al.*, 2025b]. While simple, these static strategies are inherently suboptimal for lack of context awareness and environment adaptiveness.

Heuristic-based Data Selection Recognizing the limitations of fixed mixing ratios, there have been works toward selective sampling. Balanced replay and following works prioritizes "near on-policy" offline samples [Lee *et al.*, 2022; Liu *et al.*, 2025b; Song *et al.*, 2025]. Despite mitigating distribution shift, near on-policy selection heuristics reduces data selection to a form of online data augmentation, limiting the algorithm's ability to fully leverage the environmental information encoded in the offline dataset. Our empirical findings, together with prior analytical works [Li *et al.*, 2025a], indicate that no single sampling strategy generalizes universally, which motivates an adaptive, context-sensitive strategy that adapts to learning dynamics and environmental feedback [Nakamoto *et al.*, 2023; Zheng *et al.*, 2023].

3 Bi-Level Optimization for O2O RL

In this section, we formulate data mixing as an optimization problem. From a functional perspective, we offer a novel perspective on how different data mixing strategies influence the training dynamics. Under this perspective, we argue that an optimal mixing strategy should be derived by aligning the Q-learning objective with the true RL objective. Based on this formulation, we present an idealized optimization algorithm as a theoretical solution.

3.1 Objective Misalignment in O2O Fine-Tuning

We formulate the O2O RL problem within the Markov Decision Process (MDP) framework $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$. The online fine-tuning begins with a static offline dataset $\mathcal{D}^{\text{offline}}$

generated by a behavior policy π_β and an offline initialization for the policy π^{off} and the Q-function f^{off} in a function class $\mathcal{F}$. Let the initial state distribution be $\mu \in \Delta(\mathcal{S})$ and ρ be any trajectory, then the goal is to maximize the cumulative value function, denoted by $J(\pi) = \mathbb{E}_{s\sim\mu}[V^\pi(s)] = \mathbb{E}_{\rho\sim\pi, s\sim\mu}\left[\sum_{t=0}^{\infty}\gamma^t r_t | s_0 = s\right]$.

Online fine-tuning via fitted Q-iteration. Standard value-based online fine-tuning algorithms typically follow a fitted Q-iteration (FQI) scheme [Nakamoto *et al.*, 2023]. At each iteration k, the agent collects an online data batch $\mathcal{D}^k$ with the online policy π^k. Then the Q-value iteration is performed on a mixed distribution d^{sample} built with both the offline dataset and online replay buffer. Specifically, the Q-function updates with Equation (1) where $\mathcal{T}$ is the Bellman operator defined as $\mathcal{T}f(s,a) = r(s,a) + \gamma\mathbb{E}_{s'}[\max_{a'} f(s',a')]$.

$$f^{k+1} \leftarrow \arg\min_{f\in\mathcal{F}} \left\{\mathbb{E}_{d^{\text{sample}}}\left[\mathcal{T}f^k(s,a) - f(s,a)\right]^2\right\}. \quad (1)$$

Crucially, the construction of the mixed distribution is formulated via a data mixing strategy $g \in \mathcal{G}$, defined as a mapping from online and offline data sources to a sampling distribution, where $\mathcal{G}$ is the function space of all such strategies.

$$d_g^{\text{sample}} = g(\mathcal{D}^{\text{offline}}, \mathcal{D}^1, ..., \mathcal{D}^k). \quad (2)$$

The loss landscape perspective on data mixing. Equation (1) reveals a critical insight that Q-learning is essentially projecting the Bellman target onto the representable function space $\mathcal{F}$, i.e., $f^{k+1} \leftarrow \Pi_{\mathcal{F}, d_g^{\text{sample}}}\left(\mathcal{T}f^k\right)$. Here $\Pi_{\mathcal{F}, d_g^{\text{sample}}}$ denotes the projection operator defined by weighted L_2-norm $\|\cdot\|_{d_g^{\text{sample}}}$. This formulation makes explicit the mechanism through which data mixing strategies affect online RL fine-tuning, as depicted below. The data mixing strategy governs the geometry of the projection. By altering the sampling density over the state-action space, different mixing strategies define distinct loss landscapes, which further steers the optimization toward different optima $f^*(g)$. The induced variation in Q-function updates further yields different online policies and, ultimately, different RL performance levels $J(\pi)$.

$$g \rightarrow \text{loss landscape} \rightarrow f^* \rightarrow \pi \rightarrow J(\pi).$$

The misalignment between objectives. The dependency chain of variables above demonstrates that the primary obstacle to identifying ideal data mixing stems from its *indirect* influence on the training dynamics, which gives rise to an objective misalignment problem. Specifically, contrary to standard online RL, minimizing the Bellman error in online fine-tuning does not guarantee maximizing the policy return due to the extra variable g.

Existing approaches dependent on static mixing ratios and heuristic selection effectively lock the projection geometry based on fixed priors, assuming that a particular landscape is universally beneficial. However, this assumption is fundamentally flawed because the relationship between the projection geometry and online performance is dynamic, dependent on the available data, the online policy and the stage of training. As an example, geometry required for early-stage stability (anchoring to offline data to prevent catastrophic forgetting) is the inverse of that required for asymptotic plasticity (emphasizing online samples to correct offline bias). The fixed priors cut off the explicit feedback from the online policy performance, failing to satisfy these mutually exclusive requirements simultaneously.

To resolve this misalignment, we transcend static rules and treat g as a meta-decision variable to study its effect on the policy performance (our final objective) rigorously. Since aligning the Bellman objective with RL performance necessitates precise control over the loss landscape across the state-action space, we abstain from parameterizing g and formulate it directly as a functional meta-decision variable.

3.2 The Bi-Level Optimization Formulation and The Theoretical Solution

To address the objective misalignment problem, we reframe data mixing as a dynamic meta-decision process, solving an optimization problem in the space of mixing strategies to dynamically align the Q-learning objective with policy performance maximization. Notably, this optimization is inherently hierarchical because the final objective (policy performance) depends on the learned Q-function, which is itself an optimum for minimizing the Bellman error. The constraint from an optimal inner Q-learning loop naturally casts the data mixing problem into the framework of bi-level optimization.

Formally, the inner loop is defined as standard Q-learning updates, which we treat as a fixed subroutine. The objective $\mathcal{J}_{\text{in}}^k(g,f)$ is exactly the Bellman error to be minimized on the mixed dataset, as is defined in Equation (1). In order to dynamically determine the optimal mixing, the outer-loop is expected to make a decision on the meta-variable g within each FQI iteration, maximizing the expected return $J(\pi)$ of the induced policy. Using the chain of variable dependency, we can express the objective as a function of g, leading to the objective in Equation (3).

$$\max_{g\in\mathcal{G}} \mathcal{J}_{\text{out}}^k(g) = \max_{g\in\mathcal{G}} J\left(\pi\left(\arg\min_f \mathcal{J}_{\text{in}}^k(g,f)\right)\right)$$
$$\text{where } \mathcal{J}_{\text{in}}^k(g,f) = \mathbb{E}_{d_g^{\text{sample}}}\left[\mathcal{T}f^k(s,a) - f(s,a)\right]^2. \quad (3)$$

Here we assume a differentiable policy mapping from the inner-level optimal f^* to the behavior policy $\pi(f^*)$ to ensure that the gradient exists.

The theoretical solution. Inspired by recent breakthroughs in bi-level optimization theory [Petrulionytė *et al.*, 2024], we derive a theoretical solution to the proposed setting. Within each FQI iteration, we solve the optimization via gradient ascent. The critical part is the gradient calculation of the outer-level objective demonstrated in Equation (4).

$$\nabla_g \mathcal{J}_{\text{out}}(g) = -\int_{\mathcal{S}\times\mathcal{A}} w_g(s,a)\partial_{g,f}\mathcal{J}_{\text{in}}(s,a)\mathrm{d}s\mathrm{d}a. \quad (4)$$

Here w_g denotes the density ratio between the distribution of a policy induced by the online policy and its advantage function, and the sampling distribution; the mixed Fréchet derivative $\partial_{g,f}\mathcal{J}_{\text{in}}$ evaluates the sensitivity of Bellman error to the data mixing strategy. Intuitively, Equation (4) implies that the ideal g^* induces a weighted on-policy distribution biased towards the loss-sensitive samples.

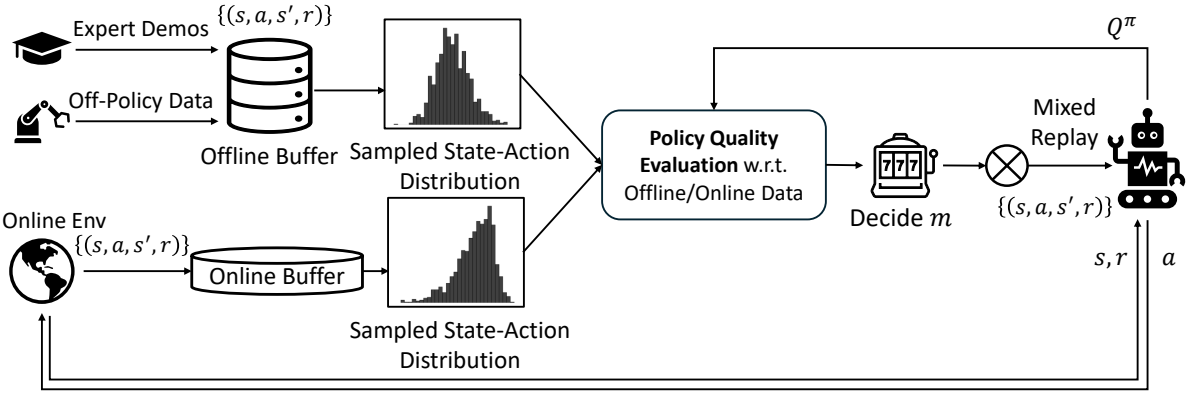

Figure 1: Illustration of the online training scheme of ROAD. In the online training phase, ROAD controls the offline data replay pattern by selecting the mixing ratio in an adaptive fashion with the derived surrogate objective for the bi-level optimization problem. The mixing ratio m adjusted by ROAD controls the mixed replay for the offline dataset and the online buffer.

4 ROAD: Towards Practical RL Fine-Tuning

Despite an idealized procedure for optimal data mixing strategy in Section 3.2, directly solving for an optimum with Equation (4) is intractable in online settings. Consequently, in this section, we present a practical algorithm for the bi-level optimization, ROAD, for online RL fine-tuning.

4.1 Bridging Theory with Practice

The primary obstacle to implementing theoretical gradient ascent problem are the *estimation bias* of the Q-function and severe *numerical instabilities* in gradient computation. First, in practice, $J(\pi(f^*))$ is a biased estimation for the outer-level objective because π is optimized against a Q-function f^* characterized by uncertainty. Maximizing this biased objective inevitably misguides the optimization process, which necessitates an unbiased estimator of $J(\pi)$. Furthermore, computing this gradient involves multiple density estimations including that for a rapidly evolving online policy d^π. These estimations are inherently unstable on online data, and the arithmetic operations further aggravate this instability. Meanwhile, the high variance associated with the mixed Fréchet derivative term easily mislead the gradient direction in sampling-based estimation. Consequently, to make unbiased estimation on the outer-level objective and meanwhile ensure computational feasibility, we adopt a zero-order optimization approach, employing a MAB mechanism to search for an optimal data mixing strategy. We establish our practical algorithm on the following two key assumptions.

Assumption 1. *The offline dataset offers sufficient number of samples within its support, leading to negligible uncertainty.*

Assumption 2. *The training dynamics varies slowly.*

Assumption 1 posits that offline data serves as an anchor. Within this "trust region", it is assumed that uncertainty does not significantly distort the perceived relative value of actions. Meanwhile, Assumption 2 is the fundamental premise for modeling non-stationary environments using MAB. It allows for a sliding-window approach to evaluate the performance of data mixing strategies.

An unbiased objective estimator. An unbiased estimation for the outer-level objective is shown in Equation (5). The empirical estimation error $\epsilon = \hat{f} - f$ is assumed to be a zero-mean random field that is independent of f and is characterized by a continuous covariance kernel K, which effectively measures the epistemic uncertainty.

$$\begin{aligned}\mathcal{J}^k_{\text{out}}(g) &= \mathbb{E}_{a\sim d^\pi}\left[\hat{f}^*\right] - \mathbb{E}_\epsilon\left[\text{Bias}_{\hat{f}^*}\right] \\ &= \mathbb{E}_{a\sim d^\pi}\left[\hat{f}^*\right] - \iint_{\mathcal{A}\times\mathcal{A}} \frac{\delta\pi(a)}{\delta\hat{f}^*(a')} K(a,a')\mathrm{d}a\mathrm{d}a' + R_3, \quad (5)\end{aligned}$$

where $\|R_3\| = o\left(\|\epsilon(s)\|^3\right)$ is the third-order remainder and the dependency on s is omitted for brevity.

Linearized mixing strategies. To tackle computational intractability, we restrict the search space to linear combinations of offline and online distributions, parameterized by an offline data replay ratio $m \in [0,1]$. Then the mixed data distribution is $d^{\text{sample}}_m(s,a) = m d^{\text{off}}(s,a) + (1-m) d^{\text{on}}(s,a)$. To ensure compatibility with the MAB search mechanism, the candidate set of mixing ratios Λ is a fixed discrete one.

4.2 The Practical Algorithm

In this section we construct a robust surrogate objective utilized as the reward within a sliding-window MAB framework [Liu *et al.*, 2024]. Our approach relies exclusively on readily available statistics and avoids complex auxiliary techniques such as density estimators or Q-ensembles [Lee *et al.*, 2022], incurring virtually no additional computational cost.

The surrogate objective. By approximating the unbiased outer-level objective with statistics directly from the training process, we formulate our surrogate objective as Equation (6), where κ is a hyperparameter. In this surrogate, Δ_{off} is the empirical estimator of RL performance improvement with respect to the offline behavior policy. By employing this advantage-like mechanism, we reduce the variance of performance estimation, leading to a stable MAB reward. Meanwhile, Δ_{on} is an approximation of the overestimation bias at

early stages, stabilizing the training process.

$R_q = \Delta_{\text{off}} - \kappa\Delta_{\text{on}}$ where

$$\begin{cases} \Delta_{\text{off}} = \mathbb{E}_{s\sim\mathcal{D}^{\text{off}},a\sim\pi_\theta}\left[Q_\phi(s,a)\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}^{\text{off}}}\left[Q_\phi(s,a)\right] \\ \Delta_{\text{on}} = \mathbb{E}_{s\sim\mathcal{D}^{\text{on}},a\sim\pi_\theta}\left[Q_\phi(s,a)\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}^{\text{on}}}\left[Q_\phi(s,a)\right]. \end{cases} \tag{6}$$

Theorem 1. *When the Q-function f is smoother than the estimation error ϵ, and the policy improvement follows a policy gradient step with small learning rate, R_q approximates of the unbiased objective with small variance.*

Proof sketch. The performance improvement term is approximated by Δ_{off} using the performance difference lemma [Kakade and Langford, 2002], while the overestimation bias equals Δ_{on} (signal) minus a true advantage term (noise). We formalize the smoothness of ϵ and f via their characteristic lengths, revealing that the signal-to-noise ratio of this overestimation bias estimate scales linearly with the ratio of these lengths and decays during training. Therefore, Δ_{on} yields a precise bias estimate initially. As training progresses, the magnitude of overestimation bias itself diminishes significantly so that Δ_{on} is effectively dominated by Δ_{off}, introducing negligible adverse effects to the training process. □

Intuitively, maximizing Δ_{off} encourages policy improvement in the offline "trust region", while minimizing Δ_{on} stabilizes online learning and prevents unlearning. Though seemingly counter-intuitive, Δ_{on} functions analogously to a min-max regularization by constraining aggressive policy updates and preventing the policy from exploiting the Q-function bias.

Implementation of ROAD. With a practical surrogate objective, we describe our algorithm ROAD as follows. ROAD works under a certain period (e.g., an episode or a certain number of steps) and maintains a recording list for the arm pulled and the surrogate objective (m^k, R_q^k) for each period k. When a period begins, ROAD selects a mixing ratio by maximizing the upper confidence bound (UCB) as is shown in Equation (7) and fixes the ratio during the period. In order to model the non-stationarity in the training dynamics, ROAD adopts a sliding window of size τ, which means that we select the best arm using only results from the recent τ periods.

$$m^k = \arg\max_{m\in\Lambda}\left\{\bar{R}_{q,m} + \sqrt{\frac{c\log(k\wedge\tau)}{N_k(\tau,m)}}\right\}, \tag{7}$$

Here $\bar{R}_{q,m}$ is the average of recorded R_q values for mixing ratio m within the sliding window; $N_k(\tau, m)$ denotes the total number of times m has been selected up to period k.

5 Experiment

We study how well ROAD can facilitate efficient training in the online phase with adaptive data mixing. To this end, we compare ROAD with several other data replay methods on a variety of offline-to-online RL benchmark tasks with various offline datasets. We also study the behavior of ROAD in adaptively adjusting its data replay pattern in various environments. Finally, we perform empirical studies to investigate the robustness of ROAD under different parameter settings.

O2O RL tasks and settings. Our benchmarking tasks and datasets include: **(1)** AntMaze tasks from D4RL [Fu *et al.*, 2020] which involves navigating an ant quadruped robot to a predetermined goal location within a maze; **(2)** Gym MuJoCo Locomotion tasks [Todorov *et al.*, 2012; Kostrikov *et al.*, 2022] that leverage the MuJoCo physics simulation engine targeted for robotic control with distinct morphologies and objectives; **(3)** FrankaKitchen tasks from D4RL [Nakamoto *et al.*, 2023] where a 9-DoF Franka robot is controlled to reach a target configuration within a kitchen setup.

Implementation details for ROAD. We examine the effect of ROAD and the baseline data replay approaches with multiple popular offline-to-online RL algorithms, including IQL [Kostrikov *et al.*, 2022], PEX [Zhang *et al.*, 2023], CQL [Kumar *et al.*, 2020], and Cal-QL [Nakamoto *et al.*, 2023], to illustrate the robustness of ROAD when integrated with various offline phase implementations. Due to the space limits, we report only the IQL-based results (Section 5.1). The candidate set Λ is $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and the hyperparameter $\kappa = 1$. The exploration parameter $c = 2.0$ in all cases to encourage adaptation to the varying training dynamics. The sliding-window size is $\tau = 1,000$ for all tasks except $\tau = k$ for Antmaze because these tasks feature sparse rewards which leads to high variance in estimating R_q.

Baselines We compare ROAD with the following baseline data mixing strategies: **(i)** **`Fixed`** [Zhang *et al.*, 2023; Nakamoto *et al.*, 2023] mixing ratios conduct online finetuning on a linear combination of online and offline distributions. We run the experiments across all fixed ratios in Λ to compare ROAD against the best one. **(ii)** **`Decreasing`** ratio adopts the heuristic that offline data stabilizes training at early stages, and thus decreases the mixing ratio linearly from 0.5 linearly to 0.1 (we do not take 0 as minimum for better performance). **(iii)** **`Balanced Replay`** (**`BR`**) [Lee *et al.*, 2022] adopts another family of heuristics that "nearly on-policy" samples help stabilize training. **(iv)** **`Uniform`** ratios select mixing ratios uniformly within the candidate set.

5.1 Empirical Results

ROAD shows solid performance and adaptivity across various environments. Table 1 provides an analysis of ROAD's performance using normalized metrics with IQL-based offline phase as an example. **`Fixed`** mixing ratios have a performance heavily dependent on the chosen ratio. While some environments are less sensitive, others exhibit pronounced differences. Heuristic-based strategies like **`Decreasing`** and **`BR`** offer improved robustness to data quality variations but lack consistent superiority. In contrast, ROAD presents consistently strong or comparable performance across all tested environments, emerging as the most stable and adaptive method evaluated. The comparison against **`Uniform`** arm selection demonstrates the efficacy of our proposed surrogate R_q.

ROAD shows adaptivity across various offline data qualities and RL algorithms. In our benchmarking environments, ROAD showcases its ability to handle diverse offline data qualities. The performances of baselines exhibit varying performance depending on dataset quality while struggling to

Tasks	Fixed Mixing Ratios						Uniform	Decreasing	BR	ROAD
	0.0	0.1	0.2	0.3	0.4	0.5				
antmaze-large-diverse	56.98±3.33	60.32±4.08	55.98±2.26	56.01±2.29	50.66±3.79	50.35±0.86	60.83±4.76	52.15±0.25	46.83±6.74	**63.13±2.24**
antmaze-large-play	58.16±2.63	46.51±1.63	53.16±5.70	53.66±0.61	46.82±2.02	55.12±3.63	46.32±9.76	52.17±1.28	38.34±1.24	**59.75±2.97**
antmaze-medium-diverse	80.00±4.55	78.68±4.48	80.50±5.00	81.66±2.01	82.67±1.85	83.29±3.71	82.30±2.01	81.82±2.52	81.84±0.75	**83.67±1.39**
antmaze-medium-play	82.50±3.11	79.17±2.40	77.32±1.92	80.16±1.66	82.00±2.95	78.62±3.53	82.67±0.26	77.16±3.02	80.99±0.99	**83.49±3.62**
antmaze-umaze-diverse	63.49±13.79	69.66±15.78	7.83±2.24	38.50±33.15	46.51±21.76	16.13±8.51	25.65±10.24	48.82±23.49	**80.66±3.99**	72.12±23.29
antmaze-umaze	92.16±0.94	93.32±0.96	93.16±0.85	93.34±1.32	94.33±0.47	94.37±0.24	91.32±1.76	92.99±0.50	94.33±0.75	**95.83±0.30**
halfcheetah-random	41.48±2.21	43.50±6.28	48.25±3.30	39.62±3.10	43.71±6.07	44.39±4.37	45.91±0.06	42.72±0.22	47.43±1.12	**49.37±3.57**
halfcheetah-medium-replay	50.70±3.38	54.54±1.81	52.55±1.72	51.55±1.90	50.11±1.14	47.41±0.34	50.07±0.05	53.55±3.13	49.28±2.29	**55.82±1.07**
halfcheetah-medium	69.61±1.58	73.53±1.74	72.23±1.72	69.47±4.25	69.39±2.61	67.87±1.08	67.56±4.61	70.46±0.30	72.49±1.51	**74.57±1.95**
halfcheetah-medium-expert	63.75±5.65	92.75±1.40	92.45±1.48	94.52±1.13	93.07±1.16	91.54±5.56	92.83±1.18	93.88±1.70	93.34±2.33	**95.06±0.55**
halfcheetah-expert	78.28±3.58	94.17±1.65	94.64±0.66	95.65±0.36	95.94±0.47	96.61±0.45	93.83±1.76	93.15±0.32	94.91±0.77	**96.86±0.39**
walker2d-random	6.57±1.50	8.72±2.36	8.13±0.38	10.00±1.52	9.20±1.08	10.65±1.66	8.58±0.31	9.56±1.00	7.58±0.86	**12.43±0.69**
walker2d-medium-replay	46.55±24.38	34.56±7.61	34.94±3.47	63.26±13.31	43.33±9.41	40.98±20.32	57.70±15.01	39.16±2.18	70.26±7.95	**78.15±15.74**
walker2d-medium	50.28±7.17	59.07±14.22	56.72±9.78	44.59±9.07	55.74±16.19	31.52±8.42	39.78±1.67	**64.30±10.13**	40.11±19.46	60.17±8.51
walker2d-medium-expert	79.40±11.42	76.15±13.66	68.21±6.43	73.54±11.20	62.06±28.48	60.48±4.65	59.12±22.35	82.93±7.30	81.88±6.36	**83.60±13.06**
walker2d-expert	86.02±12.41	**89.75±10.80**	41.59±23.56	56.49±18.25	73.62±9.23	58.19±9.67	52.01±8.22	74.28±6.07	79.64±2.37	88.06±2.09
hopper-random	17.43±10.05	18.54±6.79	19.71±2.87	12.82±3.29	13.51±5.57	15.51±5.41	13.77±3.61	19.80±1.37	21.48±2.84	**31.77±4.44**
hopper-medium-replay	54.46±29.79	80.14±25.89	67.24±13.88	71.00±25.77	67.35±9.05	77.74±29.92	97.93±13.56	77.12±9.42	88.27±24.20	**99.14±9.97**
hopper-medium	87.25±19.06	81.84±29.83	61.15±18.09	77.46±18.12	87.85±7.86	73.30±22.99	86.17±2.53	85.37±13.13	88.56±20.03	**99.23±1.05**
hopper-medium-expert	46.63±4.13	49.86±8.83	47.01±12.77	75.97±29.32	57.68±12.16	73.46±22.75	66.02±23.94	61.76±4.93	66.21±4.65	**94.08±16.78**
hopper-expert	71.28±24.84	72.74±12.96	62.33±14.68	57.80±3.67	52.69±10.61	54.07±1.54	50.81±5.70	58.80±4.32	39.62±5.91	**85.10±4.27**
kitchen-partial	4.16±13.27	38.17±11.61	39.55±4.71	23.76±9.43	18.74±14.94	41.87±18.71	22.04±0.00	32.93±28.74	29.99±29.37	**42.65±5.13**
kitchen-mixed	0.41±0.59	44.34±1.58	45.44±8.12	40.02±7.12	39.14±15.11	45.00±9.17	55.83±11.90	29.58±13.81	47.56±0.64	**56.62±6.18**
kitchen-complete	10.04±1.56	51.65±10.22	52.07±6.64	47.08±12.38	19.59±24.76	41.86±11.94	53.75±13.69	21.65±16.24	41.64±27.49	**66.25±17.63**
Average	54.07	62.15	57.28	58.66	56.49	56.26	58.45	59.00	61.80	**71.95**

Table 1: Performance of ROAD and the baseline methods for offline data replay under Antmaze tasks, Locomotion tasks and Kitchen tasks with IQL. All results are assessed across 4 random seeds. The best performance for fixed mixing ratios is underlined, and the best-performed score is bolded.

Figure 2: Heatmap for mixing ratio selection in ROAD. In some cases, ROAD could learn to converge to the mixing ratio that performs best when fixed (left three) as training progresses, while in some other cases it could adjust its mixing ratio pattern more adaptively (right two).

adapt to changing data quality. For example, when an initially inferior offline policy undergoes training and outperforms the offline dataset, it is generally intuitive to switch the data mixing strategy. By contrast, ROAD consistently matches or exceeds baseline performance across different dataset qualities. This adaptability extends to various RL algorithms, including PEX, Cal-QL and CQL. ROAD also works effectively with RLPD [Ball *et al.*, 2023], where no pretraining is involved. These results highlight ROAD's flexibility and robustness in diverse settings, adapting seamlessly to varying algorithms and offline data conditions.

5.2 Understanding ROAD

The dynamics of R_q. Our surrogate objective R_q leverages in-training statistics to capture the training dynamics and guide data mixing decisions accordingly. To intuitively illustrate the physical interpretation of our bi-level optimization theory, we construct a simple MDP and a corresponding offline dataset, and plot the curves of Δ_{off} and Δ_{on} throughout the online fine-tuning phase. As is depicted by Figure 3, Δ_{off} consistently acts as a proxy for policy improvement. Conversely, early fine-tuning is safely constrained by minimizing Δ_{on}. Δ_{on} later decays in scale, leaving it as a noise term in R_q and thus preventing it from suppressing true policy improvement. This allows ROAD to ensure training stability while simultaneously maintaining asymptotic performance.

The behavior of ROAD. To analyze how ROAD adapts mixing ratios of offline and online data across environments, we visualized its decisions during training using heatmaps. These heatmaps, shown in Figure 2, capture the evolution of mixing ratios for different environments and offline datasets. Specifically, different environments witness different type of data mixing strategies. In some cases, ROAD converges to a fixed ratio, while in other cases, ROAD adapts its strategy to the altered training dynamics. By exploring diverse ratios early in training, ROAD enables policies to become more robust to varying state-action distributions, leading to improved learning outcomes and more effective integration of data sources.

Hyperparameter choice of κ	0 ($R_q = \Delta_{\text{off}}$)	0.2	0.5	1.0	2.0	5.0	∞ ($R_q = -\Delta_{\text{on}}$)
halfcheetah-medium	72.25 ± 2.70	75.05 ± 1.10	77.09 ± 0.91	74.57 ± 1.95	76.33 ± 1.56	74.33 ± 1.48	68.31 ± 1.86
halfcheetah-medium-replay	55.08 ± 1.29	57.25 ± 2.71	53.51 ± 1.68	55.82 ± 1.07	55.16 ± 0.82	54.74 ± 1.83	47.98 ± 3.41
hopper-medium	82.88 ± 15.84	85.60 ± 16.13	91.58 ± 4.46	99.23 ± 1.05	97.11 ± 4.57	94.15 ± 4.75	73.96 ± 8.17
hopper-medium-replay	52.78 ± 14.83	98.28 ± 7.81	96.11 ± 6.49	99.14 ± 9.97	85.00 ± 14.93	93.93 ± 6.29	93.39 ± 3.73

Table 2: Ablation study on the offline and online components of relative policy quality in ROAD.

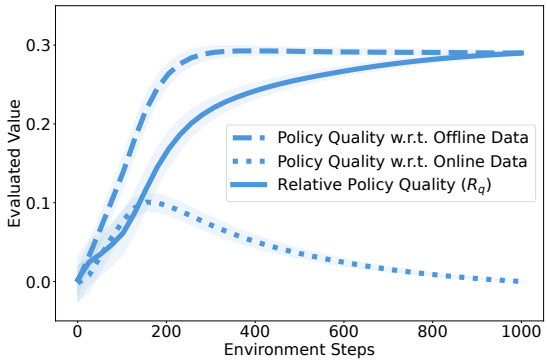

Figure 3: The R_q curve in online fine-tuning.

	τ		
c	500	1000	$\tau = t$
0.5	72.82	71.52	71.49
1.0	73.19	74.08	72.33
2.0	72.80	74.34	73.05
4.0	71.72	73.90	74.16

Table 3: ROAD performance across c and τ in Halfcheetah (mean).

5.3 Ablation and Parameter Sensitivity Study

The hyperparameter κ. This hyperparameter effectively balances aggressive policy improvement and conservative offline data utilization. Two extremes would be $\kappa = 0$ and $\kappa \to \infty$ corresponding to ablating the contributions of the offline and online components of relative policy quality. Results in Table 2 demonstrate that the full ROAD framework, combining both offline and online components, consistently outperforms settings that rely solely on one component. However, ROAD demonstrates low sensitivity to κ provided that κ falls within a moderate range. In these cases, ROAD works as expected, functioning as a stabilizer initially (dominated by Δ_{on}) and optimizes asymptotic performance later on (dominated by Δ_{off}).

The sliding-window size τ and exploration parameter c. Two critical parameters in balancing the exploration and exploitation of the ROAD framework are the exploration parameter c and the sliding-window size τ. To explore the effects of these parameters, we conducted experiments on the Halfcheetah environment with various combinations of c and τ, as detailed in Table 3. The results demonstrate that ROAD's performance remains robust across different parameter settings, highlighting its resilience to parameter variations. Typically, setting $\tau = 1000$ better captures the slowly-varying training dynamics in most environments, leading to better asymptotic performance. This indicates ROAD's broad adaptability to different environments and offline datasets.

6 Conclusions, Limitations and Future Work

In this work, we provide a novel perspective on how data mixing affects offline-to-online RL fine-tuning. Under this perspective, we identify the objective misalignment problem, where minimizing the Bellman error does not translate to maximized policy performance. To address this misalignment, we propose **R**einforcement Learning with **O**ptimized **A**daptive **D**ata-mixing, a novel, plug-and-play framework that formulates data mixing as a bi-level optimization problem. By treating data mixing strategy as a meta-decision variable, ROAD dynamically aligns the inner-level Q-learning objective with the goal of optimizing policy performance.

Our theoretical analysis of the optimization problem yields a gradient-based solution. Central to bridging the theory and practice is the derivation of a theoretically grounded surrogate objective R_q which approximates the unbiased estimator of the outer-level objective. This surrogate encourages policy improvement in trusted regions (via Δ_{off}) and suppresses value overestimation (via Δ_{on}). Extensive empirical evaluations demonstrate that ROAD consistently outperforms existing methods dependent on static and heuristic-based data replay strategies. ROAD exhibits robustness across various backbone algorithms and diverse offline data qualities.

Despite the theoretical rigor and promising empirical results, ROAD has limitations that merit discussion. The primary limitation lies in the assumption of linearized mixing strategies and further discretizations of the search space. These assumptions ensure computational feasibility but are essentially projecting our theoretical solution in the infinite-dimensional function space $\mathcal{G}$ to a low-dimensional subspace, compromising the rich expressivity of the function space. Another limitation stems from Assumption 1, indicating reliance on the offline dataset. In scenarios with extremely sparse or narrow offline datasets where the Q-value estimation is not sufficiently anchored, the reliability of R_q might degrade.

The promising directions primarily correspond to the limitation of restricted strategy search spaces. One natural extension to ROAD is to incorporate sample-level data selection (as suggested by the w_g term in Equation (4)), effectively introducing larger search spaces of mixing strategies. Another future work would be exploring methods to search for the optimal mixing strategy in a continuous space, thus allowing for more delicate control over the training dynamics. Apart from these methodological extensions, the application of this framework to real-world robotic tasks, which emphasizes the algorithm's ability to prevent unlearning while ensuring asymptotic performance, is also of interest.

Acknowledgments

This work was supported by Ant Group Research Fund.

Contribution Statement

The authors Letian Yang and Xu Liu contributed equally.

References

[Ball *et al.*, 2023] 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.

[Chaudhary *et al.*, 2025] Gaurav Chaudhary, Washim Uddin Mondal, and Laxmidhar Behera. MOORL: A framework for integrating offline-online reinforcement learning. *Transactions on Machine Learning Research*, 2025.

[Fu *et al.*, 2020] 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.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

[Hansen *et al.*, 2023] Nicklas Hansen, Yixin Lin, Hao Su, Xiaolong Wang, Vikash Kumar, and Aravind Rajeswaran. Modem: Accelerating visual model-based reinforcement learning with demonstrations. In *The Eleventh International Conference on Learning Representations*, 2023.

[Hester *et al.*, 2018] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Huang *et al.*, 2025] Xiao Huang, Xu Liu, Enze Zhang, Tong Yu, and Shuai Li. Offline-to-online reinforcement learning with classifier-free diffusion generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25581–25592. PMLR, 13–19 Jul 2025.

[Kakade and Langford, 2002] Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

[Kalashnikov *et al.*, 2018] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

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

[Kumar *et al.*, 2020] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020.

[Kumar *et al.*, 2023] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 112(7):2433–2467, 2023.

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

[Lee *et al.*, 2022] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pages 1702–1712. PMLR, 2022.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

[Li *et al.*, 2023] Jianxiong Li, Xiao Hu, Haoran Xu, Jingjing Liu, Xianyuan Zhan, and Ya-Qin Zhang. Proto: Iterative policy regularized offline-to-online reinforcement learning. *arXiv preprint arXiv:2305.15669*, 2023.

[Li *et al.*, 2025a] 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.

[Li *et al.*, 2025b] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Liu *et al.*, 2024] Xutong Liu, Siwei Wang, Jinhang Zuo, Han Zhong, Xuchuang Wang, Zhiyong Wang, Shuai Li, Mohammad Hajiesmaili, {John C.S.} Lui, and Wei Chen. Combinatorial multivariant multi-armed bandits with applications to episodic reinforcement learning and beyond. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 32139–32172, July 2024. 41st International Conference on Machine Learning (ICML 2024) ; Conference date: 21-07-2024 Through 27-07-2024.

[Liu *et al.*, 2025a] Xu Liu, Haobo Fu, Stefano V Albrecht, QIANG FU, and Shuai Li. Online-to-offline RL for agent alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Liu *et al.*, 2025b] Xuefeng Liu, Hung T. C. Le, Siyu Chen, Rick Stevens, Zhuoran Yang, Matthew Walter, and Yuxin Chen. Active advantage-aligned online reinforcement learning with offline data. In *The Exploration in AI Today Workshop at ICML 2025*, 2025.

[Mazyavkina *et al.*, 2021] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400, 2021.

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

[Nakamoto *et al.*, 2023] 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.

[Petrulionytė *et al.*, 2024] Ieva Petrulionytė, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Song *et al.*, 2022] Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

[Song *et al.*, 2025] Chihyeon Song, Jaewoo Lee, and Jinkyoo Park. Adaptive replay buffer for offline-to-online reinforcement learning. *arXiv preprint arXiv:2512.10510*, 2025.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[Vecerik *et al.*, 2017] Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

[Wagenmaker and Pacchiano, 2023] Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning*, pages 35300–35338. PMLR, 2023.

[Wang *et al.*, 2024] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincai Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 35(4):5064–5078, April 2024.

[Wu *et al.*, 2025] Mingdong Wu, Lehong Wu, Yizhuo Wu, Weiyao Huang, Hongwei Fan, Zheyuan Hu, Haoran Geng, Jinzhou Li, Jiahe Ying, Long Yang, Yuanpei Chen, and Hao Dong. Simlauncher: Launching sample-efficient real-world robotic reinforcement learning via simulation pre-training, 2025.

[Xie *et al.*, 2021] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Comput. Surv.*, 55(1), November 2021.

[Zhang *et al.*, 2023] Haichao Zhang, Wei Xu, and Haonan Yu. Policy expansion for bridging offline-to-online reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[Zhao *et al.*, 2022] Yi Zhao, Rinu Boney, Alexander Ilin, Juho Kannala, and Joni Pajarinen. Adaptive behavior cloning regularization for stable offline-to-online reinforcement learning. *arXiv preprint arXiv:2210.13846*, 2022.

[Zhao *et al.*, 2025] Jingyuan Zhao, Yuyan Wu, Rui Deng, Susu Xu, Jinpeng Gao, and Andrew Burke. A survey of autonomous driving from a deep learning perspective. *ACM Computing Surveys*, 57(10):1–60, 2025.

[Zheng *et al.*, 2023] Han Zheng, Xufang Luo, Pengfei Wei, Xuan Song, Dongsheng Li, and Jing Jiang. Adaptive policy learning for offline-to-online reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11372–11380, 2023.

[Zhou *et al.*, 2025] 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.