WOMBET: World Model-based Experience Transfer for Robust and Sample-efficient Reinforcement Learning

Mintae Kim MINTAE.KIM@BERKELEY.EDU
Koushil Sreenath KOUSHILS@BERKELEY.EDU
Hybrid Robotics, UC Berkeley
Berkeley, CA 94720, USA

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Reinforcement learning (RL) in robotics is often limited by the cost and risk of data collection, motivating experience transfer from a source task to a target task. Offline-to-online RL leverages prior data but typically assumes a given fixed dataset and does not address how to generate reliable data for transfer. We propose *World Model-based Experience Transfer* (WOMBET), a framework that jointly generates and utilizes prior data. WOMBET learns a world model in the source task and generates offline data via uncertainty-penalized planning, followed by filtering trajectories with high return and low epistemic uncertainty. It then performs online fine-tuning in the target task using adaptive sampling between offline and online data, enabling a stable transition from prior-driven initialization to task-specific adaptation. We show that the uncertainty-penalized objective provides a lower bound on the true return and derive a finite-sample error decomposition capturing distribution mismatch and approximation error. Empirically, WOMBET improves sample efficiency and final performance over strong baselines on continuous control benchmarks, demonstrating the benefit of jointly optimizing data generation and transfer.

Keywords: Offline-to-Online RL, Experience Transfer, World Models

1. Introduction

Reinforcement learning assumes abundant interaction and frequent resets–conditions rarely satisfied in real-world robotics, where data collection is expensive and unsafe Radosavovic et al. (2024); Li et al. (2025); Gupta et al. (2025). As a result, RL methods achieve strong asymptotic performance but remain sample-inefficient. Like biological systems that continually reuse past experience through internal world models, we seek to enable *experience transfer*: leveraging data collected in a source environment to improve sample efficiency and robustness in a target environment Wilson and McNaughton (1994); Lee and Wilson (2002). Existing paradigms address this only partially. Online RL adapts to new tasks but requires extensive interaction and faces distributional shift between source and target experiences Feng et al. (2023); Zhou et al. (2025); Kim and Sreenath (2026). Offline RL avoids unsafe exploration but but degrades under limited or mismatched data Yu et al. (2020); Kumar et al. (2020); Kostrikov et al. (2021); Yu et al. (2021). Applying it to real-world robotics often requires high-quality expert demonstrations or extensively tuned model-based controllers–assumptions that are unrealistic in many cases Cai et al. (2024); Kim et al. (2025); Kim (2026). Offline-to-online RL combines these by pretraining offline and fine-tuning online Nair et al. (2020); Lee et al. (2022); Rafailov et al. (2023); Smith et al. (2023); Singh et al. (2020), but assumes access to high-quality offline data without addressing how to obtain or safely reuse it. Model-based RL (MBRL) provides a principled way to generate data by learning a predictive

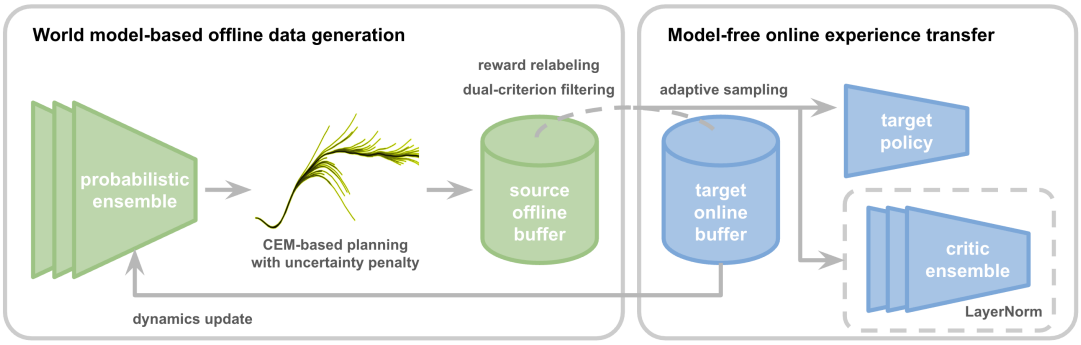

Figure 1: WOMBET pipeline: uncertainty-aware model-based data generation in the source task, followed by adaptive offline-to-online learning with iterative model updates.

world model. Planning-based approaches use the world model for control via model predictive control (MPC) Chua et al. (2018); Wang and Ba (2019), while Dyna-style methods use the model to generate synthetic rollouts for policy learning Janner et al. (2019); Sutton (1991). However, both faces limitations for transfer: MPC-based methods tend to generate low-diversity trajectories due to exploitation, and model-generated data can introduce bias when used without reliability control. Consequently, existing approaches struggle to provide reliable and adaptable data for transfer.

We propose *World Model-based Experience Transfer* (WOMBET), a unified framework that couples uncertainty-aware model-based offline data generation with online fine-tuning through iterative *model-policy co-evolution*. In each iteration, a world model generates trajectories in the source task via uncertainty-penalized MPC, and a dual criterion–high return and low epistemic uncertainty–filters reliable rollouts to form an offline dataset D_S. WOMBET then performs off-policy policy optimization in the target task, to maximize a provable lower bound on the true return, while adaptive data mixing balances bias and variance between offline (D_S from the source task) and online (D_T collected in the target task) experience. The world model is continuously refined using the aggregated dataset $D = D_S \cup D_T$, improving prediction accuracy and expanding the reliable planning region. This alternating process unifies the exploitation strength of MBRL with the exploration capability of model-free RL. Our contributions are threefold:

(1) *Coupled data generation and transfer.* We introduce a framework that jointly generates and utilizes prior data for transfer, instead of assuming a fixed offline dataset. WOMBET uses uncertainty-aware planning to construct data and iteratively refines both the world model and policy.

(2) *Reliable data curation and adaptive transfer.* We propose (i) a dual-criterion filter selecting trajectories with high return and low epistemic uncertainty, and (ii) an adaptive sampling strategy that balances offline and online data based on TD error.

(3) *Theoretical and empirical validation.* We show that uncertainty-penalized planning maximizes a lower bound on the true return and derive a finite-sample error decomposition capturing distribution mismatch and approximation error. Experiments demonstrate improved sample efficiency and final performance over strong baselines on continuous control benchmarks.

2. Related Work

Offline-to-online RL. This paradigm combines offline data efficiency with online adaptability to accelerate learning when interaction is costly Mao et al. (2022); Song et al. (2022); Rafailov et al. (2023). Early methods stabilize fine-tuning via advantage-weighted updates or constrained policy improvement Nair et al. (2020); Ball et al. (2023), while later work improves robustness through fixed offline critics, ensemble critics, and transition weighting Lee et al. (2022). Architectural com-

ponents such as symmetric replay buffers, ensemble critics, and layer normalization (LayerNorm) can further stabilize training without explicit regularization Ba et al. (2016); Ball et al. (2023). Other studies explore pretraining on related tasks to enhance exploration or show that keeping the offline dataset is unnecessary when pre-trained rollouts suffice for initialization Zhou et al. (2024). Behavior cloning and conservative value regularization are also used, but often require delicate tuning and may limit long-term performance Kumar et al. (2020); Yu et al. (2021). Learning-from-demonstration methods differ in relying on expert trajectories for imitation rather than addressing distributional shift during online fine-tuning Hester et al. (2018); Nair et al. (2018). Overall, most methods treat offline data as static and independent of the target task, leaving data generation, critic stability, and robust exploration unresolved.

MBRL for data generation and offline learning. Model-based approaches improve sample efficiency by learning a predictive world model and leveraging it either for planning or data generation. Planning-based methods perform control via MPC Chua et al. (2018); Wang and Ba (2019), while Dyna-style methods such as MBPO generate short synthetic rollouts to train model-free policies Janner et al. (2019). MOPO introduces uncertainty-based reward penalties to mitigate model exploitation Yu et al. (2020), and COMBO extends this with conservative value estimation Yu et al. (2021). The main difficulty in model-based offline RL is *model bias*: compounding prediction errors can degrade value estimation and lead to unsafe policies Rafailov et al. (2023). To address this, ensemble-based uncertainty penalizes unreliable or out-of-distribution (OOD) rollouts, often combined with behavior regularization or conservative value estimation. However, most methods use the learned model only once for rollout generation or regularization, without iterative refinement. WOMBET instead alternates between uncertainty-penalized data generation and online fine-tuning, updating both model and policy using aggregated experience. This iterative process unifies model-based offline RL and online adaptation for experience transfer.

3. Preliminaries

We consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \mu_0, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(s'|s,a)$, reward $r(s,a)$, initial distribution μ_0, and discount factor $\gamma \in [0,1)$. A policy $\pi(a|s)$ aims to maximize the discounted return $J(\pi) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. We study two tasks sharing dynamics but differing in reward and initial state: the *source task* $\mathcal{M}_S = (\mathcal{S}, \mathcal{A}, P, r_S, \mu_0^S, \gamma)$ and the *target task* $\mathcal{M}_T = (\mathcal{S}, \mathcal{A}, P, r_T, \mu_0^T, \gamma)$. The objective is to maximize the target return $J_T(\pi) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r_T(s_t, a_t)]$, using an offline dataset from $\mathcal{M}_S$ and online data from $\mathcal{M}_T$.

4. WOMBET: World Model-based Experience Transfer

Unlike standard offline-to-online RL, which assumes access to offline data, WOMBET addresses the problem of *generating* transferable experience. The challenge is not only utilizing prior data, but generating reliable data from a source task that can accelerate learning in a target task.

4.1. World Model-based Data Generation

A key component of WOMBET is the construction of the offline dataset $\mathcal{D}_S$ from a source task. Rather than assuming a given dataset, WOMBET generates $\mathcal{D}_S$ via world model-based planning.

A probabilistic ensemble $\hat{P}_S = \{\hat{P}_S^{(i)}\}_{i=1}^{E}$ approximates $P(s'|s,a)$ Chua et al. (2018). Under MPC, the optimal horizon-H action sequence solves

$$a^*_{t:t+H-1} = \arg\max_{a_{t:t+H-1}} \mathbb{E}_{\hat{P}_S}\left[\sum_{k=0}^{H-1} \gamma^k r_S(s_{t+k}, a_{t+k})\right]. \tag{1}$$

Only the first action a_t^* is executed. Repeating this procedure yields model-based rollouts $\mathcal{D}_S$.

In WOMBET, this process is iterative–the model $\hat{P}$ is continually refined with $\mathcal{D} = \mathcal{D}_S \cup \mathcal{D}_T$, progressively improving predictive accuracy and uncertainty estimation over iterations.

4.2. Robust Planning with a Provable Performance Lower Bound

Source experience $\mathcal{D}_S$ provides a strong initialization, but rollouts under the world model $\hat{P}$ remain prone to bias and over-optimism. To ensure reliability, WOMBET introduces an *uncertainty penalty* during MPC, forming a conservative surrogate that lower-bounds the true return under dynamics P. The true return $J_P(\pi) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{H-1} r(s_t, a_t)]$ deviates from the model return $J_{\hat{P}}(\pi) = \mathbb{E}_{\pi,\hat{P}}[\sum_{t=0}^{H-1} r(s_t, a_t)]$ due to model bias. Let $V_P^\pi(s) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{H-1} r(s_t, a_t) \,|\, s_0 = s]$ denote the value function of MPC under true dynamics. We assume that V_P^π is L_v-Lipschitz in state, i.e., $|V_P^\pi(s_1) - V_P^\pi(s_2)| \le L_v \|s_1 - s_2\|$ for all (s_1, s_2). We also assume that the ensemble uncertainty $u(s, a)$ upper-bounds the Wasserstein distance between transition kernels, $W_1(P, \hat{P}) \le u(s, a)$ for all (s, a). Define the one-step model bias $G^\pi(s, a) = \mathbb{E}_{\hat{P}}[V_P^\pi(s')] - \mathbb{E}_P[V_P^\pi(s')]$. Then

$$|G^\pi(s, a)| \le L_v \, W_1(P, \hat{P}) \le L_v \, u(s, a). \tag{2}$$

This follows from the Lipschitz property of V_P^π and the definition of W_1 (see Yu et al. (2020)).

Expanding the Bellman telescoping yields

$$J_P(\pi) \ge \mathbb{E}_{\pi,\hat{P}}\Big[\sum_{t=0}^{H-1} (r(s_t, a_t) - \lambda \, u(s_t, a_t))\Big], \tag{3}$$

for any $\lambda \ge L_v$. Since $J_P(\pi) - J_{\hat{P}}(\pi) = \sum_t \mathbb{E}_{\pi,\hat{P}}[G^\pi(s_t, a_t)]$, (2) and bounding each term by $L_v u(s_t, a_t)$ yields (3). Defining the penalized reward $\tilde{r}(s, a) = r(s, a) - \lambda u(s, a)$ gives

$$J_P(\pi) \ge \mathbb{E}_{\pi,\hat{P}}\Big[\sum_{t=0}^{H-1} \tilde{r}(s_t, a_t)\Big] =: \tilde{J}_{\hat{P}}(\pi), \tag{4}$$

so maximizing $\tilde{J}_{\hat{P}}$ under $\hat{P}$ maximizes a certified lower bound on J_P. Since (4) holds at each time horizon, the receding-horizon controller that greedily maximizes the penalized return satisfies

$$J_P(\pi_{\mathrm{MPC}}) \ge \mathbb{E}_{\pi_{\mathrm{MPC}},\hat{P}}\Big[\sum_{t=0}^{H-1} \tilde{r}(s_t, a_t^\star)\Big], \tag{5}$$

which is precisely the quantity optimized online. Thus, uncertainty-penalized MPC maximizes a provable lower bound on the true return, balancing reward exploitation under $\hat{P}$ with avoidance of high-uncertainty regions where $u(s, a)$ is large.

4.3. Dual-Criterion Filtering for Reliable Offline Data from MBRL

A central component of WOMBET is the generation of the offline dataset $\mathcal{D}_S$. Unlike prior work that assumes a given dataset, WOMBET generates $\mathcal{D}_S$ from rollouts of a source-trained world model $\hat{P}$, via model-based planning and explicitly controls its reliability through filtering. Since $\hat{P} \neq P$,

naïvely using all synthetic trajectories induces model bias. WOMBET controls dataset *reliability* at generation time via MPC planning and uncertainty-aware filtering. A probabilistic ensemble $\hat{P} = \{\hat{P}^{(i)}\}_{i=1}^{E}$ is used within MPC to optimize short-horizon returns under $\hat{P}$, synthesizing diverse high-return behaviors without costly real interaction Chua et al. (2018). Epistemic uncertainty–estimated by ensemble predictive variance–serves as a proxy for model error. Rather than correcting this error only during learning (e.g., MOPO Yu et al. (2020)), WOMBET *preemptively* removes unreliable trajectories by a dual-criterion filter: a trajectory τ is accepted and appended to $\mathcal{D}_S$ iff

$$\bar{u}(\tau) := \frac{1}{H}\sum_{t=0}^{H-1} u(s_t, a_t) \leq u_{\text{th}}, \qquad J(\tau) := \sum_{t=0}^{H-1} \gamma^t r_S(s_t, a_t) \geq J_{\text{th}}, \tag{6}$$

where u_{th} and J_{th} are user-specified thresholds controlling uncertainty and return levels, respectively. Accepted rollouts are relabeled with the target reward $r_T(s, a)$ to form $\mathcal{D}_S$. This step (i) suppresses bias by excluding high-uncertainty regions and (ii) yields a compact, high-value dataset that accelerates transfer. As online data accumulates, the model is refined and filtering is repeated, gradually expanding $\mathcal{D}_S$ toward a broader yet reliable state-action region.

4.4. Mixed Data Training

Given the constructed dataset $\mathcal{D}_S$ and online data $\mathcal{D}_T$, WOMBET performs policy optimization using a mixture of offline and online experience.

Let $\mathcal{D}_S$ and $\mathcal{D}_T$ denote offline and online datasets collected in the source and the target tasks respectively. Batches are sampled from $\mathcal{D}_{\text{mix}} = \alpha\mathcal{D}_S + (1-\alpha)\mathcal{D}_T$, where $\alpha \in [0, 1]$ adaptively balances bias and variance. An ensemble critic $\{Q_{\phi_i}\}_{i=1}^{N}$ is trained with a robust Bellman target:

$$\mathcal{L}_{\text{critic}}^{(i)} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}_{\text{mix}}}\left[\left(Q_{\phi_i}(s, a) - \mathcal{T}^{\pi}Q_{\bar{\phi}}(s, a)\right)^2\right], \tag{7}$$

where $\mathcal{T}^{\pi}Q(s, a) = r_T(s, a) - \mathbb{I}_{s\in\mathcal{D}_S}\lambda u(s, a) + \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[\min_i Q_{\bar{\phi}_i}(s', a')]$. Here $u(s, a)$ estimates epistemic uncertainty via a model ensemble Chua et al. (2018), and $\lambda > 0$ scales the uncertainty penalty, inducing a robust lower bound on the return. LayerNorm Ba et al. (2016) stabilizes the critic, and the actor maximizes the conservative value: $\mathcal{L}_{\text{actor}} = -\mathbb{E}_{s\sim\mathcal{D}_{\text{mix}}}[\min_i Q_{\phi_i}(s, \pi_\phi(s))]$.

4.5. Bounding Extrapolation Error via Implicit Regularization

When the policy π starts interacting with the target environment, it encounters (s, a) pairs that lie outside the support of the offline dataset $\mathcal{D}_S$ Lee et al. (2020); Park et al. (2024). On such OOD inputs, function approximators often overestimate Q-values–*extrapolation error*–causing value explosion and unstable updates. WOMBET mitigates this not by explicit penalties but through *implicit regularization* inherited from RL with prior data (RLPD) Ball et al. (2023), a SAC-based algorithm that incorporates LayerNorm, ensuring conservative and stable value estimates through two complementary mechanisms: architectural bounding and algorithmic robustness. Each critic employs LayerNorm across hidden layers to constrain activations and thereby bound Q-value magnitudes. For a critic $Q_\phi(s, a) = w^\top \text{ReLU}(\psi_\theta(s, a))$, where $\psi_\theta(s, a)$ is LayerNorm-activated,

$$\|Q_\phi(s, a)\| = \|w^\top \text{ReLU}(\psi_\theta(s, a))\| \leq \|w\|\,\|\psi_\theta(s, a)\|. \tag{8}$$

Since LayerNorm normalizes $\|\psi_\theta(s, a)\|$ across inputs, $\|Q_\phi(s, a)\|$ remains bounded by $\|w\|$, even for OOD samples. This architectural constraint prevents value explosion and stabilizes gradient propagation during learning. Complementing this bound, WOMBET uses an ensemble critic

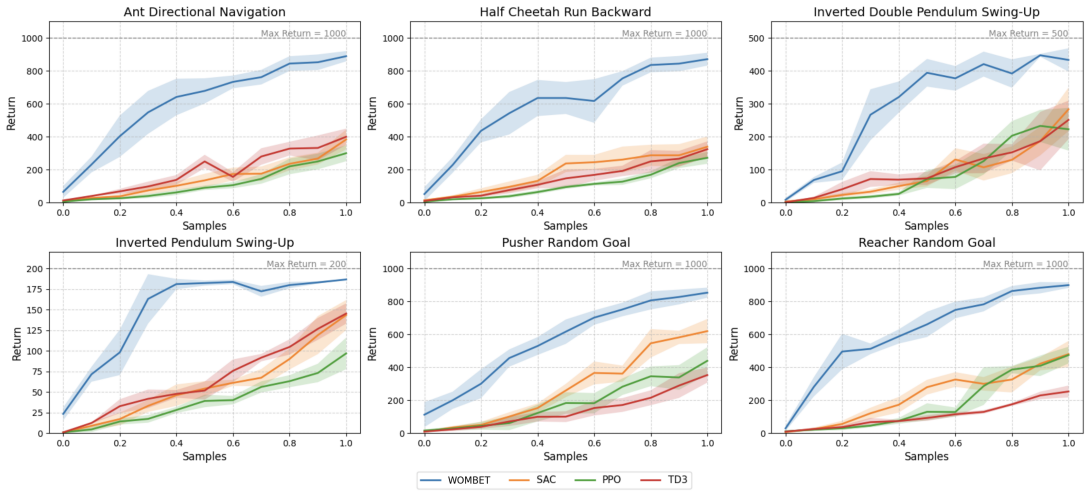

Figure 2: Sample efficiency of WOMBET vs. online RL baselines across target tasks.

$\{Q_{\phi_i}\}_{i=1}^{N}$ and computes Bellman targets using the ensemble minimum:

$$y(r, s') = r + \gamma \min_i Q_{\bar{\phi}_i}(s', a'), \qquad a' \sim \pi(\cdot|s'). \tag{9}$$

Taking the ensemble minimum reduces overestimation, particularly where critic disagreement indicates uncertainty. This ensemble-induced pessimism acts as an implicit regularizer, keeping extrapolated values conservative without extra penalty terms. Together, these regularizations maintain bounded and stable value estimates as WOMBET transitions from offline to online learning.

4.6. Mitigating Distributional Shift with Adaptive Sampling

While implicit regularization limits local extrapolation error, a more fundamental issue arises from the *global* shift in the state-action distribution as the policy evolves. As π_k adapts, its visitation distribution d_{π_k} diverges from the offline distribution $d_{\mathcal{D}_S}$, which biases value estimation. WOMBET mitigates this through *adaptive sampling*, dynamically balancing bias and variance in critic updates. At iteration k, critic updates draw samples from a mixture distribution $d_{\text{mix}}^{(k)} = \alpha_k d_{\mathcal{D}_S} + (1-\alpha_k) d_{\mathcal{D}_T^{(k)}}$, where $\mathcal{D}_T^{(k)}$ is the online replay buffer and $\alpha_k \in [0,1]$ controls the contribution of offline and online data. A larger α_k stabilizes training using reliable offline data, while a smaller α_k reduces bias by emphasizing online experience. Let $\hat{Q}$ denote the learned critic that approximates the true value function Q^{π_k}. The trade-off follows the domain adaptation bound $\mathbb{E}_{d_{\pi_k}}\big[|\hat{Q} - Q^{\pi_k}|\big] \leq \mathbb{E}_{d_{\text{mix}}^{(k)}}\big[|\hat{Q} - Q^{\pi_k}|\big] + L\, W_1(d_{\pi_k}, d_{\text{mix}}^{(k)})$, where L is the Lipschitz constant of the pointwise error map $f(s,a) = |\hat{Q}(s,a) - Q^{\pi_k}(s,a)|$ with respect to (s,a). The optimal mixture coefficient α_k^* minimizes this upper bound, defined as $\alpha_k^* = \arg\min_{\alpha\in[0,1]}\{\epsilon_{\text{approx}}\big(\mathcal{D}_{\text{mix}}(\alpha)\big) + L\, W_1\big(d_{\pi_k}, \mathcal{D}_{\text{mix}}(\alpha)\big)\}$, where $\epsilon_{\text{approx}}(\mathcal{D}) := \mathbb{E}_{d_{\mathcal{D}}}\big[\,|\hat{Q} - Q^{\pi_k}|\,\big]$ is the critic's approximation error under dataset/distribution $\mathcal{D}$, and $\mathcal{D}_{\text{mix}}(\alpha)$ denotes the mixture dataset induced by $d_{\text{mix}}^{(k)} = \alpha\, d_{\mathcal{D}_S} + (1-\alpha)\, d_{\mathcal{D}_T^{(k)}}$.

WOMBET measures critic reliability using the mean absolute TD error on recent online data. When the TD error is large (implies underfitting), more offline samples are used to stabilize updates. As the TD error decreases, sampling shifts toward online data to improve adaptation. At iteration k, the TD error is $\delta_T^{(k)} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{B}_T}\big[|Q_\phi(s,a) - y(r,s')|\big]$, where $\mathcal{B}_T$ denotes the target task replay buffer. It is smoothed via exponential averaging $\bar{\delta}_T^{(k)} = (1-\beta_{\text{ema}})\bar{\delta}_T^{(k-1)} + \beta_{\text{ema}}\delta_T^{(k)}$, and used

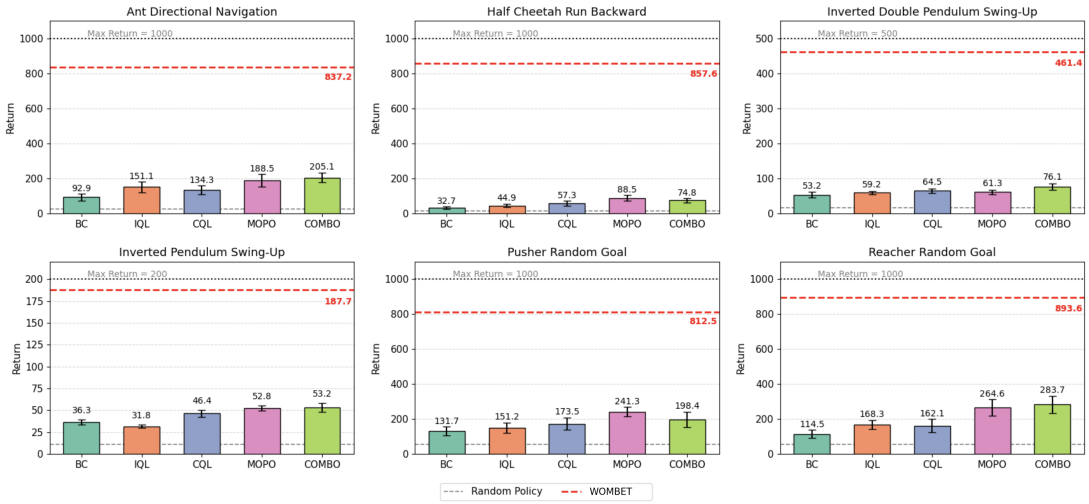

Figure 3: Comparison of offline RL baselines and WOMBET across target tasks. Bars show offline performance on $\mathcal{D}_S$ and the red dashed line denotes WOMBET's post-adaptation return.

to update the mixing ratio $\alpha_k = \text{clip}\big(\lambda_{\text{gain}}\bar{\delta}_T^{(k)}, \alpha_{\min}, \alpha_{\max}\big)$. This rule approximates the gradient of the error bound with respect to α_k, adjusting data composition based on critic uncertainty. As learning progresses, $\bar{\delta}_T^{(k)}$ stabilizes and α_k converges to a balanced ratio, keeping WOMBET stable in early training and increasingly adaptive later, enabling robust and efficient transfer learning.

4.7. Robust Surrogate Optimization and Performance Guarantee

Building on the preceding analysis, we establish a unified performance bound for WOMBET. Its two-layer design–*uncertainty-penalized offline data generation* and *adaptive fine-tuning*–ensures reliable offline initialization and stable online transfer. In the offline phase, MBRL with an uncertainty penalty produces conservative trajectories for $\mathcal{D}_S$, which are later relabeled with r_T to align with the target task. The online phase is entirely model-free and optimizes unpenalized target returns from real interactions. The remaining suboptimality arises from how accurately the learned critic $\hat{Q}$ approximates the true target value Q_T^π during fine-tuning.

The total estimation error can be decomposed into two sources:

$$\sup_{(s,a)} |Q_T^\pi(s,a) - \hat{Q}(s,a)| \leq \underbrace{|Q_T^\pi - Q_{\text{mix}}^\pi|}_{\text{(a) Distribution mismatch}} + \underbrace{|Q_{\text{mix}}^\pi - \hat{Q}|}_{\text{(b) Finite-sample approximation}}, \tag{10}$$

where Q_{mix}^π denotes the ideal action-value under the mixed visitation distribution $d_{\text{mix}} = \alpha\, d_{\mathcal{D}_S} + (1-\alpha)\, d_{\mathcal{D}_T}$. WOMBET minimizes both components through its model-based data generation and adaptive sampling mechanisms.

Given the mixed dataset $\mathcal{D}_{\text{mix}}$ and critic class $\mathcal{F}$, a PAC-style Agarwal et al. (2019) bound gives

$$|Q_{\text{mix}}^\pi(s,a) - \hat{Q}(s,a)| \leq \mathcal{O}\left(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{|\mathcal{D}_{\text{mix}}|}}\right), \tag{11}$$

where $r_{\max} := \sup_{s,a} |r_T(s,a)|$, $\gamma \in (0,1)$ is the discount factor, and $\delta \in (0,1)$ is the confidence parameter. Intuitively, PAC bounds state that with probability at least $1-\delta$, the critic's generalization error decreases as the dataset size grows. WOMBET mitigates this term via (i) *adaptive sampling*, which balances bias and variance, and (ii) *implicit regularization* (LayerNorm and ensemble-min targets), which improve generalization and prevent value divergence.

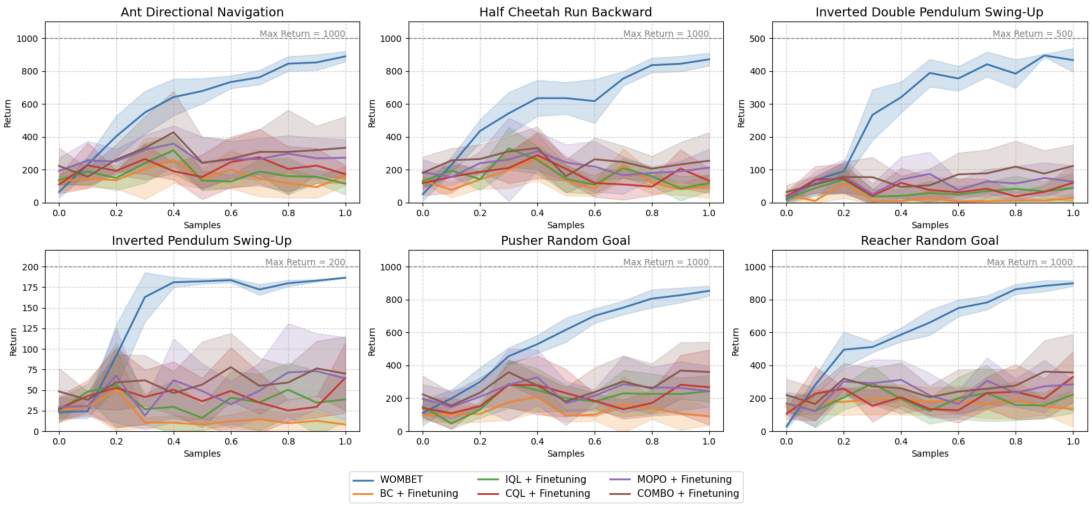

Figure 4: WOMBET vs. offline-to-online baselines (offline RL with fine-tuning) across target tasks.

Deviation between the mixed-data critic and the true target critic satisfies

$$|Q_T^\pi(s,a) - Q_{\text{mix}}^\pi(s,a)| \leq \tfrac{\gamma}{1-\gamma} L_V \Delta_P, \tag{12}$$

where $\Delta_P := \sup_{s,a} W_1\big(P_T(\cdot|s,a), P_S(\cdot|s,a)\big)$ quantifies the dynamics discrepancy between target and source, and L_V is the Lipschitz constant of the value function V_P^π with respect to state. WOMBET reduces this term through *dual-criterion filtering*, which constrains $\mathcal{D}_S$ to high-return, low-uncertainty trajectories. As online data accumulates and the mixing weight α_k decreases, the influence of mismatched source samples further diminishes.

Combining both bounds, WOMBET optimizes a robust surrogate objective while explicitly controlling approximation and distributional errors. The resulting policy satisfies

$$J_T(\pi_{\text{WOMBET}}) \geq \tilde{J}_T(\pi_{\text{WOMBET}}) - \mathcal{O}\left(\tfrac{\gamma}{1-\gamma} L_V \Delta_P + \tfrac{r_{\max}}{1-\gamma} \sqrt{\tfrac{\ln(|\mathcal{F}|/\delta)}{|\mathcal{D}_{\text{mix}}|}} \right), \tag{13}$$

showing that WOMBET achieves conservative yet asymptotically consistent transfer by coupling model-based uncertainty-aware data generation with model-free fine-tuning.

5. Experiments

We evaluate WOMBET on diverse MuJoCo benchmarks to examine: (1) **sample efficiency**: whether experience transfer accelerates learning over online RL trained from scratch, (2) **necessity of fine-tuning**: whether transferred data alone can solve the target task without fine-tuning, (3) **transfer effectiveness**: how WOMBET compares with existing offline-to-online baselines, and (4) **component contribution**: the impact of dual-criterion filtering and adaptive sampling. We measure normalized return and shaded regions indicate standard deviation across 5 seeds.

5.1. Sample Efficiency: WOMBET vs. Online RL

We evaluate whether WOMBET improves sample efficiency over standard online RL. We compare against SAC, PPO, and TD3 trained from scratch on the target task $\mathcal{M}_T$ Schulman et al. (2017); Haarnoja et al. (2018); Fujimoto et al. (2018). As shown in Figure 2, WOMBET learns faster and achieves higher asymptotic returns. The filtered dataset $\mathcal{D}_S$ provides a strong prior, allowing

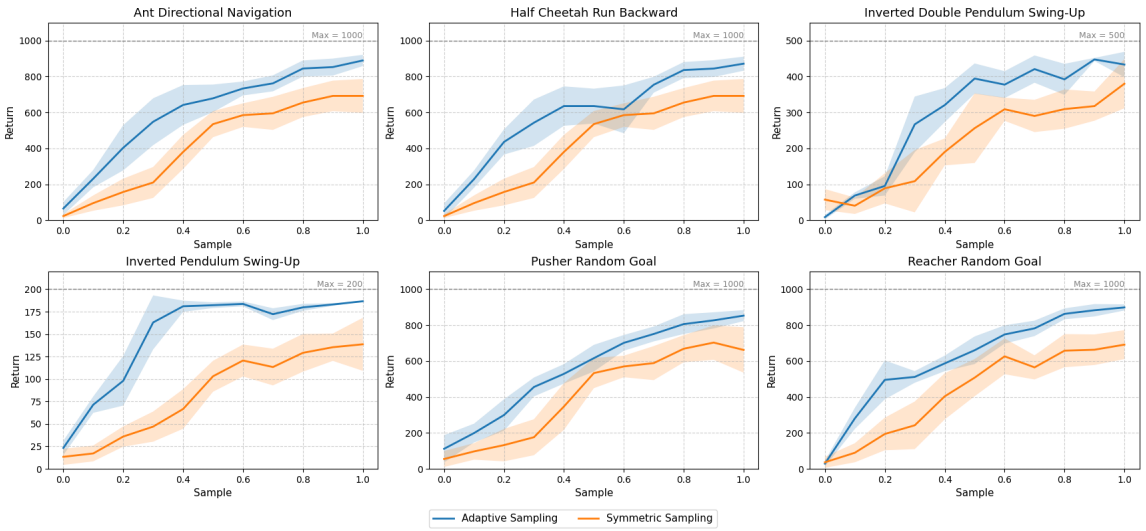

Figure 5: Comparison between WOMBET with adaptive sampling (blue) and a symmetric fixed-ratio variant (orange) across target tasks.

WOMBET to focus on refinement rather than initial exploration. This gain comes from uncertainty-aware planning and filtering, which produce reliable and transferable experience. In contrast, SAC, PPO, and TD3 rely on uninformed exploration, resulting in slower convergence. Across six environments, WOMBET attains comparable or better final returns with less than half the interaction budget, demonstrating improved sample efficiency.

5.2. Necessity of Online Fine-tuning: WOMBET vs. Offline RL

We evaluate the necessity of online fine-tuning by comparing WOMBET to BC, IQL, CQL, MOPO, and COMBO, each trained only on the model-generated dataset $\mathcal{D}_S$ Kostrikov et al. (2021); Kumar et al. (2020); Yu et al. (2020, 2021). These methods are evaluated on the target task $\mathcal{M}_T$ without further interaction, forming a zero-shot transfer setting. As shown in Figure 3, offline-only methods perform well when $\mathcal{M}_S$ and $\mathcal{M}_T$ are closely aligned but degrade under moderate shifts. Although $\mathcal{D}_S$ is reliable, it reflects source dynamics and rewards rather than those of $\mathcal{M}_T$. Without online updates, policies cannot adapt to new rewards or unseen state-action regions. WOMBET addresses this by refining both policy and model through interaction, enabling adaptation beyond the offline data support. The consistent gap shows that gains come not only from pretraining but from effectively leveraging prior data for exploration and refinement.

5.3. Effectiveness of Experience Transfer: WOMBET vs. Offline-to-Online Baselines

We next compare WOMBET with state-of-the-art offline-to-online RL methods to evaluate its transfer effectiveness. Each baseline is pretrained offline on the same model-generated dataset $\mathcal{D}_S$ and then fine-tuned online in the target environment $\mathcal{M}_T$. Note that all methods are initialized with the same dataset $\mathcal{D}_S$ generated by WOMBET, which isolates the effect of the transfer mechanism while controlling for data construction. As shown in Figure 4, WOMBET consistently achieves higher sample efficiency and asymptotic return than all baselines. This improvement stems from its integrated transfer design: dual-criterion filtering provides reliable initialization, and adaptive sampling enables a smooth offline-to-online transition, maintaining a balanced bias-variance trade-off from the outset. In contrast, conventional pretrain-finetune pipelines face abrupt distributional

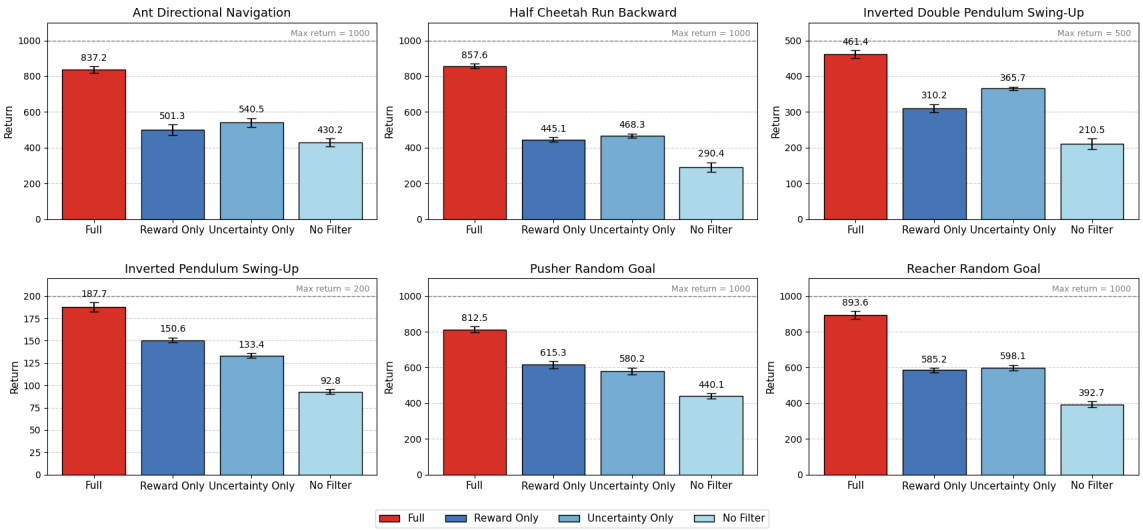

Figure 6: Ablation of WOMBET's dual-criterion filter across target tasks.

shifts–their offline-trained critics and actors often fail to adapt to online data, causing unstable or inefficient exploration. By coupling model-based pretraining with robust online transfer, WOMBET avoids this mismatch and achieves stable, data-efficient policy improvement, confirming its ability to transfer model-based experience into high-performing online control.

5.4. Importance of WOMBET's Core Components for Offline Data Generation

We evaluate whether WOMBET's performance arises from its key components: adaptive sampling and dual-criterion filtering. For adaptive sampling, we compare the full method–where α_k is adjusted using the critic's TD error–to a fixed-ratio baseline ($\alpha_k = 0.5$). As shown in Figure 5, the fixed schedule is stable but learns slower and reaches lower returns. The adaptive rule shifts weight from offline to on-policy data, maintaining a bias-variance balance during training. For dual-criterion filtering, we compare the full rule (high reward and low uncertainty) with reward-only, uncertainty-only, and unfiltered variants (Figure 6). All degraded variants reduce performance: reward-only includes unreliable samples, uncertainty-only is overly conservative, and no filtering amplifies model bias. The dual-criterion filter yields trajectories that are both reliable and task-relevant. Across tasks, these components are complementary and necessary for stable transfer and data-efficient fine-tuning.

6. Discussion and Conclusion

WOMBET presents a unified framework for offline-to-online RL that combines conservative model-based data generation with model-free fine-tuning. Unlike prior methods that assume a fixed offline dataset, WOMBET constructs transferable experience from a source task. It addresses a key limitation of MBRL–policy exploitation of model errors–by integrating uncertainty-aware data generation with adaptive policy updates. In the source task, uncertainty-penalized MPC with dual-criterion filtering produces reliable, high-return trajectories that form the offline dataset. In the target task, model-free learning with adaptive sampling balances stability from offline data and exploration through interaction. Regularized updates and normalization further stabilize value estimation. Overall, WOMBET couples conservative data generation with adaptive learning, mitigating model bias, improving sample efficiency, and enabling robust transfer across continuous control tasks.

Acknowledgments

This work was supported in part by NSF CMMI-2140650 and in part by Design of Robustly Implementable Autonomous and Intelligent Machines, DARPA award number HR00112490425.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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.

Jiaze Cai, Vishnu Sangli, Mintae Kim, and Koushil Sreenath. Learning-based trajectory tracking for bird-inspired flapping-wing robots, 2024. URL https://arxiv.org/abs/2411.15130.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Gilbert Feng, Hongbo Zhang, Zhongyu Li, Xue Bin Peng, Bhuvan Basireddy, Linzhu Yue, Zhitao Song, Lizhi Yang, Yunhui Liu, Koushil Sreenath, et al. Genloco: Generalized locomotion controllers for quadrupedal robots. In *Conference on Robot Learning*, pages 1893–1903. PMLR, 2023.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1582–1591, 2018.

Bibek Gupta, Mintae Kim, Albert Park, Eric Sihite, Koushil Sreenath, and Alireza Ramezani. Estimation of aerodynamics forces in dynamic morphing wing flight. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7210–7215. IEEE, 2025.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 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.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Mintae Kim. Finite memory belief approximation for optimal control in partially observable markov decision processes. *arXiv preprint arXiv:2601.03132*, 2026.

Mintae Kim and Koushil Sreenath. Robust adversarial policy optimization under dynamics uncertainty, 2026. URL https://arxiv.org/abs/2604.10974.

Mintae Kim, Jiaze Cai, and Koushil Sreenath. Roverfly: Robust and versatile implicit hybrid control of quadrotor-payload systems. *arXiv preprint arXiv:2509.11149v2*, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

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.

Albert K Lee and Matthew A Wilson. Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194, 2002.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Addressing distribution shift in online reinforcement learning with offline datasets. 2020.

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.

Zhongyu Li, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, 0(0):02783649241285161, 2025. doi: 10.1177/02783649241285161. URL https://doi.org/10.1177/02783649241285161.

Yihuan Mao, Chao Wang, Bin Wang, and Chongjie Zhang. Moore: Model-based offline-to-online reinforcement learning. *arXiv preprint arXiv:2201.10070*, 2022.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

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

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline rl? *Advances in Neural Information Processing Systems*, 37:79029–79056, 2024.

Ilija Radosavovic, Tete Xiao, Bike Zhang, Trevor Darrell, Jitendra Malik, and Koushil Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89): eadi9579, 2024.

Rafael Rafailov, Kyle Beltran Hatch, Victor Kolev, John D Martin, Mariano Phielipp, and Chelsea Finn. Moto: Offline to online fine-tuning for model-based reinforcement learning. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Avi Singh, Albert Yu, Jonathan Yang, Jesse Zhang, Aviral Kumar, and Sergey Levine. Cog: Connecting new skills to past experience with offline reinforcement learning. *arXiv preprint arXiv:2010.14500*, 2020.

Laura Smith, J Chase Kew, Tianyu Li, Linda Luu, Xue Bin Peng, Sehoon Ha, Jie Tan, and Sergey Levine. Learning and adapting agile locomotion skills by transferring experience. *arXiv preprint arXiv:2304.09834*, 2023.

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.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Tingwu Wang and Jimmy Ba. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*, 2019.

Matthew A Wilson and Bruce L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.

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

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. *arXiv preprint arXiv:2412.07762*, 2024.

Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. In *ICLR*, 2025. arXiv preprint arXiv:2412.07762.