BAYESIAN OFFLINE MULTI-AGENT REINFORCEMENT LEARNING WITH LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-based offline multi-agent reinforcement learning (MARL) provides a promising approach to multi-agent decision-making under partial observability, where the global state is unavailable from offline data. However, limited dataset coverage can lead to substantial out-of-distribution (OOD) model errors, which may mislead policy optimization and degrade downstream performance. Existing model-based offline MARL methods typically rely on access to global states or mitigate model errors through pessimistic or conservative policy optimization, leaving the reliability and refinement of learned world models less explored. We propose a Bayesian offline MARL framework that infers latent global states from decentralized observation-action histories and iteratively refines a posterior world model using only offline data. Building on posterior information loss, we extend this criterion to the multi-agent setting as multi-agent posterior information loss (MPIL), which provides a reliability measure for world model refinement and offline evaluation. Theoretically, we establish an upper bound on policy regret that is monotonic in MPIL and further characterize its dependence on the size and state–joint-action coverage of the offline dataset. Experiments on partially observable multi-agent variants of D4RL MuJoCo tasks demonstrate competitive policy performance across diverse offline data conditions, while MPIL-guided refinement improves world model prediction and downstream policy performance. These results support a principled connection between world model reliability, offline data coverage, and policy performance in partially observable offline MARL.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has achieved remarkable success in complex cooperative tasks (Lowe et al., 2017). However, conventional MARL typically requires extensive online interaction during training, which can incur substantial safety risks and trial-and-error costs when a sufficiently accurate simulator is unavailable. Offline MARL addresses this limitation by learning policies exclusively from historical datasets and has been widely studied in fields such as healthcare, industrial automation, and autonomous driving (Emerson et al., 2023; Ma et al., 2024; Hu et al., 2025; Lee & Kwon, 2025). Nevertheless, in many practical partially observable multi-agent systems, the global state is inaccessible, leaving agents with only decentralized local observations. This information deficiency undermines centralized policy learning, since local observations may be insufficient to characterize the underlying environment and support reliable coordination (Xu et al., 2022). Therefore, how to learn effective policies solely from offline data when global states are unavailable remains a fundamental challenge in offline MARL.

Model-based offline RL offers a natural paradigm for fully offline policy learning and evaluation, and its core ideas have also been extended to multi-agent settings. By learning environment dynamics from offline data, a world model can serve as a surrogate environment for policy optimization and evaluation. However, limited data coverage can make the learned model unreliable, causing policy optimization to exploit model errors in out-of-distribution (OOD) regions. Existing methods mainly address this issue through pessimistic or conservative policy optimization (Yu et al., 2020; Kidambi et al., 2020; Yu et al., 2021; Barde et al., 2024). While effective, these methods mainly mitigate the impact of model errors during policy optimization, rather than improving the world model itself; moreover, optimizing policies against an unreliable model can incur substantial and avoidable computational cost. Moreover, offline RL lacks a principled criterion linking world model reliability to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

policy performance and safety. SOReL (Fellows et al., 2026) takes an important step by introducing posterior information loss (PIL), which links world model discrepancy to policy regret, although it is limited to the single-agent setting and does not employ PIL for world model refinement.

To address these challenges, we propose a latent world model framework for fully offline MARL under decentralized partial observability. The framework infers latent global states from agents' local observation-action histories and models posterior uncertainty over transition, reward, and observation dynamics. Building on this model, we introduce multi-agent posterior information loss (MPIL) to quantify world model reliability and further integrate empirical MPIL into world model learning as a fine-tuning objective and reliability criterion. Starting from an initially trained world model, the framework iteratively refines the model until convergence, after which the resulting world model is used for policy learning and fully offline regret estimation, forming a closed loop of world model refinement, policy learning, and performance assessment without further interaction with the true environment.

Within this framework, we prove that the true regret is upper-bounded by a monotonic function of MPIL, and further characterize its dependence on offline dataset size and state–joint-action coverage. These results establish a unified theoretical connection among world model reliability, offline data availability, and policy-performance guarantees. We validate the proposed theory and framework on partially observable multi-agent MuJoCo (Schroeder de Witt et al., 2020) benchmarks, where MPIL-guided world model refinement improves model prediction accuracy and policy performance, and the regret estimated from the refined world model follows a trend consistent with the true regret as the amount of offline data increases.

2 PRELIMINARIES

The decentralized partially observable Markov decision process (Dec-POMDP) provides a general framework for modeling cooperative multi-agent decision-making problems in which each agent has access only to local observations, while all agents share a common system-level objective (Bernstein et al., 2002; Bellman, 1956). A Dec-POMDP is formally defined by the tuple $G = \langle \mathcal{I}, S, \{A_i\}, T, R, \{Z_i\}, \{O_i\}, \gamma \rangle$. Here, $\mathcal{I} = \{1, 2, \ldots, I\}$ denotes the finite set of agents, and S denotes the environment state space with an initial state distribution P_0. At each time step $t = 0, 1, 2, \ldots$, the environment occupies a global state $s_t \in S$, and each agent $i \in \mathcal{I}$ selects an action $a_{i,t}$ from its action space A_i according to its policy π_i. The individual actions form the joint action $\tilde{a}_t = (a_{1,t}, \ldots, a_{I,t})$. The transition function $T(s_{t+1}|s_t, \tilde{a}_t) = \Pr(s_{t+1} \mid s_t, \tilde{a}_t)$ characterizes the stochastic evolution of the environment. The reward function is defined as $R(r_t|s_t, \tilde{a}_t) = \Pr(r_t \mid s_t, \tilde{a}_t)$, where the resulting global reward is shared among all agents. Under partial observability, each agent i receives a local observation $z_{i,t}$ from its observation space Z_i rather than directly accessing the global state s_t, where the observation function $O_i(z_{i,t+1}|s_t, \tilde{a}_t) = \Pr(z_{i,t+1} \mid s_t, \tilde{a}_t)$ characterizes the generation of local observations. The policy π_i maps the local history $h_{i,t} = (z_{i,0}, a_{i,0}, \ldots, a_{i,t-1}, z_{i,t})$ to an action distribution $\pi_i(a_{i,t} \mid h_{i,t})$, and the joint policy is denoted by $\tilde{\pi} = (\pi_1, \ldots, \pi_I)$. The discount factor $\gamma \in [0, 1)$ determines the relative importance of future rewards.

To model partially observable multi-agent environments, we construct a dynamics model $D_{T,R}$ that jointly captures state transitions and shared rewards, together with an observation model $\tilde{O} = \{O_i\}_{i \in \mathcal{I}}$ for local-observation generation. The resulting world model is defined as $\mathcal{M}_{\mathrm{WM}} = (D_{T,R}, \tilde{O})$. For model learning, T and R are jointly parameterized through a transition-reward distribution $P_{T,R}$, whereas the observation distribution are represented separately by $P_{\tilde{O}}$. In this work, the true global state s_t is unavailable during both training and execution, and the offline dataset therefore contains only local observations, joint actions, and shared rewards, $\mathcal{D}_N = \{(\tilde{z}_t, \tilde{a}_t, r_t, \tilde{z}_{t+1})\}_{t=0}^{N-1}$, where N is the number of transitions in the offline dataset.

3 RELATED WORK

Existing model-free offline RL methods mitigate distribution shift through policy constraints, conservative value estimation, and pessimistic learning (Kumar et al., 2020; Fujimoto et al., 2019; Kumar et al., 2019; Fujimoto & Gu, 2021; Bai et al., 2022), with multi-agent extensions further ad-

dressing coordination through CTDE, decentralized optimization, and global-to-local regularization (Yang et al., 2021; Pan et al., 2022; Wang et al., 2023; Shao et al., 2023). However, these methods do not explicitly model environment dynamics, limiting their ability to explicitly reason about long-horizon consequences through learned environment dynamics. Model-based offline RL provides an alternative research direction by learning environment dynamics and leveraging imagined trajectories for policy optimization. Existing methods mainly mitigate the impact of model errors on policy optimization through pessimistic modeling (Yu et al., 2020; Kidambi et al., 2020; Wu et al., 2024), uncertainty estimation (Qiao et al., 2025; Sun et al., 2023), and conservative learning (Yu et al., 2021). Most of these methods assume that the global state required for dynamics learning is directly available, including recent model-based extensions to offline MARL (Barde et al., 2024). In partially observable multi-agent environments, however, the global state may not be directly accessible from offline data. A common workaround is to concatenate agents' local observations as a surrogate global representation (Lobos-Tsunekawa et al., 2022; Barde et al., 2024), while another line of work learns latent state representations from observation-action histories to capture hidden system dynamics (Gregor et al., 2019; Venugopal et al., 2023; Shi et al., 2025). Although these approaches enable dynamics modeling without direct access to the true global state, how to assess and refine the learned world model solely from fixed offline data remains underexplored.

Fully offline policy assessment has been studied from several perspectives. Off-policy evaluation (OPE) estimates target-policy performance from logged data using importance weighting, value-function estimation, or learned dynamics models (Thomas & Brunskill, 2016a; Fu et al., 2021; Uehara et al., 2026), while offline policy-selection methods use such estimates or alternative validation criteria to rank candidate policies and tune hyperparameters without environment interaction (Zhang & Jiang, 2021; Da et al., 2024). These approaches primarily assess or select policies and value functions, rather than characterizing whether a learned world model is sufficiently reliable for downstream policy learning. More recently, SOReL (Fellows et al., 2026) introduces PIL to connect posterior world model discrepancy with policy regret, enabling fully offline model assessment and tuning. However, current PIL-based approaches are mainly developed for single-agent scenarios and do not explicitly leverage PIL to update world model parameters.

In this work, we develop an MPIL-guided latent world model framework for partially observable offline MARL, which infers latent system states from local observation-action histories without access to global states. MPIL is used to assess and iteratively refine the world model fully offline, while its connection to policy regret enables principled offline evaluation of multi-agent policy performance.

4 METHODOLOGY

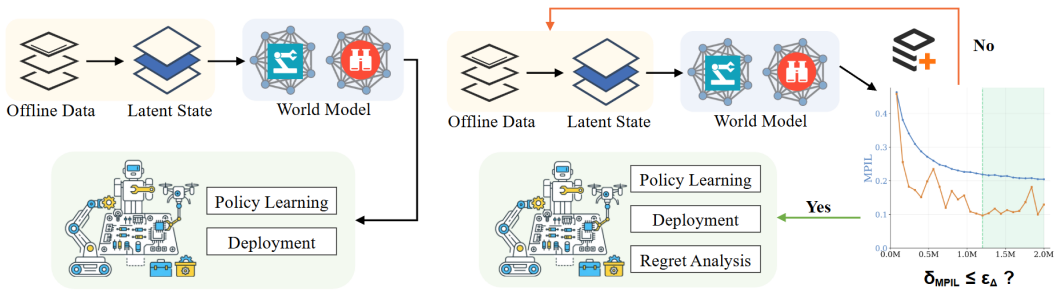
In this section, we present a Bayesian world model-based offline MARL framework for partially observable environments, as illustrated in Fig. 1. To obtain a system-level representation for world model learning and cooperative policy optimization, the proposed framework first infers a latent global state $\hat{s}_t$ from offline local observation-action histories and then learns a Bayesian world model. The world model is further optimized using an empirical multi-agent posterior information loss (MPIL) objective and assessed using an MPIL-based reliability criterion. Once the criterion is satisfied, the world model is used for cooperative policy optimization under the CTDE paradigm (Lowe et al., 2017; Rashid et al., 2020).

4.1 LATENT GLOBAL STATE INFERENCE

Although concatenating local observations provides a straightforward approximation of global information, the resulting representation can become high-dimensional as the number of agents increases and may still omit information that is unobserved by all agents (Yu et al., 2022). We therefore infer a compact history-dependent latent global state from joint observation-action histories (Venugopal et al., 2023). Let $\tilde{z}_t$ and $\tilde{a}_t$ denote the joint observation and joint action of all agents at time step t, respectively. A gated recurrent unit (GRU) is employed as a recurrent history encoder to infer the latent state $\hat{s}_t$ (Igl et al., 2018),

$$\hat{s}_t = \mathrm{GRU}_\xi\left(\hat{s}_{t-1}, [\tilde{z}_t, \tilde{a}_{t-1}]\right), \tag{1}$$

where ξ denotes the trainable parameters of the GRU encoder.



(a) Traditional Model-Based Offline MARL (b) MPIL-Guided Model-Based Offline MARL

Figure 1: Comparison of Traditional and MPIL-Guided Model-Based Offline MARL Frameworks.

To regularize the inferred latent representation and encourage it to preserve observation-relevant information, we introduce an auxiliary observation decoder that reconstructs the joint observation from the latent state $\check{z}_t = D_\psi(\hat{s}_t)$, where ψ denotes the parameters of the auxiliary decoder. The encoder and auxiliary decoder are jointly trained by minimizing the observation reconstruction loss,

$$\mathcal{L}_{\text{aux}}(\xi, \psi) = \frac{1}{N} \sum_{t=0}^{N-1} \|\tilde{z}_t - \check{z}_t\|_2^2 . \tag{2}$$

After representation learning, the offline data are mapped into the latent state space. The resulting state-augmented dataset is $\hat{\mathcal{D}}_N = \{(\hat{s}_t, \tilde{z}_t, \tilde{a}_t, r_t, \hat{s}_{t+1}, \tilde{z}_{t+1})\}_{t=0}^{N-1}$.

4.2 World Model Formulation

In this subsection, we adopt a Bayesian formulation for the transition-reward parameters θ and observation parameters $\tilde{\omega}$ (Folch et al., 2024; Duff, 2002).

Based on the inferred latent global state, the parameterized transition-reward model is defined as $P_{T,R} = p(\hat{s}_{t+1}, r_t \mid \hat{s}_t, \tilde{a}_t, \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^d$. For each agent i, the observation model is defined as $P_{O_i} = p(z_{i,t+1} \mid \hat{s}_t, \tilde{a}_t, \omega_i)$, where $\omega_i \in \Omega_i \subseteq \mathbb{R}^d$. We assume conditional independence of local observations given $(\hat{s}_t, \tilde{a}_t)$, yielding $P_{\tilde{O}} = p(\tilde{z}_{t+1} \mid \hat{s}_t, \tilde{a}_t, \tilde{\omega}) = \prod_{i=1}^{I} p(z_{i,t+1} \mid \hat{s}_t, \tilde{a}_t, \omega_i)$. We collect $\tilde{\omega} = (\omega_1, \ldots, \omega_I) \in \tilde{\Omega}$, where $\tilde{\Omega} = \Omega_1 \times \cdots \times \Omega_I$, and use the factorized prior $p(\tilde{\omega}) = \prod_{i=1}^{I} p(\omega_i)$. Prior distributions P_Θ and P_{Ω_i} encode the initial beliefs over the transition-reward and observation models, respectively. Given $\hat{\mathcal{D}}_N$, Bayes' rule (Ross et al., 2011) yields the joint posterior over the complete world model parameters,

$$p(\theta, \tilde{\omega} \mid \hat{\mathcal{D}}_N) \propto p(\theta) \prod_{i=1}^{I} p(\omega_i) \prod_{t=0}^{N-1} \left[p(\hat{s}_{t+1}, r_t \mid \hat{s}_t, \tilde{a}_t, \theta) \prod_{i=1}^{I} p(z_{i,t+1} \mid \hat{s}_t, \tilde{a}_t, \omega_i) \right]. \tag{3}$$

Together, these components define $P_{\Theta,\tilde{\Omega}}(\cdot \mid \hat{\mathcal{D}}_N)$, which characterizes posterior uncertainty over the complete world model. We train the world model using the corresponding posterior-averaged negative log-likelihood (NLL), defined as

$$\mathcal{L}_{\text{NLL}} = \mathbb{E}_{(\hat{s},\tilde{a}) \sim \nu_{\hat{\mathcal{D}}_N}} \mathbb{E}_{(\theta,\tilde{\omega}) \sim P_{\Theta,\tilde{\Omega}}(\cdot \mid \hat{\mathcal{D}}_N)} \Big[\mathbb{E}_{(s',r) \sim P^\star_{T,R}(\cdot \mid \hat{s},\tilde{a})} [-\log P_{T,R}(s', r \mid \hat{s}, \tilde{a}, \theta)] + \mathbb{E}_{\tilde{o}' \sim P^\star_{\tilde{O}}(\cdot \mid \hat{s},\tilde{a})} [-\log P_{\tilde{O}}(\tilde{o}' \mid \hat{s}, \tilde{a}, \tilde{\omega})] \Big]. \tag{4}$$

This objective evaluates how well posterior world model hypotheses assign likelihood to transitions generated by the true environment, and serves as the world model learning objective.

To evaluate world model reliability before policy learning, we extend posterior information loss (PIL) (Fellows et al., 2026) to the multi-agent setting and further reformulate it as a dataset-weighted criterion. Specifically, we define multi-agent posterior information loss (MPIL) as the posterior-averaged discrepancy between the true and learned world model distributions over state-action pairs

covered by the offline dataset:

$$\mathcal{I}_{\hat{\mathcal{D}}_N} = \mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot|\hat{\mathcal{D}}_N)} \mathbb{E}_{(\hat{s}_j,\tilde{a}_j)\sim\nu_{\hat{\mathcal{D}}_N}} \Big[D_{\mathrm{KL}}\left(P^\star_{T,R}(\cdot \mid \hat{s}_j,\tilde{a}_j) \| P_{T,R}(\cdot \mid \hat{s}_j,\tilde{a}_j,\theta)\right) + D_{\mathrm{KL}}\left(P^\star_{\tilde{O}}(\cdot \mid \hat{s}_j,\tilde{a}_j) \| P_{\tilde{O}}(\cdot \mid \hat{s}_j,\tilde{a}_j,\tilde{\omega})\right) \Big]. \quad (5)$$

Here, $P^\star_{T,R}$ and $P^\star_{\tilde{O}}$ denote the transition–reward and joint-observation distributions of the true environment, respectively. $\nu_{\hat{\mathcal{D}}_N}$ denotes the underlying state-action distribution associated with the offline dataset $\hat{\mathcal{D}}_N$; in practice, expectations with respect to $\nu_{\hat{\mathcal{D}}_N}$ are approximated by uniformly sampling transitions from the dataset. The following proposition establishes the theoretical connection between MPIL and the NLL objective.

Proposition 4.1. *For a fixed offline data distribution $\nu_{\hat{\mathcal{D}}_N}$,*

$$\mathcal{I}_{\hat{\mathcal{D}}_N} = \mathcal{L}_{\mathrm{NLL}} - \mathcal{H}^\star_{\hat{\mathcal{D}}_N}, \quad (6)$$

where $\mathcal{H}^\star_{\hat{\mathcal{D}}_N} = \mathbb{E}_{(\hat{s},\tilde{a})\sim\nu_{\hat{\mathcal{D}}_N}} \Big[H\big(P^\star_{T,R}(\cdot \mid \hat{s},\tilde{a})\big) + H\Big(P^\star_{\tilde{O}}(\cdot \mid \hat{s},\tilde{a})\Big) \Big]$ denotes the conditional entropy of the true environment under the offline data distribution.

The proof, based on the KL–cross-entropy identity, is provided in Appendix A.3.1. This relation provides a tractable way to estimate MPIL through the posterior-averaged NLL. Since $\mathcal{H}^\star_{\hat{\mathcal{D}}_N}$ is independent of the world model parameters, minimizing the posterior-averaged NLL is equivalent to minimizing MPIL, and NLL changes across refinement stages directly track changes in MPIL.

Exact posterior inference is generally intractable. We therefore employ an ensemble of K world models, $\{\mathcal{M}_k\}_{k=1}^K$, where $\mathcal{M}_k = (T_{\theta_k}, \tilde{O}_{\tilde{\omega}_k})$, to approximate the joint parameter posterior as

$$P_{\Theta,\tilde{\Omega}}\left(\cdot \mid \hat{\mathcal{D}}_N\right) \approx \widehat{P}_\Phi(\cdot) := \frac{1}{K}\sum_{k=1}^K \delta_{(\theta_k,\tilde{\omega}_k)}(\cdot), \quad (7)$$

where $\Phi = \{(\theta_k, \tilde{\omega}_k)\}_{k=1}^K$ denotes the collection of ensemble parameters. Each parameter pair represents a candidate world model hypothesis, and the ensemble members are independently initialized to encourage model diversity.

Under this ensemble approximation, the posterior expectation in Eq. (4) is replaced by an empirical average over the K world models. For an offline subset $\hat{\mathcal{D}}$, the empirical NLL is computed as

$$\widehat{\mathcal{L}}_{\mathrm{NLL}}(\Phi;\hat{\mathcal{D}}) = -\frac{1}{K|\hat{\mathcal{D}}|}\sum_{k=1}^K \sum_{\tau\in\hat{\mathcal{D}}} \Big[\log P_{T,R}(s'_i, r_i \mid \hat{s}_i,\tilde{a}_i,\theta_k) + \log P_{\tilde{O}}(\tilde{o}'_i \mid \hat{s}_i,\tilde{a}_i,\tilde{\omega}_k)\Big], \quad (8)$$

where $\tau = (\hat{s}_t, \tilde{a}_t, \hat{s}_{t+1}, r_t, \tilde{o}_{t+1})$ denotes an offline transition in $\hat{\mathcal{D}}$. We use $\widehat{\mathcal{L}}_{\mathrm{NLL}}$ as the practical world model training loss and as an empirical surrogate for monitoring MPIL during iterative world model refinement.

4.3 Learning an Approximately Bayes-Optimal Policy

We denote the parameters of the initially learned world model ensemble by $\Phi^{(1)}$. We assess world model convergence using the change in MPIL between consecutive refinement stages. A fixed held-out validation set $\hat{\mathcal{D}}_{\mathrm{val}}$ is used throughout refinement, and we define

$$\delta^{(m+1)}_{\mathrm{MPIL}} := \left| \mathcal{I}_{\hat{\mathcal{D}}_{\mathrm{val}}}\left(\Phi^{(m+1)}\right) - \mathcal{I}_{\hat{\mathcal{D}}_{\mathrm{val}}}\left(\Phi^{(m)}\right) \right|. \quad (9)$$

By Proposition 4.1, the entropy term remains constant under the same validation distribution, so $\delta^{(m+1)}_{\mathrm{MPIL}}$ equals the corresponding change in posterior-averaged NLL and can be computed entirely offline using the empirical estimator in Eq. (8).

Let $\{\mathcal{D}_m\}_{m=1}^R$ denote the raw data subsets incorporated during refinement, and let $\hat{\mathcal{D}}^{(m)}$ denote the cumulative latent-state-augmented training data constructed from the first m subsets. We use $\Phi^{(m)}$ to denote the ensemble parameters at refinement stage m, with the initially learned ensemble

indexed by $m = 1$. At stage m, the next subset $\mathcal{D}_{m+1}$ is incorporated to construct $\hat{\mathcal{D}}^{(m+1)}$, and the ensemble is warm-started from $\Phi^{(m)}$ and refined by minimizing the empirical NLL objective

$$\Phi \leftarrow \Phi - \eta \nabla_\Phi \hat{\mathcal{L}}_{\mathrm{NLL}} \left(\Phi; \hat{\mathcal{D}}^{(m+1)}_{\mathrm{train}} \right), \qquad \Phi|_{\mathrm{start}} = \Phi^{(m)}, \tag{10}$$

where η denotes the learning rate. Repeating this update yields $\Phi^{(m+1)}$, which induces the refined empirical posterior $P^{(m+1)}_{\Theta,\tilde{\Omega}}$. After each refinement stage, both $\Phi^{(m)}$ and $\Phi^{(m+1)}$ are evaluated on the validation set to compute $\delta^{(m+1)}_{\mathrm{MPIL}}$. The refinement procedure terminates when $\delta^{(m+1)}_{\mathrm{MPIL}} \leq \epsilon_\Delta$, where ϵ_Δ is a convergence tolerance, or when all R data subsets have been incorporated. The terminating refinement stage is denoted by $m^\star$, and its empirical posterior $P^{(m^\star)}_{\Theta,\tilde{\Omega}}$ is subsequently used as the surrogate environment distribution for policy learning.

To generate each imagined trajectory, we sample $(\theta, \tilde{\omega}) \sim P^{(m^\star)}_{\Theta,\tilde{\Omega}}$ and roll out the joint policy $\tilde{\pi}$ in the corresponding world model. Let $P_{\infty,\tilde{\pi}}(\theta, \tilde{\omega})$ denote the induced imagined-trajectory distribution, and let $\check{h}_\infty$ denote the corresponding interaction history. The Bayesian policy objective is defined as

$$J^{\tilde{\pi}}_{\mathrm{Bayes}} \left(P^{(m^\star)}_{\Theta,\tilde{\Omega}} \right) = \mathbb{E}_{(\theta,\tilde{\omega}) \sim P^{(m^\star)}_{\Theta,\tilde{\Omega}}, \check{h}_\infty \sim P_{\infty,\tilde{\pi}}(\theta,\tilde{\omega})} \left[\sum_{t=0}^{\infty} \gamma^t \check{r}_t \right]. \tag{11}$$

We optimize this objective using SAC (Haarnoja et al., 2018) under CTDE (Lowe et al., 2017), where the centralized critic conditions on inferred latent states and joint actions, while each decentralized actor uses only its local observation history. The resulting policy is then evaluated fully offline using the regret. The complete procedure is summarized in Algorithm 1 in the Appendix.

5 Regret Analysis

This section relates MPIL to policy performance by bounding the Bayes-optimal policy regret as a monotone function of MPIL and characterizing its dependence on offline data size and coverage.

5.1 Controlling Regret with the MPIL

For a finite N, the true regret measures the difference in expected return between the Bayes-optimal policy $\tilde{\pi}^\star_{\mathrm{Bayes}}$ learned from the world model and the optimal policy $\tilde{\pi}^\star$ in the true environment $\mathcal{M}^\star$,

$$\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N \right) := J^{\tilde{\pi}^\star}\left(\mathcal{M}^\star \right) - J^{\tilde{\pi}^\star_{\mathrm{Bayes}}}\left(\mathcal{M}^\star \right). \tag{12}$$

Since the $\mathcal{M}^\star$ is unknown, this regret cannot be directly evaluated. Leveraging the theoretical properties of MPIL, the following theorem further establishes its connection to the true regret.

Theorem 5.1. *Under Assumption A.1 and bounded coverage, the regret satisfies*

$$\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N \right) \leq 2\mathcal{R}_{\max} \cdot \sqrt{1 - \exp\left(-\frac{C\mathcal{I}_{\hat{\mathcal{D}}_N}}{1-\gamma} \right)}. \tag{13}$$

Here, $C < \infty$ denotes a concentrability coefficient that bounds the distribution shift between the policy-induced state–action distribution and the offline data distribution, i.e., the relevant occupancy-density ratio is uniformly bounded by C. This assumption is closely related to the bounded importance-weight condition used in off-policy evaluation (Thomas & Brunskill, 2016b).

The proof is provided in Appendix A.3.2. Theorem 5.1 shows that MPIL upper-bounds the regret of the policy learned from the world model, with a smaller MPIL yielding a tighter performance guarantee. This provides support for using MPIL in offline world model refinement and selection.

5.2 Expected Regret under MPIL Contraction

Theorem 5.1 relates MPIL to policy regret. We further characterize its dependence on offline dataset size under an explicit MPIL-contraction assumption and a standard coverage condition (Ghosal et al., 2000; Wei & Mukerjee, 2003; Uehara & Sun, 2022; Zhan et al., 2022). Throughout this subsection, expectations are taken over the randomness of the offline dataset and any algorithmic randomness, when applicable.

Theorem 5.2. *Assume that the expected MPIL contracts at the parametric rate $\mathbb{E}_{\hat{\mathcal{D}}_N}[\mathcal{I}_{\hat{\mathcal{D}}_N}] \leq K/N$, where $K > 0$ is independent of N. Assume further that the concentrability coefficient $C = C(\hat{\mathcal{D}}_N)$ in Theorem 5.1 is almost surely bounded, uniformly in N, by a deterministic constant $C_{\mathrm{cov}} < \infty$. Then, the expected policy regret satisfies*

$$\mathbb{E}_{\hat{\mathcal{D}}_N}\left[\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right)\right] \leq 2\mathcal{R}_{\max}\sqrt{\frac{C_{\mathrm{cov}}K}{(1-\gamma)N}}. \tag{14}$$

Posterior contraction together with bounded coverage yields an expected regret rate of $O(N^{-1/2})$.

The proof is provided in Appendix A.3.3. Theorem 5.2 separates two sources that determine the finite-data guarantee. The contraction assumption controls how quickly the posterior world model discrepancy decreases as more offline data become available, whereas C_{cov} captures the distribution mismatch between the offline data and the state-joint-action regions visited by $\tilde{\pi}^\star$ and $\tilde{\pi}^\star_{\mathrm{Bayes}}$. Therefore, increasing the dataset tightens the regret guarantee only when these policy-relevant regions remain sufficiently covered.

5.3 Empirical Regret Approximation

The regret in Eq. (12) cannot be evaluated directly because the true environment and its optimal policy are unknown. We therefore introduce an empirical approximation based on the return predicted by the learned posterior world model. Given a policy $\tilde{\pi}^\star_{\mathrm{Bayes}}$, its model-based return is estimated through finite-horizon rollouts as $\widehat{J}^{\tilde{\pi}^\star_{\mathrm{Bayes}}}(\hat{\mathcal{M}}_N) = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{H-1}\hat{r}_t^{(\ell)}$, where $\hat{\mathcal{M}}_N$ denotes the posterior world model learned from $\hat{\mathcal{D}}_N$, and L is the number of model rollouts. Let J^+_{ref} and J^-_{ref} denote fixed upper and lower reference returns, respectively. We define the empirical regret approximation as

$$\widehat{\mathrm{Regret}}\left(\hat{\mathcal{M}}_N, \hat{\mathcal{D}}_N\right) := \frac{J^+_{\mathrm{ref}} - \widehat{J}^{\tilde{\pi}^\star_{\mathrm{Bayes}}}(\hat{\mathcal{M}}_N)}{J^+_{\mathrm{ref}} - J^-_{\mathrm{ref}}}. \tag{15}$$

This normalization provides a comparable measure of the predicted performance gap, with a smaller value indicating better policy performance. Although it does not replace the theoretical regret in Eq. (12), it provides an approximate method for evaluating policy quality entirely offline.

6 Experiments

We adopt three classic MuJoCo continuous-control tasks from OpenAI Gym, including Halfcheetah, Hopper, and Walker2d, as the experimental environments. All experiments are conducted using the D4RL benchmark (Fu et al., 2020) for offline RL. For each task, four types of offline datasets are considered, including random, medium, full-replay, and expert, resulting in a total of 12 task-dataset combinations. To investigate offline multi-agent cooperation under partial observability, we extend each original single-agent dataset to a multi-agent setting. Each agent has access only to its local observations, while the joint actions of all agents collectively determine the environment dynamics. To ensure a consistent information setting across different methods, the global state is excluded from the offline datasets and remains unavailable to all methods throughout training.

6.1 Offline Policy Performance

Policy performance is evaluated in terms of normalized average return and compared with behavior cloning (BC), OMAR (Pan et al., 2022), and MOMA-PPO (Barde et al., 2024). Once the world model satisfies the MPIL criterion, we fix the model posterior and generate imagined trajectories for cooperative multi-agent policy optimization under the CTDE paradigm. Table 1 reports the normalized returns across the three tasks and four dataset types, using the standard D4RL reference scores. Results are reported as mean $\pm$ standard deviation over 5 random seeds, with each final policy evaluated over 10 episodes. Implementation details and hyperparameter settings for all methods are provided in Appendix A.2.

Overall, the proposed method demonstrates strong and consistent policy performance across different tasks and dataset types, achieving the highest normalized return in eight of the twelve experimental settings and ranking second in the remaining four, with particularly clear advantages on the

Table 1: Normalized performance comparison on the multi-agent variants of the D4RL datasets.

Task	Dataset	BC	OMAR	MOMA-PPO	Our Work
Hopper	expert	$\mathbf{37.40 \pm 4.97}$	1.82 ± 0.01	18.86 ± 2.97	29.16 ± 4.48
	full-replay	14.73 ± 2.30	1.86 ± 0.01	31.30 ± 2.98	$\mathbf{34.72 \pm 4.66}$
	medium	21.79 ± 4.65	11.79 ± 0.91	13.92 ± 3.13	$\mathbf{28.63 \pm 3.01}$
	random	3.24 ± 0.60	9.49 ± 0.06	16.96 ± 2.49	$\mathbf{31.16 \pm 0.66}$
Halfcheetah	expert	5.59 ± 3.65	8.56 ± 5.14	2.61 ± 0.76	$\mathbf{19.80 \pm 2.34}$
	full-replay	25.91 ± 5.94	33.47 ± 9.61	$\mathbf{50.13 \pm 0.64}$	43.32 ± 3.51
	medium	44.62 ± 9.61	41.80 ± 1.12	$\mathbf{50.35 \pm 0.53}$	46.09 ± 3.40
	random	2.25 ± 0.01	16.93 ± 0.74	15.18 ± 0.27	$\mathbf{24.38 \pm 0.31}$
Walker2d	expert	$\mathbf{106.02 \pm 2.61}$	1.79 ± 0.25	75.19 ± 14.66	85.35 ± 11.44
	full-replay	22.45 ± 6.57	1.03 ± 0.15	26.52 ± 7.28	$\mathbf{63.57 \pm 11.93}$
	medium	20.28 ± 7.81	21.32 ± 3.98	7.96 ± 0.79	$\mathbf{29.62 \pm 6.02}$
	random	1.51 ± 0.18	1.77 ± 0.57	1.18 ± 0.14	$\mathbf{2.22 \pm 0.77}$

random and full-replay datasets. In comparison, BC performs well on the expert datasets of Hopper and Walker2d but degrades substantially on the random and some full-replay datasets, reflecting its strong dependence on the quality of the offline behavior data. MOMA-PPO achieves the best performance on the Halfcheetah full-replay and medium datasets, but its performance is less consistent across the other tasks. OMAR also exhibits noticeable performance variation across environments and yields relatively low returns in most of the settings considered. These results indicate that the proposed method adapts well to variations in offline data quality and coverage, and can extract decision-relevant information from offline trajectories rather than relying primarily on high-quality behavior demonstrations.

6.2 Validation of MPIL and Approximate Regret

To validate the proposed offline evaluation mechanism, we jointly analyze δ_{MPIL} and approximate regret across different offline dataset sizes. As shown in Fig. 2, both δ_{MPIL} and approximate regret decrease substantially at the early stage with increasing offline data in the Walker2d environment. This indicates that, as the posterior world model becomes more accurate, the downstream policy learned from it also exhibits improved performance. Once the dataset size reaches approximately 0.5M, both metrics gradually stabilize, suggesting that the information provided by additional offline data yields only limited further improvement to either the world model or the learned policy. Moreover, the approximate regret follows a trend consistent with the true regret and stabilizes at a similar data scale, supporting its validity as an offline proxy for policy performance. The consistent stabilization of δ_{MPIL}, approximate regret, and true regret further provides a practical indication of data sufficiency, allowing the data scale required for stable policy performance to be estimated without online interaction.

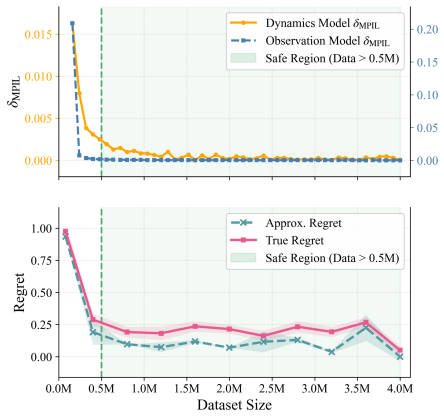


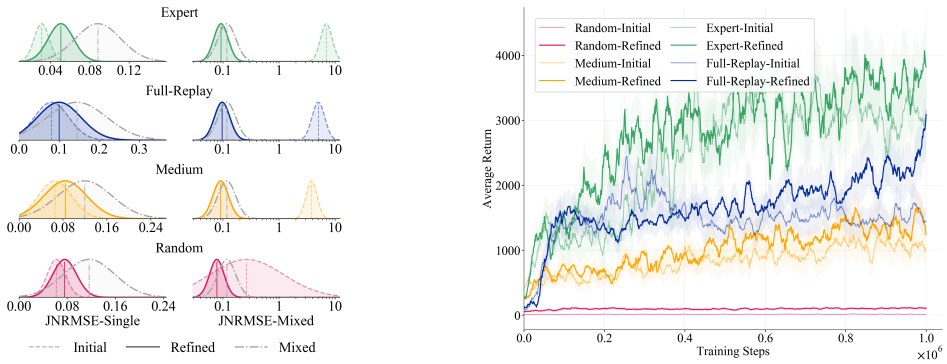
Figure 2: Relationship between δ_{MPIL}, Approximate Regret, and True Regret under Different Offline Dataset Sizes on Walker2d. The True Regret Is Normalized Using the Standard D4RL Reference Scores.

6.3 Effect of MPIL-Guided World Model Refinement

To further examine the effect of MPIL-guided world model refinement, we compare the predictive performance of the world models and their corresponding downstream policies before and after refinement under the same data budget. For each of the four dataset types, the baseline model is trained directly on the full set of 0.5M samples, whereas the refined model is initialized with a subset

and progressively updated with additional samples according to the MPIL criterion. Both models are evaluated on the corresponding validation set and on a mixed validation set constructed from all four dataset types. As an additional reference, we also train a separate world model on 0.5M samples from the mixed dataset and evaluate it using the same validation settings. World model prediction performance is measured by the joint normalized root mean square error (JNRMSE), which jointly characterizes observation and reward prediction errors. Since simulator global states are excluded and not predicted, state prediction error is omitted.

As shown in Fig. 3a, MPIL-guided refinement in the Walker2d environment generally reduces the prediction error of the world models trained on individual datasets. The improvement is particularly evident on the mixed validation set, where the initial models exhibit substantially larger errors under the shifted data distribution. After refinement, their JNRMSE values are markedly reduced and are generally lower than those of the reference model trained directly on 0.5M mixed samples. These results indicate that MPIL-guided refinement can utilize a fixed amount of offline data more effectively, yielding more accurate and robust world model predictions across different data distributions. The benefits of world model refinement are further reflected in policy learning. As shown in Fig. 3b, across all four Walker2d datasets, policies trained with the refined world models achieve higher returns, with the improvement being particularly pronounced on the full-replay dataset. This indicates that improved world model predictions can provide more reliable imagined trajectories and thereby enhance offline policy optimization.



(a) World model predictive performance

(b) Policy performance

Figure 3: Effect of world model refinement on policy performance in Walker2d.

7 Conclusion

This paper studies fully offline MARL under decentralized partial observability, where the global state is unavailable during training. We develop a Bayesian world model approach that infers latent states from local observation histories and models transition-reward and observation dynamics through posterior distributions. MPIL is further introduced to refinement of the posterior world model and enable fully offline evaluation, while posterior-consistent imagined trajectories are used to optimize cooperative multi-agent policies. Theoretical analysis establishes a monotone upper bound on policy regret in terms of MPIL and further characterizes its dependence on the size and coverage of the offline dataset. Experiments on partially observable multi-agent MuJoCo tasks show that the proposed method achieves competitive policy performance across datasets of varying quality. MPIL decreases consistently with policy regret as offline information increases, while MPIL-guided refinement further improves world model prediction and downstream policy performance.

Future work will investigate zero-shot generalization to unseen agent configurations, interaction structures, and environment dynamics. We will also explore shared or conditioned world models that can reuse learned representations and dynamics across different task settings rather than retraining a new model from scratch. When model refinement is required, MPIL may further be used to identify unreliable regions and select the most informative offline samples for targeted updates, improving the data efficiency of world model refinement.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, manuscript organization and presentation, clarification and checking of mathematical derivations and theoretical exposition, and interpretation of experimental results. Generative AI tools were also used to assist with LaTeX formatting and technical writing. All AI-assisted theoretical statements, mathematical derivations, and experimental analyses were independently reviewed and verified by the authors. Experimental data, results, and conclusions were produced and validated by the authors, and all AI-assisted text was manually reviewed and revised. Generative AI tools were not used to generate or alter experimental data or results. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study does not involve human participants, personally identifiable information, or sensitive private data.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed framework, model formulation, training procedure, and experimental settings in the main paper and appendix. The key hyperparameters, network configurations, offline data generation process, baseline implementations, and evaluation metrics are reported to facilitate reproduction of the experimental results. Complete derivations and proofs of the theoretical results are also provided in the appendix.

REFERENCES

Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhihong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. *arXiv preprint arXiv:2202.11566*, 2022.

Paul Barde, Jakob N. Foerster, Derek Nowrouzezahrai, and Amy Zhang. A model-based solution to the offline multi-agent reinforcement learning coordination problem. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 141–150, 2024.

Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3/4):221–229, 1956.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840, 2002.

Longchao Da, Porter Jenkins, Trevor Schwantes, Jeffrey Dotson, and Hua Wei. Probabilistic offline policy ranking with approximate bayesian computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20370–20378, 2024.

Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Harry Emerson, Matthew Guy, and Ryan McConville. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, 2023.

Mattie Fellows, Clarisse Wibault, Uljad Berdica, Johannes Forkel, Michael A. Osborne, and Jakob Nicolaus Foerster. Fully offline reinforcement learning. *Reinforcement Learning Journal*, 7, 2026.

Jose P Folch, Calvin Tsay, Robert M Lee, Behrang Shafei, Weronika Ormaniec, Andreas Krause, Mark van der Wilk, Ruth Misener, and Mojmír Mutnỳ. Transition constrained bayesian optimization via markov decision processes. *Advances in Neural Information Processing Systems*, 37: 88194–88235, 2024.

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.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000. doi: 10.1214/aos/1016218228.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, pp. 1861–1870. PMLR, 2018.

Yulong Hu, Tingting Dong, and Sen Li. Coordinating ride-pooling with public transit using reward-guided conservative q-learning: An offline training and online fine-tuning reinforcement learning framework. *Transportation Research Part C: Emerging Technologies*, 174:105051, 2025.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International Conference on Machine Learning*, pp. 2117–2126. PMLR, 2018.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 21810–21823, 2020.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Dongsu Lee and Minhae Kwon. Scenario-free autonomous driving with multi-task offline-to-online reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 26(9):13317–13330, 2025.

Kenzo Lobos-Tsunekawa, Akshay Srinivasan, and Michael Spranger. Ma-dreamer: Coordination and communication through shared imagination. *arXiv preprint arXiv:2204.04687*, 2022.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Chengzhong Ma, Deyu Yang, Tianyu Wu, Zeyang Liu, Houxue Yang, Xingyu Chen, Xuguang Lan, and Nanning Zheng. Improving offline reinforcement learning with in-sample advantage regularization for robot manipulation. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (6):11215–11227, 2024.

Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International Conference on Machine Learning*, pp. 17221–17237. PMLR, 2022.

Zhongjian Qiao, Jiafei Lyu, Kechen Jiao, Qi Liu, and Xiu Li. Sumo: Search-based uncertainty estimation for model-based offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20033–20041, 2025.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Stéphane Ross, Joelle Pineau, Brahim Chaib-Draa, and Pierre Kreitmann. A bayesian approach for learning and planning in partially observable markov decision processes. *Journal of Machine Learning Research*, 12(5), 2011.

Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip H. S. Torr, Wendelin Böhmer, and Shimon Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, 2020.

Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative q learning for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:77290–77312, 2023.

Zifeng Shi, Meiqin Liu, Senlin Zhang, Ronghao Zheng, Shanling Dong, and Ping Wei. Gawm: Global-aware world model for multi-agent reinforcement learning. *arXiv preprint arXiv:2501.10116*, 2025.

Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-bellman inconsistency for model-based offline reinforcement learning. In *International Conference on Machine Learning*, pp. 33177–33194. PMLR, 2023.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016a.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148. PMLR, 2016b.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *Statistical Science*, 41(3):561–581, 2026.

Aravind Venugopal, Stephanie Milani, Fei Fang, and Balaraman Ravindran. Mabl: Bi-level latent-variable world model for sample-efficient multi-agent reinforcement learning. *arXiv preprint arXiv:2304.06011*, 2023.

Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. *Advances in Neural Information Processing Systems*, 36:52413–52429, 2023.

Gang Wei and Ruhal Mukerjee. Predictive densities: General asymptotic results and admissibility. *Statistica Sinica*, 13(4):1165–1178, 2003.

Fan Wu, Rui Zhang, Qi Yi, Yunkai Gao, Jiaming Guo, Shaohui Peng, Siming Lan, Husheng Han, Yansong Pan, Kaizhao Yuan, et al. Ocean-mbrl: Offline conservative exploration for model-based offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15897–15905, 2024.

Zhiwei Xu, Yunpeng Bai, Dapeng Li, Bin Zhang, and Guoliang Fan. SIDE: State inference for partially observable cooperative multi-agent reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1400–1408, 2022.

Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

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.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775. PMLR, 2022.

Siyuan Zhang and Nan Jiang. Towards hyperparameter-free policy selection for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12864–12875, 2021.

A Appendix

A.1 Algorithm

Algorithm 1 MPIL-Guided Bayesian Offline MARL

Require: Raw offline data subsets $\{\mathcal{D}_m\}_{m=1}^{R}$, convergence tolerance ϵ_Δ, priors P_Θ and $P_{\tilde{\Omega}}$

Construct $\hat{\mathcal{D}}^{(1)}$ from $\mathcal{D}_1$ via latent-state inference
Learn the initial ensemble parameters $\Phi^{(1)}$ by minimizing $\mathcal{L}_{\mathrm{WM}}^{(1)}$
Initialize the terminating stage: $m^\star \leftarrow R$
for $m = 1, \ldots, R-1$ **do**
  Incorporate $\mathcal{D}_{m+1}$ to construct $\hat{\mathcal{D}}^{(m+1)}$
  Obtain $\Phi^{(m+1)}$ by warm-starting from $\Phi^{(m)}$ and minimizing $\mathcal{L}_{\mathrm{WM}}^{(m+1)}$
  Compute $\delta_{\mathrm{MPIL}}^{(m+1)} = \left| \mathcal{I}_{\hat{\mathcal{D}}_{\mathrm{val}}}(\Phi^{(m+1)}) - \mathcal{I}_{\hat{\mathcal{D}}_{\mathrm{val}}}(\Phi^{(m)}) \right|$
  if $\delta_{\mathrm{MPIL}}^{(m+1)} \le \epsilon_\Delta$ **then**
    Record the terminating stage: $m^\star \leftarrow m+1$
    break
  end if
end for
Construct the final empirical posterior: $P_{\Theta,\tilde{\Omega}}^{(m^\star)}(\cdot) \leftarrow \frac{1}{K}\sum_{k=1}^{K} \delta_{(\theta_k^{(m^\star)}, \tilde{\omega}_k^{(m^\star)})}(\cdot)$
Sample a world model and an imagined trajectory: $(\theta, \tilde{\omega}) \sim P_{\Theta,\tilde{\Omega}}^{(m^\star)}, \quad \check{h}_\infty \sim P_{\infty,\tilde{\pi}}(\theta, \tilde{\omega})$
Optimize the joint policy via CTDE: $\tilde{\pi}_{\mathrm{Bayes}}^\star \leftarrow \arg\max_{\tilde{\pi}} J_{\mathrm{Bayes}}^{\tilde{\pi}}\left(P_{\Theta,\tilde{\Omega}}^{(m^\star)}\right)$
return Bayes-optimal joint policy $\tilde{\pi}_{\mathrm{Bayes}}^\star$

A.2 Implementation Details

A.2.1 World Model Implementation

The latent-state inference module uses a recurrent encoder to summarize the joint observation–action history into a centralized latent state. At each time step, the joint observation and previous joint action are first projected through a linear layer with ReLU activation and then passed to a GRUCell. Given the inferred latent state and current joint action, an observation decoder predicts the next joint observation, while a shared transition–reward network predicts the next latent state and reward. The implementation details are summarized in Table 2.

Table 2: Latent-state inference model hyperparameters.

Hyperparameter	Value
Observation dimension	17
Joint action dimension	6
Latent state dimension	64
GRU hidden dimension	64
Encoder projection dimension	256
Decoder hidden dimension	256
Transition–reward hidden dimension	256
Decoder MLP layers	2
Transition–reward MLP layers	2
Sequence length	32
Input normalization	Yes
Optimizer	Adam
Pretraining learning rate	3×10^{-4}
Fine-tuning learning rate	1×10^{-4}
Activation function	ReLU; GRU sigmoid/tanh
λ_recon	1.0
λ_z	1.0
λ_reward	1.0

The inferred latent states are then used to construct the posterior world model described in Section 4.2. The posterior over the transition–reward and observation-model parameters is approximated by an ensemble of probabilistic neural networks, with elite members selected according to holdout performance for posterior prediction and imagined rollouts. The ensemble is optimized using the empirical MPIL objective defined in Section 4.2. The main implementation hyperparameters are summarized in Table 3.

Table 3: Posterior world model implementation hyperparameters.

Hyperparameter	Value
Ensemble size	7
Number of elite models	5
Hidden dimension	300
Probabilistic output	Mean and variance
Holdout ratio	0.2
Optimizer	Adam
Learning rate	1×10^{-3}
Activation function	Swish; linear output; bounded log-variance

A.2.2 World Model Refinement

The refinement procedure follows Section 4.3 and uses the same world model ensemble architecture as A.2.1. Each stage is warm-started from the preceding ensemble and optimized using the empirical MPIL objective in Section 4.2, with holdout-based validation and early stopping. The main hyperparameters are summarized in Table 4.

A.2.3 Policy Learning with the Refined World Model

Policy optimization follows the SAC-based CTDE formulation in Section 4.3. Each decentralized actor is implemented as a two-layer MLP with a Gaussian policy output, while the centralized critic uses twin Q-networks with the same backbone structure. Imagined transitions generated from the refined world model ensemble are combined with offline data for actor and critic updates. The main policy-learning and rollout hyperparameters are summarized in Table 5.

Table 4: Hyperparameters for MPIL-based world model refinement.

Hyperparameter	**Value**
Maximum refinement rounds	50
Epochs per round	10
Batch size	512
Holdout ratio	0.2
Early-stopping patience	5
Optimizer	Adam
Learning rate	1×10^{-3}
Activation function	Swish; linear output; bounded log-variance

Table 5: Policy learning and model rollout hyperparameters.

Hyperparameter	**Value**
Actor hidden dimensions	$[256, 256]$
Critic hidden dimensions	$[256, 256]$
Optimizer	Adam
Learning rate	5×10^{-5}
Activation function	ReLU; linear output; tanh squash
Discount factor γ	0.99
Target update coefficient	0.005
Batch size	256
Total training steps	1×10^{6}
Rollout length	1
Rollout batch size	50,000
Rollout frequency	1000 steps
Conservative loss weight	5.0
Evaluation episodes	10

A.2.4 Baseline Implementation

We compare our method with behavior cloning (BC), OMAR, and MOMA-PPO (Barde et al., 2024) under the same D4RL tasks and multi-agent action decomposition. The main baseline hyperparameters are summarized in Table 6.

BC is implemented as deterministic behavior cloning with a mean-squared action loss. OMAR follows the conservative critic regularization and candidate-action-based actor rectification of Pan et al. (2022), with a conservative critic weight of 2.0, a behavior-cloning weight of 1.0, an OMAR objective weight of 0.5, a candidate-action number of 8, and a critic-target clipping threshold of 100.

MOMA-PPO follows the model-based offline MARL procedure of Barde et al. (2024). A learned world model is used to generate synthetic multi-agent trajectories initialized from the offline dataset. At each rollout step, actions are sampled from the current decentralized policies and the world model predicts the subsequent transition and reward. Model uncertainty is used to penalize unreliable synthetic transitions and to terminate rollouts when the model becomes insufficiently reliable. The resulting synthetic trajectories are then used to update the agents with MAPPO under centralized training and decentralized execution.

Table 6: Main hyperparameters of the baseline methods for Walker2d-random.

Hyperparameter	**BC**	**OMAR**	**MOMA-PPO**
Training steps	1×10^6	1×10^6	1×10^6
Batch size	2048	512	256
Hidden dimension	128	300	$[256, 256]$
Learning rate	3×10^{-5}	3×10^{-5}	5×10^{-5}
Discount factor	–	0.99	0.99
Entropy coefficient	–	0.02	0.05
Gradient clipping	5.0	5.0	–

A.2.5 Multi-Agent, Local-Observation, and Dataset Configuration

Table 7 gives the exact task decomposition. All indices are zero-based, and entries within each cell are ordered by agent ID. At time t, actor i receives only $o_t^i = s_t[\mathcal{O}_i]$ and writes its output to the specified coordinates of the joint action a_t; the environment then executes the complete joint action.

Table 7: Multi-agent action and observation decomposition. The task dimensions are shown as (observation, action).

Task (dims.)	**No. agents**	**Action dim.: indices**	**Local observation indices $\mathcal{O}_i$**	**Shared/overlap indices**
HalfCheetah-v2 $(17, 6)$	3	$2 : [0, 1]$ $2 : [2, 3]$ $2 : [4, 5]$	$\mathcal{O}_0 = [0, 1, 2, 3, 8, 9, 10, 11, 12]$ $\mathcal{O}_1 = [0, 1, 4, 5, 8, 9, 10, 13, 14]$ $\mathcal{O}_2 = [0, 1, 6, 7, 8, 9, 10, 15, 16]$	$[0, 1, 8, 9, 10]$
Hopper-v2 $(11, 3)$	3	$1 : [0]$ $1 : [1]$ $1 : [2]$	$\mathcal{O}_0 = [0, 1, 2, 5, 6, 7, 8]$ $\mathcal{O}_1 = [0, 1, 3, 5, 6, 7, 9]$ $\mathcal{O}_2 = [0, 1, 4, 5, 6, 7, 10]$	$[0, 1, 5, 6, 7]$
Walker2d-v2 $(17, 6)$	2	$3 : [0, 1, 2]$ $3 : [3, 4, 5]$	$\mathcal{O}_0 = [0, 1, 2, 3, 4, 8, 9, 10, 11, 12, 13]$ $\mathcal{O}_1 = [0, 1, 5, 6, 7, 8, 9, 10, 14, 15, 16]$	$[0, 1, 8, 9, 10]$

Thus, partial observability is imposed on the decentralized actor inputs: the union of local observations reconstructs the original D4RL observation, whereas no single actor observes it in full.

Dataset IDs and source. We use the D4RL Gym-MuJoCo `v2` datasets (Fu et al., 2020) listed in Table 8.

Table 8: D4RL Gym-MuJoCo `v2` dataset IDs used in our experiments.

Task	**Random**	**Medium**	**Expert**	**Full-replay**
HalfCheetah	`halfcheetah-random-v2`	`halfcheetah-medium-v2`	`halfcheetah-expert-v2`	`halfcheetah-full-replay-v2`
Hopper	`hopper-random-v2`	`hopper-medium-v2`	`hopper-expert-v2`	`hopper-full-replay-v2`
Walker2d	`walker2d-random-v2`	`walker2d-medium-v2`	`walker2d-expert-v2`	`walker2d-full-replay-v2`

The `full-replay-v2` datasets are the complete replay buffers distributed by D4RL, rather than renamed `medium-replay` datasets or mixtures constructed by us. They contain neither model rollouts generated by our method nor trajectories resampled from the evaluation environments.

A.3 Proof

Assumption A.1 (Latent sufficiency and factorization)**.** *For the theoretical analysis, we consider the realizable idealization in which $\hat{s}_t$ is predictively sufficient—future environment variables depend on the joint history only through $\hat{s}_t$. Let $x = (\hat{s}, \tilde{a})$, $u = (\hat{s}', r)$, and $v = \tilde{z}'$. The true and learned one-step kernels used in the analysis are*

$$K^\star(\mathrm{d}u, \mathrm{d}v \mid x) := P^\star_{T,R}(\mathrm{d}u \mid x) P^\star_{\tilde{O}}(\mathrm{d}v \mid x),$$
$$K_{\theta,\tilde{\omega}}(\mathrm{d}u, \mathrm{d}v \mid x) := P_{T,R}(\mathrm{d}u \mid x, \theta) P_{\tilde{O}}(\mathrm{d}v \mid x, \tilde{\omega}). \tag{16}$$

For every $\mu \in \{\nu_{\hat{\mathcal{D}}_N}, \rho^{\tilde{\pi}^\star}, \rho^{\tilde{\pi}^\star_{\mathrm{Bayes}}}\}$, we further assume that $K^\star(\cdot \mid x) \ll K_{\theta,\tilde{\omega}}(\cdot \mid x)$ for μ-almost every x and posterior-almost every $(\theta, \tilde{\omega})$, and that

$$\mathbb{E}_{x\sim\mu}\mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot|\hat{\mathcal{D}}_N)}[D_{\mathrm{KL}}(K^\star(\cdot \mid x)\|K_{\theta,\tilde{\omega}}(\cdot \mid x))] < \infty. \tag{17}$$

These conditions define the scope of the theoretical analysis and are not implied by the recurrent encoder. Equation (1) is a practical deterministic history encoder; Theorem 5.1 does not quantify deviations from this idealized setting.

A.3.1 Proof of Proposition 4.1

Proof. For notational compactness, let $x = (\hat{s}, \tilde{a})$, let $y_{T,R} = (s', r)$, and let $y_{\tilde{O}} = \tilde{o}'$. For any fixed x and posterior parameter realization θ, the definition of the KL divergence gives

$$\begin{aligned} &D_{\mathrm{KL}}\big(P^\star_{T,R}(\cdot \mid x)\big\|P_{T,R}(\cdot \mid x,\theta)\big) \\ &\quad= \mathbb{E}_{y_{T,R}\sim P^\star_{T,R}(\cdot|x)}\left[\log P^\star_{T,R}(y_{T,R} \mid x) - \log P_{T,R}(y_{T,R} \mid x,\theta)\right] \\ &\quad= \mathbb{E}_{y_{T,R}\sim P^\star_{T,R}(\cdot|x)}\left[-\log P_{T,R}(y_{T,R} \mid x,\theta)\right] - H\big(P^\star_{T,R}(\cdot \mid x)\big). \end{aligned} \tag{18}$$

Likewise, for any fixed x and $\tilde{\omega}$,

$$\begin{aligned} &D_{\mathrm{KL}}\big(P^\star_{\tilde{O}}(\cdot \mid x)\big\|P_{\tilde{O}}(\cdot \mid x,\tilde{\omega})\big) \\ &\quad= \mathbb{E}_{y_{\tilde{O}}\sim P^\star_{\tilde{O}}(\cdot|x)}\left[-\log P_{\tilde{O}}(y_{\tilde{O}} \mid x,\tilde{\omega})\right] - H\big(P^\star_{\tilde{O}}(\cdot \mid x)\big). \end{aligned} \tag{19}$$

These identities hold for probability mass functions and, with H interpreted as differential entropy, for probability densities, whenever the displayed quantities are well defined.

Substituting Eqs. (18) and (19) into the definition of MPIL in Eq. (5), and then taking expectations with respect to the fixed offline data distribution and the posterior, yields

$$\begin{aligned} \mathcal{I}_{\hat{\mathcal{D}}_N} = \mathbb{E}_{x\sim\nu_{\hat{\mathcal{D}}_N}}\mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot|\hat{\mathcal{D}}_N)}\Big[&\mathbb{E}_{y_{T,R}\sim P^\star_{T,R}(\cdot|x)}[-\log P_{T,R}(y_{T,R} \mid x,\theta)] \\ &+ \mathbb{E}_{y_{\tilde{O}}\sim P^\star_{\tilde{O}}(\cdot|x)}[-\log P_{\tilde{O}}(y_{\tilde{O}} \mid x,\tilde{\omega})]\Big] \\ - \mathbb{E}_{x\sim\nu_{\hat{\mathcal{D}}_N}}\Big[&H\big(P^\star_{T,R}(\cdot \mid x)\big) + H\big(P^\star_{\tilde{O}}(\cdot \mid x)\big)\Big]. \end{aligned} \tag{20}$$

The first term on the right-hand side of Eq. (20) is exactly the posterior-averaged NLL $\mathcal{L}_{\mathrm{NLL}}$ defined in Eq. (4), while the second term is $\mathcal{H}^\star_{\hat{\mathcal{D}}_N}$. Consequently,

$$\mathcal{I}_{\hat{\mathcal{D}}_N} = \mathcal{L}_{\mathrm{NLL}} - \mathcal{H}^\star_{\hat{\mathcal{D}}_N}, \tag{21}$$

which proves Eq. (6). Because the true-environment entropy term depends only on $P^\star_{T,R}$, $P^\star_{\tilde{O}}$, and the fixed distribution $\nu_{\hat{\mathcal{D}}_N}$, it is independent of the learned world model parameters. Hence differences in posterior-averaged NLL between world models evaluated under the same offline data distribution equal the corresponding differences in MPIL. □

A.3.2 Proof of Theorem 5.1

Assumption A.2. *For each relevant policy $\tilde{\pi} \in \{\tilde{\pi}^\star, \tilde{\pi}^\star_{\mathrm{Bayes}}\}$, assume that the policy-induced state-action distribution $\rho^{\tilde{\pi}}$ is absolutely continuous with respect to the offline data distribution $\nu_{\hat{\mathcal{D}}_N}$ and that*

$$C := \max_{\tilde{\pi}\in\{\tilde{\pi}^\star,\tilde{\pi}^\star_{\mathrm{Bayes}}\}} \operatorname*{ess\,sup}_{(\hat{s},\tilde{a})} \frac{\rho^{\tilde{\pi}}(\hat{s},\tilde{a})}{\nu_{\hat{\mathcal{D}}_N}(\hat{s},\tilde{a})} < \infty, \tag{22}$$

where the essential supremum is taken with respect to $\nu_{\hat{\mathcal{D}}_N}$.

Assumption A.2 requires the offline data distribution to cover the state–action regions visited by the two policies relevant to the regret comparison, but does not require uniform coverage over all possible policies.

This coverage condition is closely related to the bounded importance-weight assumption commonly adopted in off-policy evaluation (Thomas & Brunskill, 2016b): both require the target policies to remain within the support of the offline data-generating distribution. We use it here only as an explicit conditional assumption and do not claim that the policy-learning algorithm enforces it automatically.

Proof. For a joint policy $\tilde{\pi}$, write its posterior-predictive value as

$$J_B^{\tilde{\pi}} := J_{\mathrm{Bayes}}^{\tilde{\pi}}\left(P_{\Theta,\tilde{\Omega}}(\cdot \mid \hat{\mathcal{D}}_N)\right). \tag{23}$$

Let $\tilde{\pi}^\star$ maximize the return in the true environment $\mathcal{M}^\star$, and let $\tilde{\pi}^\star_{\mathrm{Bayes}}$ maximize $J_B^{\tilde{\pi}}$. Adding and subtracting the two corresponding posterior-predictive values in Eq. (12) gives

$$\begin{aligned}
&\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right) \\
&= J^{\tilde{\pi}^\star}(\mathcal{M}^\star) - J_B^{\tilde{\pi}^\star} + J_B^{\tilde{\pi}^\star} - J_B^{\tilde{\pi}^\star_{\mathrm{Bayes}}} + J_B^{\tilde{\pi}^\star_{\mathrm{Bayes}}} - J^{\tilde{\pi}^\star_{\mathrm{Bayes}}}(\mathcal{M}^\star) \\
&\le \left|J^{\tilde{\pi}^\star}(\mathcal{M}^\star) - J_B^{\tilde{\pi}^\star}\right| + \left|J_B^{\tilde{\pi}^\star_{\mathrm{Bayes}}} - J^{\tilde{\pi}^\star_{\mathrm{Bayes}}}(\mathcal{M}^\star)\right|,
\end{aligned} \tag{24}$$

where the inequality follows from $J_B^{\tilde{\pi}^\star} - J_B^{\tilde{\pi}^\star_{\mathrm{Bayes}}} \le 0$. Thus, it suffices to control the return discrepancy for these two policies rather than uniformly over all policies.

We next establish such a bound for an arbitrary fixed joint policy $\tilde{\pi}$. Define its posterior-predictive trajectory-prefix law by

$$\widehat{P}_{t+1,\tilde{\pi}}(\hat{\mathcal{D}}_N) := \mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot\mid\hat{\mathcal{D}}_N)}\left[P_{t+1,\tilde{\pi}}(\theta,\tilde{\omega})\right]. \tag{25}$$

We assume that the true and posterior-predictive processes have the same initial latent-state and initial-observation distribution; an additional initial-distribution divergence term is required otherwise. Since rewards lie in $[r_{\min}, r_{\max}]$, the variational characterization of total variation implies

$$\begin{aligned}
&\left|J^{\tilde{\pi}}(\mathcal{M}^\star) - J_B^{\tilde{\pi}}\right| \\
&\le (r_{\max} - r_{\min}) \sum_{t=0}^{\infty} \gamma^t D_{\mathrm{TV}}\left(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star), \widehat{P}_{t+1,\tilde{\pi}}(\hat{\mathcal{D}}_N)\right) \\
&= \mathcal{R}_{\max}\mathbb{E}_{t\sim\mathcal{G}(\gamma)}\left[D_{\mathrm{TV}}\left(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star), \widehat{P}_{t+1,\tilde{\pi}}(\hat{\mathcal{D}}_N)\right)\right],
\end{aligned} \tag{26}$$

where $\mathcal{R}_{\max} = (r_{\max} - r_{\min})/(1-\gamma)$ and $\Pr_{\mathcal{G}(\gamma)}(T = t) = (1-\gamma)\gamma^t$.

Let $g(x) := \sqrt{1 - \exp(-x)}$. The Bretagnolle–Huber inequality, convexity of KL divergence in its second argument, and monotonicity and concavity of g yield

$$\begin{aligned}
&\mathbb{E}_{t\sim\mathcal{G}(\gamma)}\left[D_{\mathrm{TV}}\left(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star), \widehat{P}_{t+1,\tilde{\pi}}(\hat{\mathcal{D}}_N)\right)\right] \\
&\le g\left(\mathbb{E}_{t\sim\mathcal{G}(\gamma)}\mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot\mid\hat{\mathcal{D}}_N)}\left[D_{\mathrm{KL}}(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star)\|P_{t+1,\tilde{\pi}}(\theta,\tilde{\omega}))\right]\right).
\end{aligned} \tag{27}$$

For $x = (\hat{s}, \tilde{a})$, define the posterior-averaged one-step model discrepancy

$$\begin{aligned}
d_{\hat{\mathcal{D}}_N}(x) := \mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot\mid\hat{\mathcal{D}}_N)}\Big[&D_{\mathrm{KL}}\big(P^\star_{T,R}(\cdot \mid x)\big\|P_{T,R}(\cdot \mid x,\theta)\big) \\
&+ D_{\mathrm{KL}}\big(P^\star_{\tilde{O}}(\cdot \mid x)\big\|P_{\tilde{O}}(\cdot \mid x,\tilde{\omega})\big)\Big].
\end{aligned} \tag{28}$$

Under Assumption A.1, the KL identity for product measures gives, for each posterior-supported $(\theta,\tilde{\omega})$,

$$\begin{aligned}
&D_{\mathrm{KL}}(K^\star(\cdot \mid x)\|K_{\theta,\tilde{\omega}}(\cdot \mid x)) \\
&\quad = D_{\mathrm{KL}}\big(P^\star_{T,R}(\cdot \mid x)\big\|P_{T,R}(\cdot \mid x,\theta)\big) + D_{\mathrm{KL}}\big(P^\star_{\tilde{O}}(\cdot \mid x)\big\|P_{\tilde{O}}(\cdot \mid x,\tilde{\omega})\big).
\end{aligned} \tag{29}$$

Thus, $d_{\hat{\mathcal{D}}_N}(x)$ is exactly the posterior-averaged KL divergence between the true and learned one-step kernels; no mutual-information term appears because both kernels factorize in Assumption A.1.

For a fixed joint policy $\tilde{\pi}$, the true and learned trajectory-prefix laws factorize as

$$\begin{aligned}
P_{t+1,\tilde{\pi}}(\mathcal{M}^\star) &= P_0 \prod_{j=0}^{t} \tilde{\pi}(\tilde{a}_j \mid \tilde{h}_j) K^\star(u_j, v_j \mid x_j), \\
P_{t+1,\tilde{\pi}}(\theta,\tilde{\omega}) &= P_0 \prod_{j=0}^{t} \tilde{\pi}(\tilde{a}_j \mid \tilde{h}_j) K_{\theta,\tilde{\omega}}(u_j, v_j \mid x_j),
\end{aligned} \tag{30}$$

where $u_j = (\hat{s}_{j+1}, r_j)$, $v_j = \tilde{z}_{j+1}$, and the common P_0 is the initial latent-state and observation law specified above. The common initial and policy factors therefore cancel in the likelihood ratio. Applying the KL chain rule and then Eq. (29) gives

$$\mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot|\hat{\mathcal{D}}_N)}\left[D_{\mathrm{KL}}(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star)\|P_{t+1,\tilde{\pi}}(\theta,\tilde{\omega}))\right] = \sum_{j=0}^{t}\mathbb{E}_{x_j\sim P_j^{\tilde{\pi}}(\mathcal{M}^\star)}\left[d_{\hat{\mathcal{D}}_N}(x_j)\right]. \tag{31}$$

Define the normalized policy-induced state–action distribution by

$$\rho^{\tilde{\pi}} := \mathbb{E}_{t\sim\mathcal{AG}(\gamma)}\mathbb{E}_{j\sim\mathcal{U}_t}\left[P_j^{\tilde{\pi}}(\mathcal{M}^\star)\right], \tag{32}$$

where $\mathcal{U}_t$ is uniform on $\{0,\dots,t\}$ and $\Pr_{\mathcal{AG}(\gamma)}(T=t) = (1-\gamma)^2(t+1)\gamma^t$. Combining Eqs. (31) and (32), we obtain

$$\mathbb{E}_{t\sim\mathcal{G}(\gamma)}\mathbb{E}_{(\theta,\tilde{\omega})\sim P_{\Theta,\tilde{\Omega}}(\cdot|\hat{\mathcal{D}}_N)}\left[D_{\mathrm{KL}}(P_{t+1,\tilde{\pi}}(\mathcal{M}^\star)\|P_{t+1,\tilde{\pi}}(\theta,\tilde{\omega}))\right] = \frac{1}{1-\gamma}\mathbb{E}_{x\sim\rho^{\tilde{\pi}}}\left[d_{\hat{\mathcal{D}}_N}(x)\right]. \tag{33}$$

Let

$$\mathcal{I}^{\tilde{\pi}} := \mathbb{E}_{x\sim\rho^{\tilde{\pi}}}\left[d_{\hat{\mathcal{D}}_N}(x)\right] \tag{34}$$

denote the policy-dependent information loss used only as an intermediate quantity in the proof. Equations (26)– (33) imply

$$\left|J^{\tilde{\pi}}(\mathcal{M}^\star) - J_B^{\tilde{\pi}}\right| \le \mathcal{R}_{\max}\sqrt{1-\exp\left(-\frac{\mathcal{I}^{\tilde{\pi}}}{1-\gamma}\right)}. \tag{35}$$

It remains to connect the policy-dependent quantity in Eq. (34) to the dataset-weighted MPIL. Under the bounded-coverage assumption above, $d_{\hat{\mathcal{D}}_N}(x) \ge 0$, so a change of measure gives, for either relevant policy,

$$\begin{aligned}\mathcal{I}^{\tilde{\pi}} &= \int d_{\hat{\mathcal{D}}_N}(x)\frac{d\rho^{\tilde{\pi}}}{d\nu_{\hat{\mathcal{D}}_N}}(x)\,d\nu_{\hat{\mathcal{D}}_N}(x)\\ &\le C\int d_{\hat{\mathcal{D}}_N}(x)\,d\nu_{\hat{\mathcal{D}}_N}(x) = C\mathcal{I}_{\hat{\mathcal{D}}_N},\end{aligned} \tag{36}$$

where the last equality is the dataset-weighted MPIL definition in Eq. (5). Substituting Eq. (36) into Eq. (35) for the two policies in Eq. (24), and using the monotonicity of g, yields

$$\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right) \le 2\mathcal{R}_{\max}\sqrt{1-\exp\left(-\frac{C\mathcal{I}_{\hat{\mathcal{D}}_N}}{1-\gamma}\right)}. \qquad \square$$

A.3.3 Proof of Theorem 5.2

Proof. Since $1-\exp(-x) \le x$ for every $x \ge 0$, Theorem 5.1 and the uniform bound $C \le C_{\mathrm{cov}}$ imply

$$\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right) \le 2\mathcal{R}_{\max}\sqrt{\frac{C_{\mathrm{cov}}\mathcal{I}_{\hat{\mathcal{D}}_N}}{1-\gamma}}. \tag{37}$$

Taking expectations over the offline dataset and any algorithmic randomness, when applicable, and applying Jensen's inequality to the concave square-root function gives

$$\begin{aligned}\mathbb{E}_{\hat{\mathcal{D}}_N}\left[\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right)\right] &\le \frac{2\mathcal{R}_{\max}\sqrt{C_{\mathrm{cov}}}}{\sqrt{1-\gamma}}\mathbb{E}_{\hat{\mathcal{D}}_N}\left[\sqrt{\mathcal{I}_{\hat{\mathcal{D}}_N}}\right]\\ &\le \frac{2\mathcal{R}_{\max}\sqrt{C_{\mathrm{cov}}}}{\sqrt{1-\gamma}}\sqrt{\mathbb{E}_{\hat{\mathcal{D}}_N}\left[\mathcal{I}_{\hat{\mathcal{D}}_N}\right]}.\end{aligned} \tag{38}$$

Finally, substituting the assumed contraction rate $\mathbb{E}_{\hat{\mathcal{D}}_N}[\mathcal{I}_{\hat{\mathcal{D}}_N}] \le K/N$ into Eq. (38) yields

$$\mathbb{E}_{\hat{\mathcal{D}}_N}\left[\mathrm{Regret}\left(\mathcal{M}^\star, \hat{\mathcal{D}}_N\right)\right] \le 2\mathcal{R}_{\max}\sqrt{\frac{C_{\mathrm{cov}}K}{(1-\gamma)N}}, \tag{39}$$

which is Eq. (14) and completes the proof. $\square$

A.4 World Model Predictive Performance

We evaluate next-observation and reward predictions on a held-out test set. Here, the 17/11/17 prediction dimensions correspond to the original D4RL observation vectors for HalfCheetah, Hopper, and Walker2d, respectively, rather than simulator global states or inferred latent states. To further examine whether a reduction in MPIL corresponds to more accurate modeling of the environment dynamics, we compare the next-observation and reward predictions of the initial world model with those of the world model optimized under MPIL guidance on the Walker2d task. As shown in Figure 4, the dashed line $y = x$ represents ideal predictions, where the model outputs exactly match the ground-truth. For next-observation prediction, although the initial world model generally captures the trends across different observation dimensions, noticeable prediction dispersion remains in some regions. After the MPIL-guided posterior updates, the predictions across observation dimensions become more concentrated around $y = x$, with substantially reduced dispersion. A similar trend is observed for reward prediction. These results are consistent with the decreasing trend of MPIL during training, indicating that lower MPIL reflects improved characterization of next-observation and reward dynamics by the world model, thereby providing a more reliable model foundation for subsequent policy learning based on imagined trajectories.

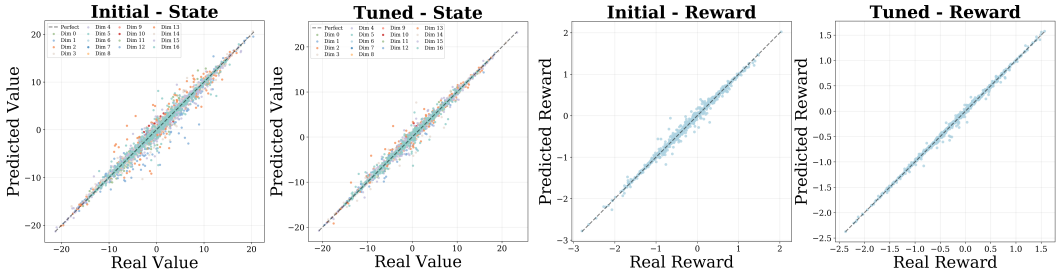
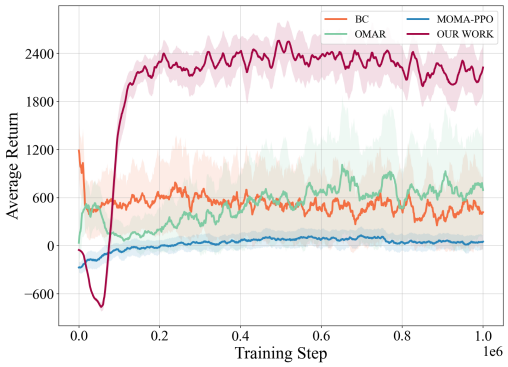


Figure 4: Comparison of World Model Predictions Before and After MPIL-Guided Optimization.

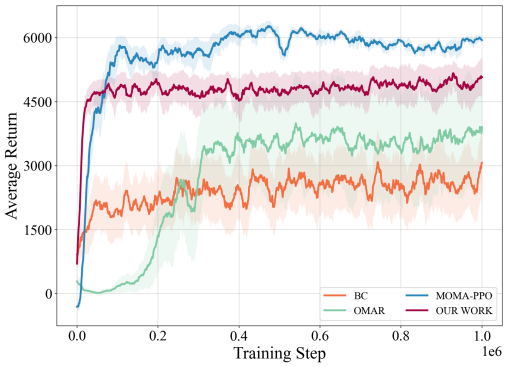
A.5 Policy Performance

Table 9: Comparison of prediction performance before and after refinement on different MuJoCo tasks.

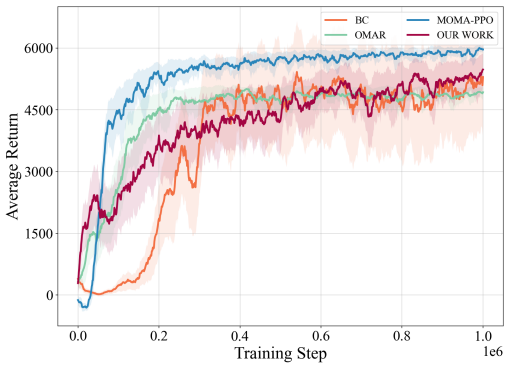
Task	Dimension	MAE_Initial	MAE_Tuned	MAE_Improvement (%)	R^2_Inital	R^2_Tuned	R^2_Improvement (%)
Halfcheetah	0	0.0029	0.0012	58.93	0.9994	0.9999	0.05
	1	0.0054	0.0022	58.69	0.9999	1.0000	0.00
	2	0.0076	0.0030	59.99	0.9978	0.9996	0.18
	3	0.0088	0.0033	62.69	0.9967	0.9995	0.28
	4	0.0101	0.0040	60.54	0.9945	0.9991	0.45
	5	0.0100	0.0045	55.14	0.9973	0.9993	0.20
	6	0.0093	0.0043	54.13	0.9972	0.9992	0.20
	7	0.0099	0.0053	46.44	0.9959	0.9985	0.26
	8	0.0641	0.0313	51.20	0.9778	0.9928	1.53
	9	0.1036	0.0499	51.86	0.9474	0.9793	3.37
	10	0.2063	0.1000	51.53	0.9483	0.9807	3.41
	11	0.2736	0.1337	51.14	0.9918	0.9973	0.55
	12	0.4100	0.1884	54.04	0.9832	0.9956	1.26
	13	0.3691	0.1684	54.38	0.9836	0.9957	1.23
	14	0.3865	0.2009	48.03	0.9871	0.9951	0.81
	15	0.4084	0.2355	42.34	0.9873	0.9932	0.60
	16	0.3936	0.2276	42.18	0.9803	0.9904	1.03
	Reward	0.0380	0.0180	52.72	0.9903	0.9973	0.71
Hopper	0	0.0054	0.0006	88.17	0.1732	0.9791	465.28
	1	0.0079	0.0007	90.73	0.8290	0.9918	19.64
	2	0.0084	0.0015	82.22	0.8145	0.9720	19.34
	3	0.0093	0.0011	88.74	0.8433	0.9921	17.64
	4	0.0219	0.0014	93.51	0.6166	0.9946	61.30
	5	0.0401	0.0051	87.41	0.9029	0.9961	10.32
	6	0.0411	0.0049	87.99	0.7564	0.9951	31.55
	7	0.1453	0.0164	88.73	0.9245	0.9960	7.74
	8	0.1492	0.0153	89.72	0.8710	0.9964	14.40
	9	0.1367	0.0135	90.11	0.8883	0.9972	12.26
	10	0.2005	0.0153	92.37	0.8473	0.9976	17.74
	Reward	0.0280	0.0024	91.50	0.9671	0.9999	3.39
Walker2d	0	0.0077	0.0023	70.09	0.7090	0.9576	35.06
	1	0.0525	0.012	77.11	0.5628	0.9598	70.55
	2	0.0451	0.0109	75.78	0.5502	0.9674	75.84
	3	0.0349	0.0129	62.93	0.8363	0.9758	16.68
	4	0.0475	0.0222	53.16	0.8854	0.9833	11.06
	5	0.0449	0.0111	75.37	0.5612	0.9656	72.04
	6	0.0342	0.0118	65.51	0.8541	0.9791	14.63
	7	0.0428	0.0211	50.75	0.9016	0.9855	9.31
	8	0.0881	0.0265	69.99	0.8335	0.9806	17.65
	9	0.1897	0.0529	72.11	0.6446	0.9651	49.73
	10	1.0088	0.2438	75.83	0.7850	0.9815	25.02
	11	1.0388	0.3302	68.21	0.8802	0.9854	11.96
	12	0.8285	0.3426	58.65	0.9159	0.9837	7.40
	13	1.4659	0.882	39.83	0.8874	0.9539	7.50
	14	0.9653	0.3115	67.73	0.8907	0.9862	10.72
	15	0.8146	0.3546	56.47	0.9144	0.9847	7.68
	16	1.3732	0.7956	42.06	0.8975	0.9590	6.85
	Reward	0.0296	0.0095	67.80	0.9897	0.9992	0.95



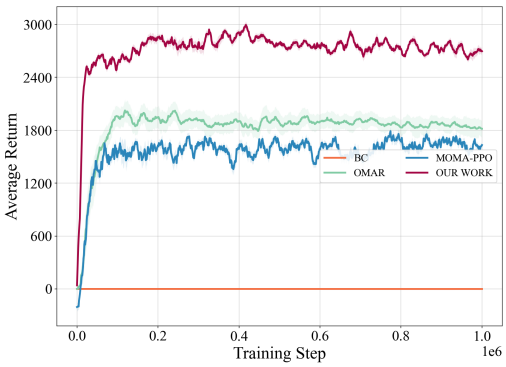
(a) HalfCheetah Expert



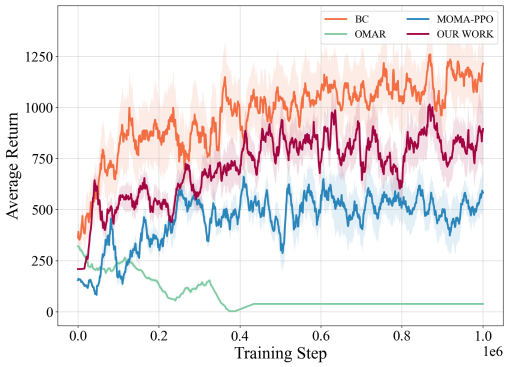
(b) HalfCheetah Full



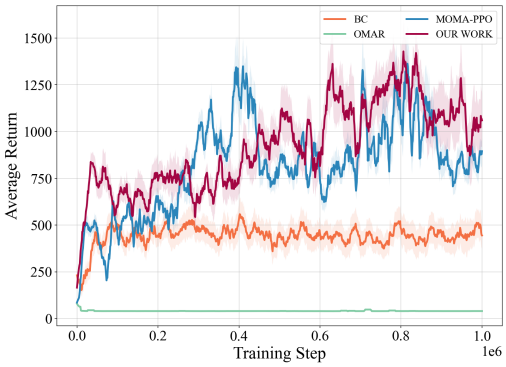
(c) HalfCheetah Medium



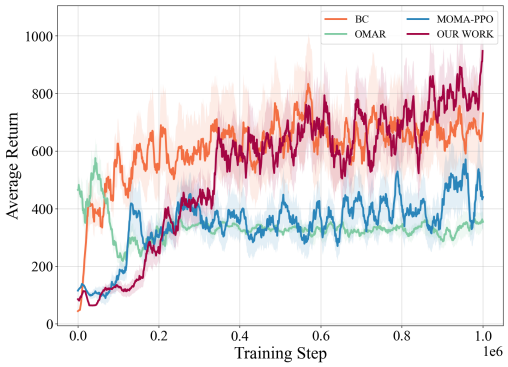
(d) HalfCheetah Random



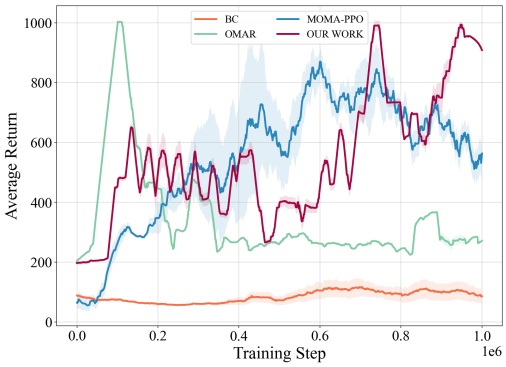
(e) Hopper Expert



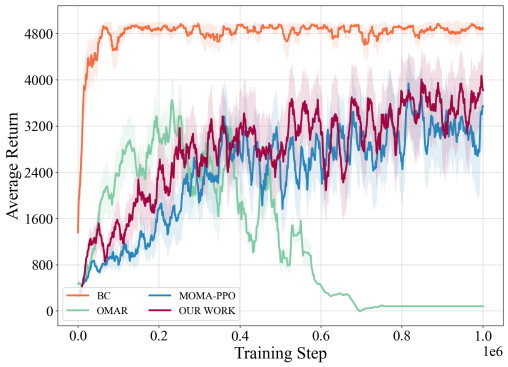
(f) Hopper Full



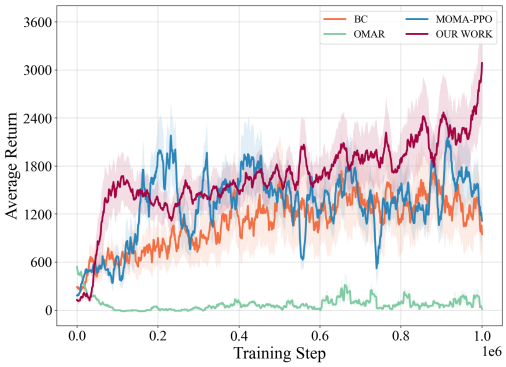
(g) Hopper Medium



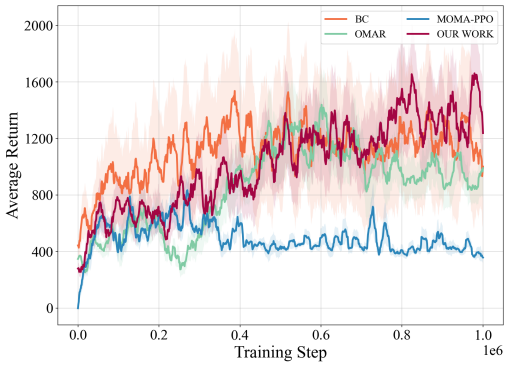
(h) Hopper Random



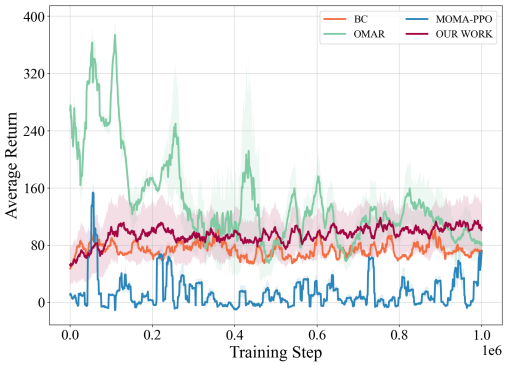
(i) Walker2d Expert



(j) Walker2d Full



(k) Walker2d Medium



(l) Walker2d Random

Figure 5: Performance comparison across different environments and dataset qualities.