SeMOPO: Learning High-quality Model and Policy from Low-quality Offline Visual Datasets

Shenghua Wan [1 2] **Ziyuan Chen** [3] **Le Gan** [1 2] **Shuai Feng** [4] **De-Chuan Zhan** [1 2]

Abstract

Model-based offline reinforcement Learning (RL) is a promising approach that leverages existing data effectively in many real-world applications, especially those involving high-dimensional inputs like images and videos. To alleviate the distribution shift issue in offline RL, existing model-based methods heavily rely on the uncertainty of learned dynamics. However, the model uncertainty estimation becomes significantly biased when observations contain complex distractors with non-trivial dynamics. To address this challenge, we propose a new approach - *Separated Model-based Offline Policy Optimization* (SeMOPO) - decomposing latent states into endogenous and exogenous parts via conservative sampling and estimating model uncertainty on the endogenous states only. We provide a theoretical guarantee of model uncertainty and performance bound of SeMOPO. To assess the efficacy, we construct the Low-Quality Vision Deep Data-Driven Datasets for RL (LQV-D4RL), where the data are collected by non-expert policy and the observations include moving distractors. Experimental results show that our method substantially outperforms all baseline methods, and further analytical experiments validate the critical designs in our method. The project website is https://sites.google.com/view/semopo.

[1]School of Artificial Intelligence, Nanjing University, China [2]National Key Laboratory for Novel Software Technology, Nanjing University, China [3]School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing, China [4]School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China. Correspondence to: De-Chuan Zhan <zhandc@nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Offline reinforcement learning (RL) (Lange et al., 2012; Levine et al., 2020), which learns policies from fixed datasets without the need for costly interactions with online environments, has been increasingly applied in real-world tasks such as drug discovery (Trabucco et al., 2022) and autonomous driving (Iroegbu & Madhavi, 2021; Diehl et al., 2023). Offline RL saves the cost of interacting with the environment and improves sample efficiency. Real-world RL datasets typically exhibit two main characteristics: (1) they are often collected by non-expert or random policies (Jin et al., 2020; Rashidinejad et al., 2021), and (2) they originate from real environments with high-dimensional observations, such as images or videos (Rafailov et al., 2021; Prudencio et al., 2022), containing complex noise like moving backgrounds (Lu et al., 2022). Learning high-quality policies from low-quality datasets poses a significant challenge.

Concerning the first characteristic, previous research has demonstrated the efficacy of offline RL with highly suboptimal or random datasets (Jin et al., 2020; Rashidinejad et al., 2021; Kumar et al., 2022). Regarding the second characteristic, some studies have employed model-based RL (MBRL) methods to address the challenges of high-dimensional inputs. MBRL learns a low-dimensional surrogate model of the high-dimensional environment (Ha & Schmidhuber, 2018; Hafner et al., 2019), allowing the agent to interact with this model to gather additional trajectories in the low-dimensional state space. Offline MBRL methods improve sample efficiency and reduce storage costs. However, an unavoidable gap exists between the actual environment and the learned model from offline datasets. Offline MBRL methods (Yu et al., 2020) mitigate the distribution shift problem by incorporating the model prediction's uncertainty as a penalty in the reward function.

In real-world decision-making scenarios, uncertainty arises not only from task-relevant dynamics but also from irrelevant distractors in observations, such as moving backgrounds. However, previous offline visual RL works (Yu et al., 2020; Rafailov et al., 2021; Lu et al., 2022) do not differentiate between these two types of uncertainty. If both are treated as model uncertainty in offline policy training and used as a penalty in the reward function, the learned policies

may become overly conservative. Therefore, it is crucial to consider task-irrelevant dynamics uncertainty during offline model training.

To address these challenges, we propose the Separated Model-based Offline Policy Optimization (SeMOPO) method. We first analyze the performance lower bound under the Exogenous Block MDP (EX-BMDP) assumption (Efroni et al., 2021) in the case of offline learning from noisy visual datasets. We find that such a lower bound is empirically tighter than that under the POMDP assumption (Kaelbling et al., 1998) commonly used in previous offline RL research (Rafailov et al., 2021; Lu et al., 2022). We replace the typical random sampling method in offline MBRL (Yu et al., 2020; Rafailov et al., 2021; Lu et al., 2022) with our proposed conservative sampling method, training the separated model on trajectories collected by relatively deterministic behavior policies. After obtaining the task-relevant model from the offline dataset, we train the policy on the endogenous states imagined by this model. To evaluate our method, we construct the Low-Quality Vision Datasets for Deep Data-Driven RL (LQV-D4RL), including 15 different settings from DMControl Suite (Tassa et al., 2018) and Gym (Brockman et al., 2016) environments. SeMOPO achieves significantly better performance than all baseline methods on the LQV-D4RL. Our analytical experiments confirm the effectiveness of the conservative sampling method in identifying task-relevant information and the superiority of estimating uncertainty estimation on it during offline policy training.

The main contributions of our work are: (i) We propose a new approach named Separated Model-based Offline Policy Optimization (SeMOPO), which aims to solve offline visual RL tasks from low-quality datasets collected by sub-optimal policies and with complex distractors in observations. (ii) To establish a benchmark for the practical setting, we construct Low-Quality Vision Datasets for Deep Data-Driven RL (LQV-D4RL), which offers new research opportunities for practitioners in the field. (iii) We provide a theoretical analysis of the lower performance bound of policies learned on the endogenous state space and the superiority of our proposed conservative sampling method in differentiating task-relevant and irrelevant information. (iv) We show excellent performance of SeMOPO on LQV-D4RL, with analytical experiments validating the efficacy of each component of the method.

2. Preliminaries

In our work, we focus on learning skills from the offline dataset consisting of image observations, actions, and rewards, denoted as $\mathcal{B} = \{o^i_{1:T}, a^i_{1:T}, r^i_{1:T}\}^n_{i=1}$. The dataset contains moving distractors within the visual observations and is collected by policies π_B with suboptimal or random behaviors. To better model the environment, we consider the Exogenous Block Markov Decision Process (EX-BMDP) setting (Efroni et al., 2021), an adaptation of the Block MDP (Du et al., 2019). A Block MDP consists of a set of observations $\mathcal{O}$; a set of latent states, $\mathcal{Z}$ with cardinality Z; a finite set of actions, $\mathcal{A}$ with cardinality A; a transition function, $T : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$; a reward function $R : \mathcal{O} \times \mathcal{A} \to [0, 1]$; an emission function $\mathcal{U} : \mathcal{Z} \to \Delta(O)$; and an initial state distribution $\mu_0 \in \Delta(\mathcal{Z})$. The agent has no access to the latent states but can only receive the observations. The block structure holds if the support of the emission distributions of any two latent states are disjoint, $\text{supp}(\mathcal{U}(\cdot|z_1)) \cap \text{supp}(\mathcal{U}(\cdot|z_2)) = \emptyset$ when $z_1 \neq z_2$, where $\text{supp}(\mathcal{U}(\cdot|z)) = \{o \in \mathcal{O} | \mathcal{U}(o|z) > 0\}$, distinguishing the BMDP from the Partially Observable MDP (Kaelbling et al., 1998). We now restate the definition of EX-BMDP:

Definition 2.1. (Exogenous Block Markov Decision Process). An EX-BMDP is a BMDP such that the latent state can be decoupled into two parts $z = (s^+, s^-)$ where $s^+ \in \mathcal{S}^+$ is the endogenous state and $s^- \in \mathcal{S}^-$ is the exogenous state. For $z \in \mathcal{Z}$ the initial distribution and transition functions are decoupled, that is: $\mu(z) = \mu(s^+)\mu(s^-)$, and $T(z'|z, a) = T(s^{+\prime}|s^+, a)T(s^{-\prime}|s^-)$

EX-BMDP decomposes the dynamics and separates the exogenous noise from the endogenous state. This noise is not controlled by the agent, but it may have a non-trivial dynamic. EX-BMDP provides a natural way to characterize this type of noise.

Model-based Offline Policy Optimization (Yu et al., 2020) (MOPO) is a typical offline RL method, showing the efficacy of model-based methods to learn policies from offline datasets. MOPO learns the dynamics model $\widehat{T}(\cdot|s, a)$ and the reward model $\hat{r}(s, a)$ from the fixed dataset $\mathcal{B}$, and constructs an uncertainty-penalized MDP whose reward function is defined as $\tilde{r}(s, a) = r(s, a) - u(s, a)$, where $u(s, a)$ is the model uncertainty estimation. MOPO proves that optimizing the policy in the uncertainty-penalized MDP with an admissible uncertainty estimation $u(s, a)$ is equivalent to optimizing it in the true MDP. We adopt a similar pipeline to MOPO to learn the model and policy from offline datasets.

3. SeMOPO: Separated Model-based Offline Policy Optimization

Model-based offline RL methods use model uncertainty to address the distribution shift problem between the online environment and the offline dataset. In Section 3.1, we first provide a theoretical justification of the performance bound under the Exogenous Block MDP assumption when learning from offline datasets with visual distractors and illustrate the inherent limitations of employing the POMDP framework in this scenario. In Section 3.2, we propose a sampling strategy

to help the model differentiate task-relevant and irrelevant components and give a corresponding theoretical analysis. In Section 3.3, we detail a practical implementation of SeMOPO, which integrates these advancements.

3.1. Uncertainty Estimation and Performance Bound under Exogenous Block MDP

In this section, we theoretically analyze the lower bound of policy performance under EX-BMDP. We also provide empirical evidence that, in some specific cases, the uncertainty under EX-BMDP is lower than that under POMDP, resulting in a tighter lower performance bound. We start our analysis by extending the well-known telescoping lemma (Luo et al., 2019) from the state space $\mathcal{S}$ to the endogenous state space $\mathcal{S}^+$:

Lemma 3.1. *(Telescoping lemma in the endogenous state space). Let M and $\widetilde{M}$ be two MDPs with the same reward function r, but different dynamics T and $\widetilde{T}$ respectively. Let $G^{\pi}_{\widetilde{M}}(s^+, a) := \mathop{\mathbb{E}}_{s^{+\prime}\sim\widetilde{T}(s^+,a)}\left[V^{\pi}_{M}\left(s^{+\prime}\right)\right] - \mathop{\mathbb{E}}_{s^{+\prime}\sim T(s^+,a)}\left[V^{\pi}_{M}\left(s^{+\prime}\right)\right]$. Then,*

$$\eta_{\widetilde{M}}(\pi) - \eta_M(\pi) = \gamma \mathop{\overline{\mathbb{E}}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[G^{\pi}_{\widetilde{M}}(s^+, a)\right]$$

Note that if $\mathcal{F}$ is a set of functions mapping $\mathcal{S}^+$ to $\mathbb{R}$ that contains V^{π}_{M} for all π, then,

$$\begin{aligned}
&|G^{\pi}_{\widetilde{M}}(s^+, a)| \\
&\leq \sup_{f\in\mathcal{F}}\left|\mathbb{E}_{s^{+\prime}\sim\widetilde{T}(s^+,a)}[f(s^{+\prime})] - \mathbb{E}_{s^{+\prime}\sim T(s^+,a)}[f(s^{+\prime})]\right| \\
&=: d_{\mathcal{F}}(\widetilde{T}(s^+, a), T(s^+, a))
\end{aligned}$$

where $d_{\mathcal{F}}$ is the integral probability metric (IPM) defined by $\mathcal{F}$. The telescoping lemma provides a way to measure the performance gap between policies in the true dynamics T and the estimated dynamics $\widetilde{T}$.

Assumption 3.2. Let $\mathcal{S}^+$ be the endogenous state space under the EX-BMDP. We say $\tilde{u} : \mathcal{S}^+ \times \mathcal{A} \to \mathbb{R}$ is an admissible error estimator for $\widetilde{T}$ if $d_{\mathcal{F}}(\widetilde{T}(s^+, a), T(s^+, a)) \leq \tilde{u}(s^+, a)$ for all $s \in \mathcal{S}^+, a \in \mathcal{A}$.

Theorem 3.3. *Under Assumption 3.2, we can define the uncertainty estimation $\epsilon_{\tilde{u}}(\pi)$ under the EX-BMDP as $\epsilon_{\tilde{u}}(\pi) := \mathop{\overline{\mathbb{E}}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}[\tilde{u}(s^+, a)]$. Let $\tilde{\pi}$ denote the learned optimal policy under the endogenous MDP, then,*

$$\eta_M(\tilde{\pi}) \geq \sup_{\pi}\{\eta_M(\pi) - 2\lambda\epsilon_{\tilde{u}}(\pi)\}$$

Theorem 3.3 suggests that the performance under the true MDP $\mathcal{M}$ can be bounded by learning a policy in the endogenous state space using the reward penalized by the corresponding model uncertainty, which provides a theoretical guarantee on the performance of SeMOPO. The proof is in Appendix A.1.

Remarks. The Partially Observable MDP used in prior works like LOMPO (Rafailov et al., 2021) and Offline DV2 (Lu et al., 2022) will give a biased model uncertainty estimation when learning from offline datasets containing task-irrelevant distractors in observations. These methods assume that $u : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ can serve as an admissible error estimator for $\widehat{T}$ if $d_{\mathcal{F}}(\widehat{T}(z, a), T(z, a)) \leq u(z, a)$ for all z and a, where z represents the belief state of the true state s within the latent space $\mathcal{Z}$. However, in addition to task-relevant information, there are task-irrelevant components in image observations that are beyond the agent's control. Without excluding them during model learning, these components will be absorbed into z, leading to biased uncertainty estimation based on $d_{\mathcal{F}}(\widehat{T}(z, a), T(z, a))$. Taking LOMPO as an example, the model uncertainty is estimated as the variance of the predicted probability of the next state given the current state and action, i.e., $\mathrm{Var}\{\log \widehat{T}_{\theta_i}(z_{t+1}|z_t, a_t)\}$, which is larger than the model uncertainty $\mathrm{Var}\{\log \widetilde{T}_{\theta_i}(s^+_{t+1}|s^+_t, a_t)\}$ under the EX-BMDP using the same calculation method. LOMPO then uses this model uncertainty as a reward penalty, leading to a loser lower performance bound than the EX-BMDP.

3.2. Model Training with Conservative Sampling

In Section 3.1, we know that a separated model under the EX-BMDP assumption benefits policy learning. In the following, we focus on how to learn such a separated model. Under the EX-BMDP assumption, the log-likelihood for the trajectory $\tau = \{o_1, a_1, \cdots, o_T, a_T\}$ can be decomposed as:

$$\begin{aligned}
\ln p(\tau) = \sum_{t=1}^{T}\Big[&\ln p(o_t|s^+_t, s^-_t) + \ln p(a_t|s^+_t) \\
&+ \ln p(s^+_t|s^+_{t-1}, a_{t-1}) + \ln p(s^-_t|s^-_{t-1})\Big],
\end{aligned} \tag{1}$$

We omit the terms of the initial distribution of s^+_1 and s^-_1 for convenience. The detailed derivation is in Appendix B.

From the last three terms in Equation (1), we can find that the agent's action is the key factor to help the model distinguish the task-relevant and irrelevant components. An intuitive insight is that the separated model should be updated using relatively deterministic actions. Excessively random action distributions can weaken the influence of agent actions on the endogenous transition, causing it to collapse into $p(s^+_{t+1}|s^+_t)$, making it indistinguishable from the exogenous transition. It naturally raises a question: how to sample the offline data to help decompose the endogenous and exogenous dynamics? Consider two kinds of training processes, the first is maximizing the likelihood where the data is sampled from all the mixed trajectories without re-

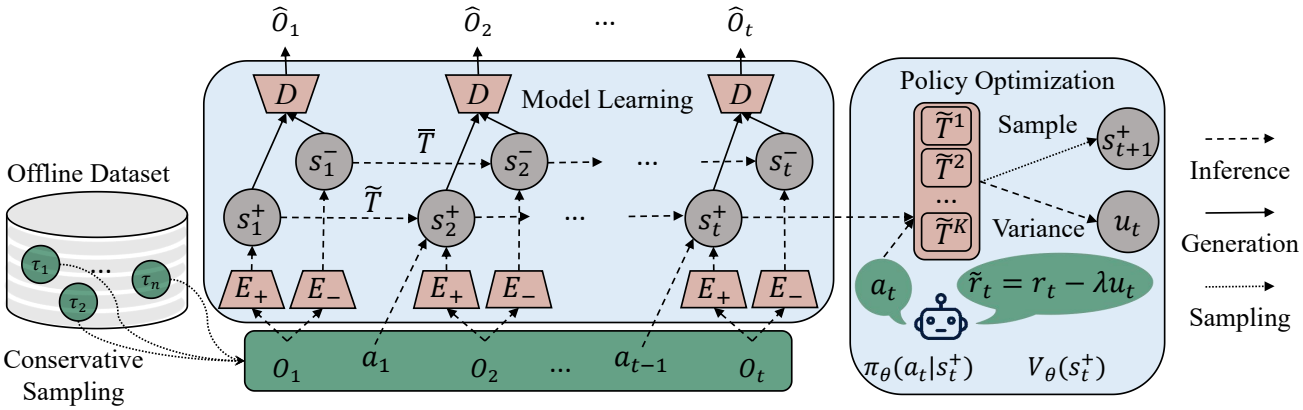

Figure 1. The SeMOPO framework encompasses two parts: model learning and policy optimization. In the model learning phase, SeMOPO employs conservative sampling to select trajectories, which are then used to train models for endogenous and exogenous dynamics, each implemented as an ensemble of Gaussian distributions. During policy optimization, SeMOPO trains a policy $\pi_\theta(a_t|s_t^+)$ and a value model $V_\theta(s_t^+)$ based on the endogenous states generated by a sampled endogenous dynamics model $\tilde{T}^j$. SeMOPO uses the reward penalized by the variance of the endogenous dynamics models' predictions to train the value model.

placing (*Random Sampling*), and the second is maximizing the likelihood where data is sampled trajectory by trajectory (*Conservative Sampling*). To formulate the relationship between these two likelihoods, we introduce Theorem 3.4.

Theorem 3.4. *Consider the likelihood optimization problem on the same offline dataset $\mathcal{B}$ but with two different sampling methods. let $\mathcal{B}_{\pi_i}$ be the dataset collected by the behavior policy π_i, where $i = 1, 2, \cdots, n$. $\mathcal{B}_{\pi_{mix}} = \mathcal{B}_{\pi_1} \cup \mathcal{B}_{\pi_1} \cup \cdots \cup \mathcal{B}_{\pi_n}$ is the mixture of the datasets of all policies. Then we have*

$$\mathbb{E}_{\tau \in \mathcal{B}_{\pi_{mix}}} \ln p(\tau) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\tau \in \mathcal{B}_{\pi_i}} \ln p(\tau),$$

where $p(\tau)$ is the true density of τ.

Theorem 3.4 suggests that the total likelihood estimated on the sampled offline trajectory of a certain policy is larger than on the mixture dataset of all policies. The proof is in Appendix A.2. Maximizing the likelihood will make the estimated distribution toward the true distribution in Equation (1), which helps distinguish the endogenous and exogenous transitions. Notably, Theorem 3.4 does not guarantee that we can obtain the maximum likelihood via the conservative sampling method. Even so, we empirically show that the model can separate the task-relevant and irrelevant components well compared to the random sampling method.

3.3. Practical Implementation of SeMOPO

Based on the above analysis, we present a practical implementation of *Separated Model-based Offline Policy Optimization*. The overall method of SeMOPO is shown in Figure 1 and summarized in Algorithm 1.

Separated Model Learning. Theorem 3.3 allows us to achieve a significantly improved lower bound of the total return if we can learn a separate model and train the policy in the endogenous state space. By bifurcating the latent state z into endogenous (s^+) and exogenous (s^-) parts, we obtain the following Evidence Lower Bound (ELBO):

$$\begin{aligned}\max_\theta \mathbb{E}\Bigg[\sum_{t=1}^{T} & \ln \mathcal{U}_\theta(o_t|s_t^+, s_t^-) \\ & - \mathbb{D}_{KL}\big(\bar{q}_\theta(s_t^-|o_{\leq t})||\overline{T}_\theta(s_t^-|s_{t-1}^-)\big) \\ & - \mathbb{D}_{KL}\big(\tilde{q}_\theta(s_t^+|o_{\leq t}, a_{<t})||\widetilde{T}_\theta(s_t^+|s_{t-1}^+, a_{t-1})\big)\Bigg].\end{aligned}$$

where $\tilde{q}_\theta$ and $\bar{q}_\theta$ represent the inference models for the endogenous and exogenous states, respectively. Likewise, $\widetilde{T}_\theta$ and $\overline{T}$ denote the corresponding transition dynamics models, and $\mathcal{U}_\theta$ is the observation model which reconstructs the observation jointly from the endogenous and exogenous states. The derivation of the ELBO is detailed in Appendix B, with the implementation specifics of each model described in Appendix D.

For the conservative sampling strategy outlined in Section 3.2, we design a simple but effective implementation. In the m-th training epoch, the SeMOPO's model is only trained on the sampled trajectory τ_j generated by a certain policy, where $j \leq \min(m, n)$ and n is the number of trajectories in the offline dataset. We provide an intuitive interpretation for this implementation: it forces the model to separate the task-relevant and irrelevant dynamics in the early training stages ($m \leq n$); as training progresses ($m > n$), the model is trained on the trajectories sampled from the entire dataset to enhance the coverage of transitions.

Table 1. Normalized test returns of SeMOPO and compared baselines on the LQV-D4RL benchmark. Mean scores (higher is better) with standard deviation are recorded across 4 seeds for each task. The original returns are shown in Table 5.

Dataset		SeMOPO	Offline DV2	LOMPO	DrQ+BC	DrQ+CQL	BC	InfoGating
Walker Walk	random	**0.77** ± **0.06**	0.27 ± 0.05	0.22 ± 0.06	0.04 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.07 ± 0.00
	medrep	**0.87** ± **0.06**	0.29 ± 0.04	0.36 ± 0.11	0.04 ± 0.01	0.03 ± 0.01	0.04 ± 0.01	0.09 ± 0.03
	medium	0.45 ± 0.07	0.11 ± 0.04	0.10 ± 0.02	0.65 ± 0.06	0.03 ± 0.01	**0.69** ± **0.05**	0.16 ± 0.06
Cheetah Run	random	**0.63** ± **0.07**	0.10 ± 0.03	0.16 ± 0.04	0.23 ± 0.08	0.00 ± 0.00	0.05 ± 0.05	0.14 ± 0.03
	medrep	0.64 ± 0.07	0.16 ± 0.07	0.19 ± 0.08	0.41 ± 0.23	0.00 ± 0.00	0.05 ± 0.05	**0.66** ± **0.12**
	medium	**0.73** ± **0.08**	0.20 ± 0.14	0.13 ± 0.09	0.64 ± 0.07	0.00 ± 0.00	0.62 ± 0.10	0.71 ± 0.09
Hopper Hop	random	0.68 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.10	0.00 ± 0.00	0.06 ± 0.08	**0.79** ± **0.13**
	medrep	**0.91** ± **0.07**	0.00 ± 0.00	0.00 ± 0.00	0.25 ± 0.18	0.00 ± 0.00	0.04 ± 0.02	0.53 ± 0.16
	medium	**1.24** ± **0.16**	0.02 ± 0.05	0.01 ± 0.04	0.81 ± 0.19	0.00 ± 0.00	0.42 ± 0.07	0.58 ± 0.09
Humanoid Walk	random	**0.01** ± **0.00**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	medrep	0.01 ± 0.01	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**0.02** ± **0.01**	0.00 ± 0.00
	medium	0.01 ± 0.01	0.01 ± 0.00	0.00 ± 0.00	**0.02** ± **0.01**	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
Car Racing	random	**0.93** ± **0.16**	0.51 ± 0.18	0.86 ± 0.18	-0.05 ± 0.05	-0.23 ± 0.02	-0.15 ± 0.01	-0.20 ± 0.01
	medrep	**0.80** ± **0.17**	0.39 ± 0.18	0.72 ± 0.40	-0.18 ± -0.01	-0.23 ± 0.02	-0.21 ± 0.02	-0.20 ± 0.01
	medium	**0.90** ± **0.34**	0.62 ± 0.32	0.69 ± 0.22	0.39 ± 0.21	-0.21 ± 0.02	-0.18 ± 0.00	-0.18 ± 0.01

Endogenous Model-based Offline Policy Optimization. We train a policy π_θ and a value model V_θ in the endogenous state space of the learned model, following the standard training algorithm in DreamerV2 (Hafner et al., 2021). To address the distribution shift issue, we estimate the model uncertainty through the discrepancy $d_{\mathcal{F}}(\widetilde{T}(s^+, a), T(s^+, a))$. Since we can not access the true dynamics T, we adopt a widely-used approach based on the model disagreements (Pathak et al., 2019), which is implemented as a penalty term:

$$\tilde{r}(s^+, a) = r(s^+, a) - \lambda \sum_{i=1}^{K} (\mu^i(s^+, a) - \bar{\mu}(s^+, a))^2 \quad (2)$$

Here, λ is a coefficient for adjusting the penalty weight, and $\bar{\mu}(s^+, a) = \frac{1}{K}\sum_{i=1}^{K} \mu^i(s^+, a)$ is the average prediction across an ensemble of K endogenous dynamics models.

4. Experiments

We conduct several experiments to answer the following scientific questions: (1) Can SeMOPO outperform the existing methods with low-quality offline visual datasets? (2) Can SeMOPO give a reasonable model uncertainty estimation? (3) How does the sampling strategy for policy data affect the separated model training? (4) Can SeMOPO generalize to online environments with different distractors?

Datasets. To evaluate the efficacy of SeMOPO in the scenario of learning with low-quality offline visual datasets, we construct a dataset named LQV-D4RL (Low-Quality Vision Datasets for Deep Data-Driven RL). A "low-quality" dataset, in this context, refers to one where the data collection policy is either sub-optimal or randomly initialized, and the observations contain complex distractors with non-trivial dynamics. With these considerations in mind, we create nine distinct subsets for evaluation. We select four locomotion tasks — *Walker Walk*, *Cheetah Run*, *Hopper Hop*, and *Humanoid Walk* — from the DMControl Suite (Tassa et al., 2018), and the *Car Racing* task from Gym (Brockman et al., 2016). Each task is represented across three different levels of policy performance:

- **random**: Trajectories collected by randomly initialized policies.
- **medium_replay** (medrep): Trajectories drawn from the replay buffer accumulated during the training of a policy with medium performance.
- **medium**: Trajectories collected by a fixed policy of medium performance.

The backgrounds of each locomotion task's observations are replaced with videos from the "driving car" category of the Kinetics dataset (Kay et al., 2017), as used in DBC (Zhang et al., 2021). The size of image observation is $64 \times 64 \times 3$. To mimic the real data collection process in natural settings, we train policies and then collect trajectories based on image observations with the aforementioned distractors. Further details about the dataset can be found in Appendix C.

Baselines. We compare several representative methods in offline visual RL literature, dividing them into model-based and model-free categories. Model-based methods like **LOMPO** (Rafailov et al., 2021) introduce a penalty term in the reward function based on model disagreements and optimize the policy in its constructed latent MDP, while **Offline DV2** (Lu et al., 2022) adapts the DreamerV2 method

with a similar penalty for offline settings. Both of these two methods show great performance in offline visual RL tasks. In model-free approaches, Behavioral Cloning (BC) (Bain & Sammut, 1995; Bratko et al., 1995) learns by imitating the behavior of the policy that collected the data, and Conservative Q-Learning (CQL) (Kumar et al., 2020) samples actions from a broad distribution while penalizing those that fall outside the support region of the offline data. Remarkably, both BC and CQL are not originally tailored for scenarios involving high-dimensional image inputs. To address this, DrQ-v2's regularization techniques (Yarats et al., 2022) are applied to BC and CQL in (Lu et al., 2022), creating **DrQ+BC** and **DrQ+CQL** methods. These modified methods, alongside the original **BC**, are compared in our study to evaluate their effectiveness in offline visual RL tasks with low-quality datasets. Additionally, we compare the offline version of the **InfoGating** method (Tomar et al., 2024), a visual reinforcement learning approach that removes task-irrelevant noise by minimizing the information required for the task. We run all experiments on four seeds and report the normalized test return after training. The normalized return is obtained by normalizing the original return with the maximum and minimum values of the three levels of datasets for each task. The detailed calculation procedure is in Appendix D.2.

4.1. Evaluation on the LQV-D4RL Benchmark

We evaluate SeMOPO against various baselines in nine scenarios within the LQV-D4RL benchmark. The results in Table 1 consistently demonstrate SeMOPO's superior performance across diverse datasets, confirming the effectiveness of uncertainty estimation in the endogenous state space, especially in low-quality visual datasets. Significantly, SeMOPO outperforms model-free approaches in nearly all environments. However, BC-based approaches also perform well when the behavior policy of the dataset is reasonably effective. Model-based methods, Offline DV2 and LOMPO, show improved results when trained on random and medium_replay datasets compared to medium datasets. This indicates that trajectories with random behaviors may provide a wider range of transitions, which benefits mitigating the distribution shift problem. This phenomenon aligns with previous research (Jin et al., 2020; Rashidinejad et al., 2021; Kumar et al., 2022) and holds even in environments with complex noise in image inputs. Moreover, our findings highlight that the DrQ+BC and BC methods outperform Offline DV2 and LOMPO in several settings. This implies that inaccuracies in model uncertainty estimations can lead to worse performance of model-based methods than direct policy imitation. SeMOPO and all baseline methods have failed in the humanoid walk task. The visual reconstruction on our project website demonstrates that SeMOPO can effectively extract task information. Thus, task failure may

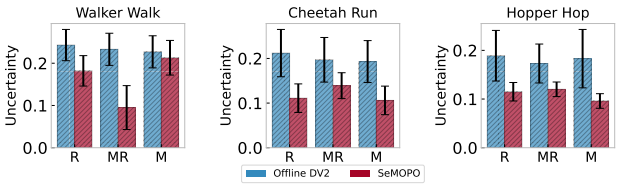

Figure 2. The model uncertainty estimation of SeMOPO and Offline DV2 on the LQV-D4RL dataset. We randomly select 1000 states and report the mean and standard deviation of uncertainty on them.

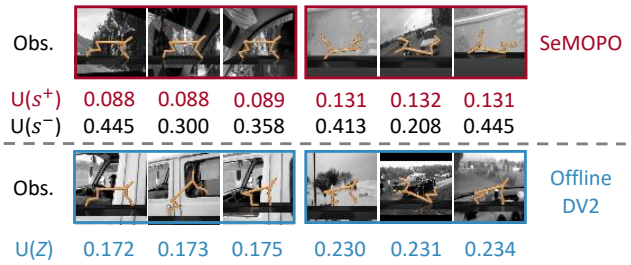

Figure 3. Examples of uncertainty estimated by SeMOPO and Offline DV2. $U(s^+)$, $U(s^-)$, and $U(Z)$ represent uncertainty estimations for the endogenous state, the exogenous state, and the belief latent state, respectively.

be attributed to the inherently complex nature of the humanoid task, which involves controlling a high-dimensional action space of up to 21 dimensions. Particularly in our experiments, the challenge is further compounded by the high-dimensional image inputs and complex moving distractors in the background, significantly increasing learning difficulty. We anticipate that future researchers will focus on learning a high-performance policy for this task within noisy visual inputs. In addition, we assess these methods on the V-D4RL benchmark, as detailed in Appendix F.2, where observations do not include distractors.

4.2. Can SeMOPO give a reasonable model uncertainty estimation?

To answer the question, we compare the model uncertainty of SeMOPO and Offline DV2 on randomly selected states, as shown in Figure 2. We find that SeMOPO exhibits lower model uncertainty than Offline DV2 across nine datasets, confirming the conclusion of Theorem 3.3. Given that SeMOPO and Offline DV2 employ the same method for uncertainty calculation, the reduced uncertainty observed in SeMOPO can be ascribed to the endogenous state transition it introduced. We also show the uncertainty of exogenous state transitions in Appendix F.1.

To further illustrate the uncertainty estimation on unique

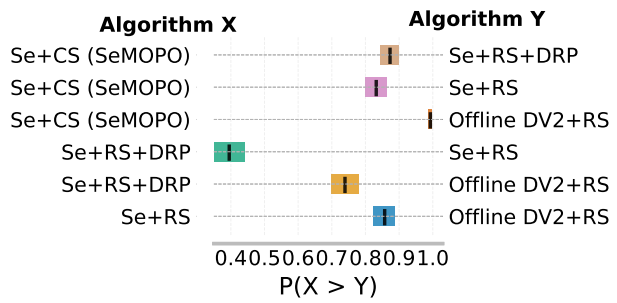

Figure 4. The performance comparison for ablated methods. Each row represents the comparative performance probabilities, complete with 95% bootstrap confidence intervals, suggesting that Algorithm X is superior to Algorithm Y (Agarwal et al., 2021). These probabilities are derived from 50 runs of 4 seeds for every task to ensure robustness in the evaluation. We show the aggregated results for all nine tasks.

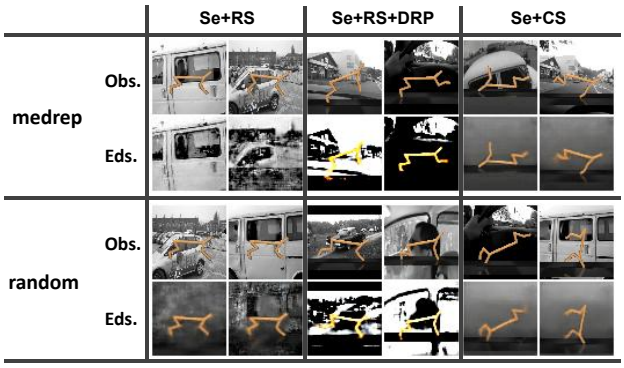

Figure 5. Original observations (Obs.) and image reconstructions from the endogenous states (Eds.) of different ablated methods in the medium_replay and random datasets of the *Cheetah Run* task. SeMOPO (Se+CS) can preserve task-relevant information well, while others can not.

states, we capture frames from the *Cheetah Run* task and record the corresponding estimated uncertainty of SeMOPO and Offline DV2 in Figure 3. SeMOPO displays lower model uncertainty on normal states as opposed to unpredictable abnormal states and shows different uncertainty of the exogenous states based on background disturbances. Offline DV2 provides similar uncertainty estimates for states with comparable background distractors but significantly different agent movements (illustrated in the three frames at the lower left corner of Figure 3). This occurs because Offline DV2 does not distinctly differentiate between endogenous and exogenous states, instead directly estimating model uncertainty on the latent states. The analysis of specific states reveals that SeMOPO can achieve reasonable model uncertainty estimation.

4.3. How does the sampling strategy for policy data affect the separated model training?

To validate the efficacy of our proposed sampling method for training the separated model, we conduct a comparison involving several ablated methods: 1) Random sampling of all policy trajectories (Se + RS); 2) Random sampling combined with additional dissociated prediction loss[1] to hinder predicting the true reward from the exogenous state (Se + RS + DRP), as used in (Fu et al., 2021); 3) Our conservative sampling method (Se + CS (SeMOPO)). The "Se" indicates the application of these methods to the training of the separated model. We also compare these with the original Offline DV2 method trained under the random sampling (Offline DV2 + RS).

[1]More information about dissociated reward prediction loss is in Appendix D.3.

As illustrated in Figure 4, SeMOPO demonstrates superior performances over other ablated methods, with a high probability (> 0.8). There is no significant difference in performance between Se+RS and Se+RS+DRP (≈ 0.4), suggesting that the model trained by dissociated reward prediction may not benefit offline policy learning. Both methods, however, outperform the original Offline DV2, indicating the effectiveness of the separated model in addressing offline visual RL challenges with noisy observations. To further analyze these sampling strategies, we visualize the observation reconstruction of the endogenous state for each training method in Figure 5. The model trained with random sampling does not guarantee comprehensive task-related information in the endogenous state, especially in the medium_replay dataset. While adding dissociated reward prediction loss to random sampling enriches task information in the endogenous state, it also introduces considerable distraction. Our conservative sampling method effectively reconstructs task-related information, ensuring the endogenous state contains only task-relevant details.

To corroborate the theoretical claims made in Section 3.2, we compute the action entropy for the first 30 epochs using both RS and CS methods across all nine datasets. As shown in Figure 6, the CS method consistently results in lower action entropy compared to RS across different environments and datasets. Lower action entropy implies more deterministic policy behavior, aiding in distinguishing endogenous states from exogenous noise. Due to little action variations among different policies in the medium dataset, the difference in action entropy between RS and CS is less than the random and medium_replay datasets. It explains why SeMOPO shows a greater advantage in the latter datasets, as described in Table 1.

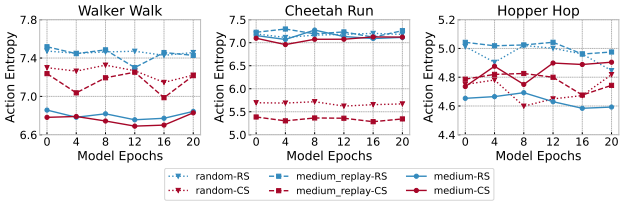

Figure 6. Action entropy of sampled data by random sampling (RS) and conservative sampling (CS). We record values of the first 30 training epochs across nine datasets of LQV-D4RL. The conservative sampling yields data with significantly lower action entropy than random sampling.

Table 2. Normalized test returns on tasks with different online distractions from offline datasets.

Dataset		SeMOPO	Offline DV2	DrQ+BC
Walker Walk	random	**0.78** ± **0.14**	0.33 ± 0.05	0.05 ± 0.01
	medrep	**0.50** ± **0.02**	0.30 ± 0.05	0.05 ± 0.02
Cheetah Run	random	**0.65** ± **0.05**	0.16 ± 0.04	0.45 ± 0.05
	medrep	**0.59** ± **0.06**	0.17 ± 0.06	0.49 ± 0.14

4.4. Can SeMOPO generalize to online environments with different distractors?

In this section, we evaluate the effectiveness of SeMOPO in addressing the "offline-to-online distraction gap" challenge. We alter the online testing environment by introducing an RGB background, starkly contrasting with the grayscale background of the LQV-D4RL dataset. To assess the methods' adaptability to distinct online visual distractions, we compared SeMOPO with Offline DV2 and DrQ+BC in two distinct environments: *Walker Walk* and *Cheetah Run*, each with two datasets - "random" and "medrep". As shown in Table 2, Offline DV2 performs better than DrQ+BC in the *Walker Walk* task, but the reverse is true for the *Cheetah Run* task. However, both are outperformed by SeMOPO, demonstrating SeMOPO's clear advantage in managing the distraction gap.

5. Related Work

Offline RL. Offline RL (Lange et al., 2012; Levine et al., 2020) aims to learn a policy from an offline dataset of trajectories for deployment in an online environment. It has been applied to various domains, such as robotic grasping (Kalashnikov et al., 2018; 2021), healthcare (Shortreed et al., 2010; Wang et al., 2018), and autonomous driving (Iroegbu & Madhavi, 2021; Diehl et al., 2023). The primary challenge in offline RL is the distribution shift (Fujimoto et al., 2018; Kumar et al., 2019), where the behavioral policy's data diverges from the data distribution in the actual online environment. Methods to address distribution shift broadly fall into two categories: (1) *policy-constraints* methods that restrict the learned policy to align closely with the behavior policy generating the dataset (Fujimoto et al., 2018; Kumar et al., 2019; Liu et al., 2019; Nachum et al., 2019; Peng et al., 2019; Siegel et al., 2020; Fujimoto & Gu, 2021), thus avoiding unexpected actions; (2) *conservative* methods (Kumar et al., 2020; Kidambi et al., 2020; Kostrikov et al., 2021; Yu et al., 2020) that construct a conservative return estimate, enhancing the learned policy's robustness against distribution shift. Previous studies have shown offline RL's efficacy even with random or sub-optimal datasets (Jin et al., 2020; Zanette, 2020; Wang et al., 2020; Rashidinejad et al., 2021). However, there is limited research on the performance of offline RL in scenarios where the observation data is noisy, particularly when learning from high-dimensional image inputs with complex distractors. Our work focuses on learning a policy from such low-quality offline visual RL datasets to effectively tackle related tasks.

Control with Noisy Observations. Real-world control tasks often involve observations that contain irrelevant noise. Recent studies that focused on learning policies from noisy image observations, can be categorized into four types: (1) separating task-relevant and irrelevant information based on key factors (actions, rewards, etc.) (Fu et al., 2021; Wang et al., 2022; Pan et al., 2022; Liu et al., 2023b); (2) learning task-related representations through bisimulation metrics (Zhang et al., 2021; Liu et al., 2023a); (3) mitigating exogenous noises via data augmentation methods (Kostrikov et al., 2020; Hansen et al., 2021; Fan & Li, 2021; Yuan et al., 2022; Bertoin et al., 2022; Huang et al., 2022); (4) extracting task-related features using auxiliary prediction tasks (Yang et al., 2015; Badia et al., 2020; Baker et al., 2022; Efroni et al., 2022; Lamb et al., 2022). These methods primarily target training policies in online environments, where additional data can be gathered to differentiate between task-relevant and irrelevant information. However, this issue has been less explored in offline visual RL. The V-D4RL benchmark dataset (Lu et al., 2022) involves only two related subtasks and does not offer a specific approach for handling noisy observations. Our method distinctively trains a separated state-space model from offline noisy data using a conservative sampling method and learns a high-performance policy.

Sampling Strategy in Offline RL. Sampling strategies in offline RL aim to improve the performance of the learning agent by optimally selecting data from the dataset. One approach involves using uncertainty estimation of the Q-Value function to guide sampling (Kumar & Kuzovkin, 2022), which allows for a better exploration-exploitation balance. Another strategy is Offline Prioritized Experience Replay (OPER) (Hong et al., 2023), which assigns weights to transi-

tions based on their normalized advantage, prioritizing those with higher rewards. Rank-Based Sampling (RBS) (Shen et al., 2021) adopts a similar way, sampling transitions with high returns more frequently. Policy Regularization with Dataset Constraint (PRDC) (Ran et al., 2023) limits the policy towards the closest state-action pair in the dataset, avoiding out-of-distribution actions. To address the imbalance dataset issue, RB-CQL (Jiang et al., 2023) utilizes a retrieval process to use related experiences effectively. Unlike these methods, which directly target policy performance improvement, our proposed conservative sampling strategy focuses on helping the model distinguish between task-relevant and irrelevant information.

6. Conclusion

In this paper, we consider the problem of learning a policy from offline visual datasets where the observations contain non-trivial distractors and the behavior policies that generate the dataset are either random or sub-optimal. To simulate this scenario, we construct the LQV-D4RL benchmark. We provide a theoretical analysis of the lower performance bound under the EX-BMDP assumption which is tighter than that under the POMDP in some specific cases. We propose a conservative sampling method to facilitate the decomposition of endogenous and exogenous states. Guided by this theoretical framework, we introduce the Separated Model-based Offline Policy Optimization (SeMOPO) method. Experimental results on the LQV-D4RL dataset indicate that SeMOPO outperforms other offline visual RL methods. Further experiments validate the theory's applicability and highlight the significance of each component within our method. Generalization experiments show SeMOPO's ability to handle variations between online and offline environmental disturbances.

Limitations and Future Work. Our study adopts the independence assumption under the EX-BMDP for the transitions of endogenous and exogenous states. However, potential interactions between these states in real-world tasks may present a promising avenue for future research.

Impact Statement

This paper presents work that aims to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

We thank Han-Jia Ye, Shaowei Zhang, Minghao Shao, Hai-Hang Sun, Kaichen Huang, Yucen Wang, and Jiayi Wu for their valuable discussions. This work was partially supported by the National Science and Technology Major Project under Grant No. 2022ZD0114805, Collaborative Innovation Center of Novel Software Technology and Industrialization, NSFC (62376118, 62006112, 62250069, 61921006), the Postgraduate Research & Practice Innovation Program of Jiangsu Province, and the National Natural Science Foundation of China (No. 123B2009).

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. G. Deep reinforcement learning at the edge of the statistical precipice. In *Neural Information Processing Systems*, 2021.

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, D., and Blundell, C. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, 2020.

Bain, M. and Sammut, C. A framework for behavioural cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, UK, July 1995]*, pp. 103–129. Oxford University Press, 1995.

Baker, B., Akkaya, I., Zhokhov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *ArXiv*, abs/2206.11795, 2022.

Bertoin, D., Zouitine, A., Zouitine, M., and Rachelson, E. Look where you look! saliency-guided q-networks for visual rl tasks. *ArXiv*, abs/2209.09203, 2022.

Bratko, I., Urbancic, T., and Sammut, C. Behavioural cloning: Phenomena, results and problems. *IFAC Proceedings Volumes*, 28:143–149, 1995.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI gym. *ArXiv*, abs/1606.01540, 2016.

Diehl, C. E., Sievernich, T., Krüger, M., Hoffmann, F., and Bertram, T. Uncertainty-aware model-based offline reinforcement learning for automated driving. *IEEE robotics and automation letters*, 2023. doi: 10.1109/LRA.2023.3236579.

Du, S. S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudík, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. *ArXiv*, abs/1901.09018, 2019.

Efroni, Y., Misra, D., Krishnamurthy, A., Agarwal, A., and Langford, J. Provable RL with exogenous distractors via multistep inverse dynamics. *ArXiv*, abs/2110.08847, 2021.

Efroni, Y., Foster, D. J., Misra, D. K., Krishnamurthy, A., and Langford, J. Sample-efficient reinforcement learning in the presence of exogenous information. In *Annual Conference Computational Learning Theory*, 2022.

Fan, J. and Li, W. Robust deep reinforcement learning via multi-view information bottleneck. *ArXiv*, abs/2102.13268, 2021.

Fu, X., Yang, G., Agrawal, P., and Jaakkola, T. Learning task informed abstractions. In *International Conference on Machine Learning*, 2021.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *ArXiv*, abs/2106.06860, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2018.

Ganin, Y. and Lempitsky, V. S. Unsupervised domain adaptation by backpropagation. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1180–1189. JMLR.org, 2015.

Ha, D. and Schmidhuber, J. World models. *ArXiv*, abs/1803.10122, 2018.

Hafner, D., Lillicrap, T. P., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering Atari with discrete world models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hansen, N., Su, H., and Wang, X. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems*, 34:3680–3693, 2021.

Hong, Z.-W., Kumar, A., Karnik, S., Bhandwaldar, A., Srivastava, A., Pajarinen, J., Laroche, R., Gupta, A., and Agrawal, P. Beyond uniform sampling: Offline reinforcement learning with imbalanced datasets. *ArXiv*, abs/2310.04413, 2023.

Huang, Y., Peng, P., Zhao, Y., Chen, G., and Tian, Y. Spectrum random masking for generalization in image-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:20393–20406, 2022.

Iroegbu, E. I. and Madhavi, D. Accelerating the training of deep reinforcement learning in autonomous driving. *IAES International Journal of Artificial Intelligence*, 2021. doi: 10.11591/IJAI.V10.I3.PP649-656.

Jiang, L., Cheng, S., Qiu, J., Xu, H., Chan, W. K. V., and Ding, Z. Offline reinforcement learning with imbalanced datasets. *ArXiv*, abs/2307.02752, 2023.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 2020.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., and Levine, S. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *ArXiv*, abs/1806.10293, 2018.

Kalashnikov, D., Varley, J., Chebotar, Y., Swanson, B., Jonschkowski, R., Finn, C., Levine, S., and Hausman, K. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *ArXiv*, abs/2104.08212, 2021.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Kostrikov, I., Yarats, D., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *ArXiv*, abs/2004.13649, 2020.

Kostrikov, I., Tompson, J., Fergus, R., and Nachum, O. Offline reinforcement learning with fisher divergence critic regularization. *ArXiv*, abs/2103.08050, 2021.

Kumar, A. and Kuzovkin, I. Offline robot reinforcement learning with uncertainty-guided human expert sampling. *ArXiv*, abs/2212.08232, 2022.

Kumar, A., Fu, J., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Neural Information Processing Systems*, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *ArXiv*, abs/2006.04779, 2020.

Kumar, A., Hong, J., Singh, A., and Levine, S. When should we prefer offline reinforcement learning over behavioral cloning? *ArXiv*, abs/2204.05618, 2022.

Lamb, A., Islam, R., Efroni, Y., Didolkar, A., Misra, D. K., Foster, D. J., Molu, L., Chari, R., Krishnamurthy, A., and Langford, J. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Trans. Mach. Learn. Res.*, 2023, 2022.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020.

Liu, Q., Zhou, Q., Yang, R., and Wang, J. Robust representation learning by clustering with bisimulation metrics for visual reinforcement learning with distractions. In *AAAI Conference on Artificial Intelligence*, 2023a.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with stationary distribution correction. *ArXiv*, abs/1904.08473, 2019.

Liu, Y.-R., Huang, B., Zhu, Z., Tian, H., Gong, M., Yu, Y., and Zhang, K. Learning world models with identifiable factorization. *ArXiv*, abs/2306.06561, 2023b.

Lu, C., Ball, P. J., Rudner, T. G. J., Parker-Holder, J., Osborne, M. A., and Teh, Y. W. Challenges and opportunities in offline reinforcement learning from visual observations. *ArXiv*, abs/2206.04779, 2022.

Luo, Y., Xu, H., Li, Y., Tian, Y., Darrell, T., and Ma, T. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *ArXiv*, abs/1912.02074, 2019.

Pan, M., Zhu, X., Wang, Y., and Yang, X. Isolating and leveraging controllable and noncontrollable visual dynamics in world models. *ArXiv*, abs/2205.13817, 2022.

Pathak, D., Gandhi, D., and Gupta, A. K. Self-supervised exploration via disagreement. *ArXiv*, abs/1906.04161, 2019.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *ArXiv*, abs/1910.00177, 2019.

Prudencio, R. F., Máximo, M. R. O. A., and Colombini, E. L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *ArXiv*, abs/2203.01387, 2022.

Rafailov, R., Yu, T., Rajeswaran, A., and Finn, C. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control*, pp. 1154–1168. PMLR, 2021.

Ran, Y., Li, Y., Zhang, F., Zhang, Z., and Yu, Y. Policy regularization with dataset constraint for offline reinforcement learning. In *International Conference on Machine Learning*, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. J. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68:8156–8196, 2021.

Shen, Y., Fang, Z., Xu, Y., Cao, Y., and Zhu, J. A rank-based sampling framework for offline reinforcement learning. *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 197–202, 2021.

Shortreed, S. M., Laber, E. B., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84:109–136, 2010.

Siegel, N., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., and Riedmiller, M. A. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *ArXiv*, abs/2002.08396, 2020.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T. P., and Riedmiller, M. A. DeepMind control suite. *ArXiv*, abs/1801.00690, 2018.

Tomar, M., Islam, R., Taylor, M., Levine, S., and Bachman, P. Ignorance is bliss: Robust control via information gating. *Advances in Neural Information Processing Systems*, 36, 2024.

Trabucco, B., Geng, X., Kumar, A., and Levine, S. Design-bench: Benchmarks for data-driven offline model-based optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21658–21676. PMLR, 2022.

Wang, L., Zhang, W., He, X., and Zha, H. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *ArXiv*, abs/2010.11895, 2020.

Wang, T., Du, S. S., Torralba, A., Isola, P., Zhang, A., and Tian, Y. Denoised mdps: Learning world models better than the world itself. *ArXiv*, abs/2206.15477, 2022.

Yang, Y., Ye, H.-J., Zhan, D.-C., and Jiang, Y. Auxiliary information regularized machine for multiple modality feature learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Yarats, D., Fergus, R., Lazaric, A., and Pinto, L. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Yuan, Z., Ma, G., Mu, Y., Xia, B., Yuan, B., Wang, X., Luo, P., and Xu, H. Don't touch what matters: Task-aware lipschitz data augmentation for visual reinforcement learning. *ArXiv*, abs/2202.09982, 2022.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *ArXiv*, abs/2012.08005, 2020.

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

A. Proof

A.1. Proof of Theorem 3.3

Lemma A.1. *(Lemma 3.1) (Telescoping lemma in the endogenous state space). Let M and $\widetilde{M}$ be two MDPs with the same reward function r, but different dynamics T and $\widetilde{T}$ respectively. Let $G^{\pi}_{\widetilde{M}}(s^+, a) := \mathop{\mathbb{E}}_{s^{+\prime}\sim\widetilde{T}(s^+,a)}\left[V^{\pi}_{M}\left(s^{+\prime}\right)\right] - \mathop{\mathbb{E}}_{s^{+\prime}\sim T(s^+,a)}\left[V^{\pi}_{M}\left(s^{+\prime}\right)\right]$. Then,*

$$\eta_{\widetilde{M}}(\pi) - \eta_M(\pi) = \gamma \mathop{\overline{\mathbb{E}}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[G^{\pi}_{\widetilde{M}}(s^+, a)\right]$$

Proof. We adopt the same proof as (Luo et al., 2019; Yu et al., 2020) to prove the telescoping lemma in the endogenous state space. Let W_j be the expected return when executing π on $\widetilde{T}$ for the first j steps, then switching to T for the remainder. That is,

$$W_j = \mathop{\mathbb{E}}_{\substack{t<j:s^+_{t+1}\sim\widetilde{T}\left(s^+_t,a_t\right)\\ t\geq j:s^+_{t+1}\sim T\left(s^+_t,a_t\right)}}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s^+_t, a_t\right)\right]$$

Note that $W_0 = \eta_M(\pi)$ and $W_\infty = \eta_{\widetilde{M}}(\pi)$, so

$$\eta_{\widetilde{M}}(\pi) - \eta_M(\pi) = \sum_{j=0}^{\infty}\left(W_{j+1} - W_j\right)$$

Write

$$W_j = R_j + \mathop{\mathbb{E}}_{s^+_j, a_j\sim\pi,\widetilde{T}}\left[\mathop{\mathbb{E}}_{s^+_{j+1}\sim T\left(s^+_t,a_t\right)}\left[\gamma^{j+1}V^{\pi}_M\left(s^+_{j+1}\right)\right]\right]$$
$$W_{j+1} = R_j + \mathop{\mathbb{E}}_{s^+_j, a_j\sim\pi,\widetilde{T}}\left[\mathop{\mathbb{E}}_{s^+_{j+1}\sim \widetilde{T}\left(s^+_t,a_t\right)}\left[\gamma^{j+1}V^{\pi}_M\left(s^+_{j+1}\right)\right]\right]$$

where R_j is the expected return of the first j time steps, which are taken with respect to $\widetilde{T}$. Then

$$\begin{aligned}W_{j+1} - W_j &= \gamma^{j+1}\mathop{\mathbb{E}}_{s^+_j, a_j\sim\pi,\widetilde{T}}\left[\mathop{\mathbb{E}}_{s^{+\prime}\sim\widetilde{T}\left(s^+_j,a_j\right)}\left[V^{\pi}_M\left(s^{+\prime}\right)\right] - \mathop{\mathbb{E}}_{s^{+\prime}\sim T\left(s^+_j,a_j\right)}\left[V^{\pi}_M\left(s^{+\prime}\right)\right]\right]\\ &= \gamma^{j+1}\mathop{\mathbb{E}}_{s^+_j, a_j\sim\pi,\widetilde{T}}\left[G^{\pi}_{\widetilde{M}}\left(s^+_j, a_j\right)\right]\end{aligned}$$

Thus

$$\begin{aligned}\eta_{\widetilde{M}}(\pi) - \eta_M(\pi) &= \sum_{j=0}^{\infty}\left(W_{j+1} - W_j\right)\\ &= \sum_{j=0}^{\infty}\gamma^{j+1}\mathop{\mathbb{E}}_{s^+_j, a_j\sim\pi,\widetilde{T}}\left[G^{\pi}_{\widetilde{M}}\left(s^+_j, a_j\right)\right]\\ &= \gamma^{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[G^{\pi}_{\widetilde{M}}(s^+, a)\right]\end{aligned}$$

□

Theorem A.2. *(Theorem 3.3) Under Assumption 3.2, we can define the uncertainty estimation $\epsilon_{\tilde{u}}(\pi)$ under the EX-BMDP as $\epsilon_{\tilde{u}}(\pi) := \mathop{\overline{\mathbb{E}}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}[\tilde{u}(s^+, a)]$. Let $\tilde{\pi}$ denote the learned optimal policy under reward-penalized the endogenous MDP, then,*

$$\eta_M(\tilde{\pi}) \geq \sup_{\pi}\{\eta_M(\pi) - 2\lambda\epsilon_{\tilde{u}}(\pi)\}$$

Proof. With Lemma 3.1, we can get the corollary that:

$$\eta_M(\pi) = \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[r(s^+,a) - \gamma G^{\pi}_{\widetilde{M}}(s^+,a)\right] \geq \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[r(s^+,a) - \gamma|G^{\pi}_{\widetilde{M}}(s^+,a)|\right]$$

With Assumption 3.2, we know that $|G^{\pi}_{\widetilde{M}}(s^+,a)| \leq \tilde{u}(s^+,a)$. Thus, we define the penalized reward function as $\tilde{r}(s^+,a) = r(s^+,a) - \lambda u(s^+,a)$, where $\lambda = \gamma c$ and c is the scalar that satisfies $|G_{\widetilde{M}^{\pi}}(s^+,a)| \leq c d_{\mathcal{F}}(\widetilde{T}(s^+,a), T(s^+,a))$. We can obtain the relationship between the policy performance under the true MDP and the reward-penalized endogenous MDP $\widetilde{M}_{\tilde{r}} = (\mathcal{S}, \mathcal{A}, \widetilde{T}, \tilde{r}, \mu_0, \gamma)$:

$$\begin{aligned}
\eta_M(\pi) &\geq \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[r(s^+,a) - \gamma|G^{\pi}_{\widetilde{M}}(s^+,a)|\right] \\
&\geq \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[r(s^+,a) - \lambda\tilde{u}(s^+,a)\right] \\
&\geq \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}\left[\tilde{r}(s^+,a)\right] \\
&= \eta_{\widetilde{M}_{\tilde{r}}}(\pi)
\end{aligned} \tag{3}$$

From Assumption 3.2 We can easily get the two-sided bound that:

$$|\eta_M(\pi) - \eta_{\widetilde{M}}(\pi)| \leq \bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}|\gamma G^{\pi}_{\widetilde{M}}(s^+,a)| \leq \lambda\bar{\mathbb{E}}_{(s^+,a)\sim\rho^{\pi}_{\widetilde{T}}}[\tilde{u}(s^+,a)] = \lambda\epsilon_{\tilde{u}}(\pi) \tag{4}$$

With the help of Equation (3) and Equation (4), we can get the performance lower bound of the learned optimal policy $\tilde{\pi}$ in the reward-penalized endogenous MDP:

$$\begin{aligned}
& \eta_M(\tilde{\pi}) \geq \eta_{\widetilde{M}_{\tilde{r}}}(\tilde{\pi}) \geq \eta_{\widetilde{M}_{\tilde{r}}}(\pi) = \eta_{\widetilde{M}}(\pi) - \lambda\epsilon_{\tilde{u}}(\pi) \geq \eta_M(\pi) - 2\lambda\epsilon_{\tilde{u}}(\pi) \\
\Rightarrow \quad & \eta_M(\tilde{\pi}) \geq \sup_{\pi}\{\eta_M(\pi) - 2\lambda\epsilon_{\tilde{u}}(\pi)\}
\end{aligned}$$

□

A.2. Proof of Theorem 3.4

Assumption A.3. There exists an underlying function $f : \mathcal{S}^+ \times \mathcal{S}^- \to \mathcal{Z}$ satisfying that for any $s_1^+, s_2^+ \in \mathcal{S}^+, s_1^-, s_2^- \in \mathcal{S}^-$, if $f(s_1^+, s_1^-) = f(s_2^+, s_2^-)$, then

$$p(o_t|s_1^+, s_1^-) = p(o_t|s_2^+, s_2^-) \tag{5}$$

This assumption is intuitive because the block structure indicates that each context o uniquely determines its generating state z. However, z can have multiple different partitions of s^+ and s^- based on specific semantics, such as task-related information, viewpoint, etc.

Assumption A.4. For each i, the distribution of action for trajectory τ_i is the same, i.e., there exist $p(a)$ such that

$$\mathbb{E}_{s^+}\pi_i(a|s^+) = p(a), \quad i = 1, 2, \cdots, n.$$

Theorem A.5. *(Theorem 3.4) Consider the likelihood optimization problem on the same offline dataset $\mathcal{B}$ but with two different sampling methods. let $\mathcal{B}_{\pi_i}$ be the dataset collected by the behavior policy π_i, where $i = 1, 2, \cdots, n$. $\mathcal{B}_{\pi_{mix}} = \mathcal{B}_{\pi_1} \cup \mathcal{B}_{\pi_1} \cup \cdots \cup \mathcal{B}_{\pi_n}$ is the mixture of the datasets of all policies. Then we have*

$$\mathbb{E}_{\tau\in\mathcal{B}_{\pi_{mix}}} \ln p(\tau) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\tau\in\mathcal{B}_{\pi_i}} \ln p(\tau),$$

where $p(\tau)$ is the true density of τ.

Proof. For that τ generated under the guide of π, the action distribution of the policy, we maximize the log-likelihood to evaluate the conditional distributions $p(o|s^+, s^-), p(a|s^+), p(s^{+'}|s^+, a)$ and $p(s^{-'}|s^-)$ which also denoted as $p_o, p_a, p_{s^+}, p_{s^-}$

respectively for convenience, i.e.,

$$\begin{aligned}\hat{p}_o, \hat{p}_a, \hat{p}_{s^+}, \hat{p}_{s^-} &= \underset{p_o, p_a, p_{s^+}, p_{s^-} \in \mathcal{H}_\Theta}{\arg\max} \mathbb{E}_{\tau \in \mathcal{B}_\pi} \ln p(\tau) \\ &= \underset{p_o, p_a, p_{s^+}, p_{s^-} \in \mathcal{H}_\Theta}{\arg\max} \mathbb{E}_{\tau \in \mathcal{B}_\pi} \sum_{t=1}^{T} \Big(\ln p_o(o_t | s_t^+, s_t^-) + \ln p_a(a_t | s_t^+) \\ &\qquad + \ln p_{s^+}(s_t^+ | s_{t-1}^+, a_{t-1}) + \ln p_{s^-}(s_t^- | s_{t-1}^-) \Big).\end{aligned}$$

Here $\mathcal{H}_\Theta$ is the approximate space and Θ is the class of parameters. In practice, $\mathcal{H}_\Theta$ is the variational posterior space with an encoder and decoder for each conditional distribution.

With Assumption A.3, we know that once the distribution of $f(s_t^+, s_t^-)$ is fixed, the variant of s_t^+ and s_t^- has no effect on the observation o_t. Therefore, if we want to separate sub-optimal distributions for s_t^+ and s_t^- from the optimal results, the key terms in the likelihood are those involving the action a_t. Formally, we treat the likelihood sequentially as

$$\mathbb{E}_{\tau \in \mathcal{B}_\pi} \ln p(\tau) := \sum_{t=1}^{T} l_\pi^{(t)}(s_t^+, s_t^-, a) + \mathbb{E}_{\tau \in \mathcal{B}_\pi} \ln p_o(o_t | s_t^+, s_t^-),$$

where

$$l_\pi^{(t)}(s_t^+, s_t^-, a) = \mathbb{E}_{a \sim \pi} \ln p_a(a_t | s_t^+) + \ln p_{s_t^+}(s_t^+ | s_{t-1}^+, a_{t-1}) + \ln p_{s_t^-}(s_t^- | s_{t-1}^-).$$

Define the equivalent class for s^+, s^- as

$$\mathcal{A}_p = \left\{ s^+ \sim p_{s^+}, s^- \sim p_{s^-} : f(s^+, s^-) \sim p \right\}.$$

Different elements in $\mathcal{A}_p$ means different distributions for s^+ and s^- that induce the same distribution for $f(s^+, s^-)$. From Assumption A.3, we know that for those s^+, s^- in a certain class $\mathcal{A}_{p_f}$, $\mathbb{E}_{\tau \in \mathcal{B}_\pi} \ln p_o(o_t | s_t^+, s_t^-)$ is only related to the distribution p_f and we denote its value as $l_o(p_f)$. For $t = 1$, we know

$$l_\pi^{(1)}(s^+, s^-, a) = \mathbb{E}_\pi \ln p_a(a | s^+) + \ln p_{s^+}(s^+) + \ln p_{s^-}(s^-).$$

Recall the conditional entropy and mutual information entropy for a and s^+ such that

$$H(a) = \int -\ln p_a(a) dp_a(a), \quad H(a|s^+) = \int dp_{s^+}(s^+) \int -\ln p_a(a|s^+) dp_a(a|s^+), \quad I(a, s^+) = H(a) - H(a|s^+).$$

Under Assumption A.4, consider two different type of policy π_i as the policy of i-th curve and π_{mix} as the mixture policy of all curves, i.e., $\pi_{mix} = \frac{1}{n} \sum_{i=1}^{n} \pi_i$. We know that

$$\begin{aligned}I(a \sim \pi_{mix}, s^+ \sim p_{s^+}) &= H(a) + \int dp_{s^+}(s^+) \int \frac{1}{n} \sum_{i=1}^{n} \pi_i(a|s^+) \ln \frac{1}{n} \sum_{i=1}^{n} \pi_i(a|s^+) da \\ &\leq H(a) + \int dp_{s^+}(s^+) \int \frac{1}{n} \sum_{i=1}^{n} \pi_i(a|s^+) \ln \pi_i(a|s^+) da \\ &= \frac{1}{n} \sum_{i=1}^{n} I(a \sim \pi_i, s^+ \sim p_{s^+}).\end{aligned}$$

Define $p_{s^+}^\star(s_{t=1}^+), p_{s^-}^\star(s_{t=1}^-) \in \mathcal{A}_{P_{t=1}^\star}$ as the true distribution of s^+ and s^- at time $t = 1$. Once the class $\mathcal{A}_{p_f}$ and the

distribution of s^+ is fixed, the distribution for s^- is also fixed. So we know that

$$\begin{aligned}
&l^{(1)}_{\pi_{mix}}(s^+ \sim p^\star_{s^+}(s^+_{t=1}), s^- \sim p^\star_{s^-}(s^-_{t=1}), a) - \frac{1}{n}\sum_{i=1}^{n} l^{(1)}_{\pi_i}(s^+ \sim p^\star_{s^+}(s^+_{t=1}), s^- \sim p^\star_{s^-}(s^-_{t=1}), a) \\
&= I(a \sim \pi_{mix}, s^+ \sim p^\star_{s^+}) + \mathbb{E}_{\pi_{mix}} \ln p^\star_{s^+}(s^+) + \ln p^\star_{s^-}(s^-) \\
&\quad - I(a \sim \pi_{mix}, s^+ \sim p^\star_{s^+}) - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\pi_i} \ln p^\star_{s^+}(s^+) + \ln p^\star_{s^-}(s^-)\right) \\
&= I(a \sim \pi_{mix}, s^+ \sim p^\star_{s^+}) - \frac{1}{n}\sum_{i=1}^{n} I(a \sim \pi_i, s^+ \sim p^\star_{s^+}) \le 0.
\end{aligned}$$

For $t > 1$, similarly define that $p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}), p^\star_{s^-}(s^-_t|s^-_{t-1}) \in \mathcal{A}_{P^\star_{t>1}}$ as the true condition distribution of s^+_t condition on s^+_{t-1} and s^-_t condition on s^-_{t-1}. So

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_{mix}} \ln p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}) &= \int \pi_{mix}(a_{t-1}|s^+_{t-1}) \ln p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}) da_{t-1} \\
&= \frac{1}{n}\mathbb{E}_{a\sim\pi_i} \ln p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}).
\end{aligned}$$

Then we also have

$$\begin{aligned}
&l^{(t)}_{\pi_{mix}}(s^+ \sim p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}), s^- \sim p^\star_{s^-}(s^-_t|s^-_{t-1}), a) - \frac{1}{n}\sum_{i=1}^{n} l^{(t)}_{\pi_i}(s^+ \sim p^\star_{s}(s^+_t|s^+_{t-1}, a_{t-1}), s^- \sim p^\star_{s^-}(s^-_t|s^-_{t-1}), a) \\
&= I(a \sim \pi_{mix}, s^+ \sim \prod_{k=1}^{t} \mathbb{E}_{s^+_{k-1}, a_{t-1}} p^\star_{s^+}(s^+_k|s^+_{k-1}, a_{t-1})) + \mathbb{E}_{\pi_{mix}} \ln p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}) + \ln p^\star_{s^-}(s^-_t|s^-_{t-1}) \\
&- \frac{1}{n}\sum_{i=1}^{n} I(a \sim \pi_i, s^+ \sim \prod_{k=1}^{t} \mathbb{E}_{s^+_{k-1}, a_{t-1}} p^\star_{s^+}(s^+_k|s^+_{k-1}, a_{t-1})) - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\pi_i} \ln p^\star_{s^+}(s^+_t|s^+_{t-1}, a_{t-1}) + \ln p^\star_{s^-}(s^-_t|s^-_{t-1})\right) \\
&= I(a \sim \pi_{mix}, s^+ \sim \prod_{k=1}^{t} \mathbb{E}_{s^+_{k-1}, a_{t-1}} p^\star_{s^+}(s^+_k|s^+_{k-1}, a_{t-1})) - \frac{1}{n}\sum_{i=1}^{n} I(a \sim \pi_i, s^+ \sim \prod_{k=1}^{t} \mathbb{E}_{s^+_{k-1}, a_{t-1}} p^\star_{s^+}(s^+_k|s^+_{k-1}, a_{t-1})) \le 0.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{\tau\in\mathcal{B}_{\pi_{mix}}} \ln p^\star(\tau) &= \sum_{t=1}^{T} l^{(t)}_{\pi_{mix}}(s^+ \sim p^\star_{s^+}(s^+_t|s^+_{t-1}), s^- \sim p^\star_{s^-}(s^-_t|s^-_{t-1}), a) + l_o(p^\star_t) \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} l^{(t)}_{\pi_i}(s^+ \sim p^\star_{s^+}(s^+_t|s^+_{t-1}), s^- \sim p^\star_{s^-}(s^-_t|s^-_{t-1}), a) + l_o(p^\star_t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tau\in\mathcal{B}_{\pi_i}} \ln p^\star(\tau).
\end{aligned}$$

Besides, if we can select a subset $\mathcal{C}$ of curves i with large mutual information entropy $I(a \sim \pi_i|s^+ \sim p^\star_{s^+})$, i.e., $\frac{1}{\#\mathcal{C}}\sum_{i\in\mathcal{C}} I(a \sim \pi_i|s^+ \sim p^\star_{s^+}) > I(a \sim \pi_{mix}|s^+ \sim p^\star_{s^+})$ and we know that

$$\mathbb{E}_{\tau\in\mathcal{B}_{\pi_{mix}}} \ln p^\star(\tau) < \frac{1}{\#\mathcal{C}}\sum_{i\in\mathcal{C}} \mathbb{E}_{\tau\in\mathcal{B}_{\pi_i}} \ln p^\star(\tau).$$

□

B. Derivation

Likelihood. For the trajectory $\tau_i = \{o_1, a_1, \cdots, o_T, a_T\}$, the log-likelihood can be expressed as:

$$\begin{aligned}
\ln p(\tau_i) &= \ln p(o_0, a_0, \cdots, o_T, a_T) \\
&= \ln \prod_{t=1}^{T} p(o_t|z_t) p(a_t|z_t) p(z_t|z_{t-1}, a_{t-1}) \\
&= \ln \prod_{t=1}^{T} p(o_t|s_t^+, s_t^-) p(a_t|s_t^+) p(s_t^+|s_{t-1}^+, a_{t-1}) p(s_t^-|s_{t-1}^-) \\
&= \sum_{t=1}^{T} \big[\ln p(o_t|s_t^+, s_t^-) + \ln p(a_t|s_t^+) + \ln p(s_t^+|s_{t-1}^+, a_{t-1}) + \ln p(s_t^-|s_{t-1}^-) \big],
\end{aligned}$$

The third equation comes from the decomposition of endogenous and exogenous dynamics, as assumed in the EX-BMDP framework.

One-step predictive distribution. The variational bound for latent dynamics models $p(o_{1:T}, z_{1:T}|a_{1:T}) = \prod_t p(s_t^+|s_{t-1}^+, a_{t-1}) p(s_t^-|s_{t-1}^-) p(o_t|z_t)$ and a variational posterior $q(z_{1:T}|o_{1:T}, a_{1:T}) = \prod_t q(s_t^+|o_{\leq t}, a_{<t}) q(s_t^-|o_{\leq t})$ follows from importance weighting and Jensen's inequality as shown:

$$\begin{aligned}
\ln p(o_{1:T}|a_{1:T}) &\triangleq \ln \mathbb{E}_{p(z_{1:T}|a_{1:T})} \Big[\prod_{t=1}^{T} p(o_t|z_t) \Big] \\
&= \ln \mathbb{E}_{q(z_{1:T}|o_{1:T}, a_{1:T})} \Big[\prod_{t=1}^{T} \frac{p(o_t|z_t) p(z_t|z_{t-1}, a_{t-1})}{q(z_t|o_{\leq t}, a_{<t})} \Big] \\
&= \ln \mathbb{E}_{q(s_{1:T}^+, s_{1:T}^-|o_{1:T}, a_{1:T})} \Big[\prod_{t=1}^{T} \frac{p(o_t|s_t^+, s_t^-) p(s_t^+|s_{t-1}^+, a_{t-1}) p(s_t^-|s_{t-1}^-)}{q(s_t^+|o_{\leq t}, a_{<t}) q(s_t^-|o_{\leq t})} \Big] \\
&\geq \mathbb{E}_{q(s_{1:T}^+, s_{1:T}^-|o_{1:T}, a_{1:T})} \Big[\sum_{t=1}^{T} \big(\ln p(o_t|s_t^+, s_t^-) + \ln p(s_t^+|s_{t-1}^+, a_{t-1}) + \ln p(s_t^-|s_{t-1}^-) \\
&\quad - \ln q(s_t^+|o_{\leq t}, a_{<t}) - \ln q(s_t^-|s_{t-1}^-) \big) \Big] \\
&= \sum_{t=1}^{T} \Big(\mathbb{E}_{q(s_t^+|o_{\leq t}, a_{<t}) q(s_t^-|o_{\leq t})} \Big[\ln p(o_t|s_t^+, s_t^-) \Big] \\
&\quad - \mathbb{E}_{q(s_{t-1}^+|o_{\leq t-1}, a_{<t-1})} \Big[\mathrm{KL}[q(s_t^+|o_{\leq t}, a_{<t}) || p(s_t^+|s_{t-1}^+, a_{t-1})] \Big] \\
&\quad - \mathbb{E}_{q(s_{t-1}^-|o_{\leq t-1})} \Big[\mathrm{KL}[q(s_t^-|o_{\leq t}) || p(s_t^-|s_{t-1}^-)] \Big] \Big).
\end{aligned}$$

C. The LQV-D4RL Benchmark

To evaluate the performance of visual RL methods with offline datasets containing noisy observations, we introduce a benchmark named Low-Quality Vision Datasets for Deep Data-Driven RL (LQV-D4RL). This benchmark comprises four typical environments from the DeepMind Control Suite and one environment from Gym:

- *Walker Walk*: A bipedal agent is trained to first stand and then walk forward as efficiently as possible.
- *Cheetah Run*: A cheetah-like bipedal model aims to run at high speeds on a straight track.
- *Hopper Hop*: The agent, with a single-legged body, must balance and hop forward, focusing on agility and stability.

- *Humanoid Walk*: A simplified humanoid with 21 joints, aims to walk stably, which is extremely difficult with many local minima.
- *Car Racing*: A highly challenging racing game where players must pass through checkpoints to score points. The faster they reach the finish line within a set time, the higher their score. The observations contain numerous distractors.

Each task is represented across three different levels of policy performance:

- **Random**: Trajectories are collected by randomly initialized policies.
- **Medium Replay** (medrep): Trajectories are drawn from the replay buffer accumulated during training of a medium-performance policy.
- **Medium**: Trajectories are collected by a fixed policy of medium performance.

For each locomotion task's observations, the backgrounds are replaced with videos from the "driving car" category of the Kinetics dataset (Kay et al., 2017), as utilized in DBC (Zhang et al., 2021). To simulate real data collection processes in natural settings, we train policies using the TIA approach (Fu et al., 2021) and then collect trajectories based on image observations with the aforementioned distractors. The "random" and "medium" datasets each contain 200 trajectories, while "medium_replay" comprises 400 trajectories, with each trajectory being 1000 steps long. Specific statistical details of the LQV-D4RL benchmark are reported in Table 3. We upload the dataset in the supplementary materials.

Table 3. Full summary statistics of per-episode return in the LQV-D4RL benchmark.

Dataset		Episodes	Mean	Std	Min	P25	Median	P75	Max
Walker Walk	random	200	86.6	48.1	5.9	51.9	70.8	128.3	199.1
	medrep	400	106.6	79.7	5.0	45.4	81.3	148.8	398.2
	medium	200	513.1	54.8	401.5	471.4	516.0	559.1	598.0
Cheetah Run	random	200	77.6	57.2	3.1	14.5	71.6	118.8	198.2
	medrep	400	145.4	113.3	1.7	42.9	125.1	232.6	396.5
	medium	200	350.3	30.8	300.5	322.2	352.5	376.0	403.2
Hopper Hop	random	200	2.1	4.9	0.0	0.0	0.0	0.1	19.5
	medrep	400	4.7	9.7	0.0	0.0	0.0	3.7	39.9
	medium	200	62.0	13.9	40.3	49.8	60.8	74.5	84.9
Humanoid Walk	random	200	1.1	0.8	0.0	0.5	1.0	1.5	5.7
	medrep	400	95.6	114.3	0.0	1.4	5.4	202.8	359.0
	medium	200	573.0	16.8	526.5	560.6	572.9	584.9	609.4
Car Racing	random	200	10.3	65.5	-82.0	-43.3	-6.5	59.8	149.2
	medrep	400	76.1	116.3	-82.0	-27.5	54.1	181.5	297.3
	medium	200	372.3	42.7	302.3	335.5	370.3	408.3	449.9

D. Implementation Details

D.1. Networks

We implement the proposed algorithm using TensorFlow 2 and conduct all experiments on an NVIDIA RTX 3090, totaling approximately 1000 GPU hours. The recurrent state-space model from DreamerV2 (Hafner et al., 2021) is employed for both forward dynamics and the posterior encoder. The hidden sizes of deterministic and stochastic parts of the model are 200 and 32, respectively. For learning an ensemble of forward dynamics, multiple MLP networks are utilized, each outputting the mean and standard deviation of the next state. The hidden size for each MLP is 1024. The reward predictor comprises 4 MLP layers, each of size 400. We use the convolutional encoder and decoder from TIA (Fu et al., 2021). All dense layers have a size of 400, and the activation function used is ELU. The ADAM optimizer is employed to train the network with batches of 64 sequences, each of length 50. The learning rate is 6e-5 for both the endogenous and exogenous models and 8e-5 for the action and value nets. We stabilize the training process by clipping gradient norms to 100 and set ($\lambda = 10$)

for the uncertainty penalty term. The imagine horizon of 5, as used in Offline DV2 (Lu et al., 2022), is adopted for policy optimization. We train a model comprising the dynamics and reward predictor for 25000 epochs using offline visual datasets, followed by policy training within the model for 100000 steps. The codes and datasets are contained in the supplementary materials.

D.2. Evaluation Metric

To compare the performance of different methods, we define the normalized return based on the statistics of the offline dataset as follows:

$$S_{\text{normalize}} = \frac{S_{\text{score}} - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}}$$

Here, S_{score} represents the original return obtained by the test method, while S_{min} and S_{max} denote the minimum and maximum episodic returns of the dataset for each task across three levels (random, medium_replay, medium), respectively.

D.3. Dissociated Reward Prediction

The Reward Dissociation used in TIA (Fu et al., 2021) for the exogenous model is achieved through the adversarial objective $\mathcal{J}^t_{Radv}$.

$$\mathcal{J}^t_{Radv} = -\lambda_{Radv} \max_q \ln q(r_t|s^-_t)$$

where λ_{Radv} for *Walker Walk*, *Cheetah Run*, and *Hopper Hop* are 20000, 20000, and 30000, respectively. This setup involves a minimax strategy, wherein the training of the exogenous model's reward prediction head is interleaved with the exogenous model's training, occurring for multiple iterations per training step. The reward prediction head is trained to minimize the reward prediction loss, represented by $-\ln q(r_t|s^-_t)$. In contrast, the exogenous model aims to maximize this objective to prevent reward-correlated information from influencing its learned features, as outlined by Ganin & Lempitsky. At the same time, TIA optimizes the endogenous model to maximize the log-likelihood of predicting rewards from endogenous states via the objective $\mathcal{J}^t_R = \ln q(r_t|s^+_t)$. The reward prediction loss is calculated using $\ln \mathcal{N}(r_t; \hat{r}_t, 1)$, where $\mathcal{N}(\cdot; \mu, \sigma^2)$ denotes the Gaussian likelihood and $\hat{r}_t$ represents the predicted reward. Notably, neither loss is used to update the endogenous and exogenous models in SeMOPO. The reward prediction loss only updates the reward predictors by stopping the backward gradients to the endogenous states.

E. Algorithm of SeMOPO

The pseudo-code of our proposed SeMOPO is provided in Algorithm 1.

Algorithm 1 Training Procedure of SeMOPO

Input: Offline datasets $\mathcal{B}$
Initialize forward dynamics model $\widetilde{T}_\theta, \bar{T}_\theta$, posterior encoder $\tilde{q}_\theta, \bar{q}_\theta$, observation decoder $\widetilde{U}_\theta$, reward predictor R_θ, policy π_θ, value model V_θ.
// Offline model training
for each training epoch $m = 1 \cdots M$ **do**
  Sample minibatch $(o_{1:T}, a_{1:T-1}, r_{1:T})_{1:b}$ from the dataset $\mathcal{B}$ via Conservative Sampling
  Update the forward dynamics model $\widetilde{T}_\theta, \bar{T}_\theta$ and the posterior encoder $\tilde{q}_\theta, \bar{q}_\theta$
  Update the observation decoder $\mathcal{U}_\theta$ and the reward predictor R_θ
end for
// Policy Optimization
for training iteration $i = 1 \cdots$ It **do**
  Imagine the endogenous latent states $s^+_{1:H}$ by policy π using the endogenous state model $\widetilde{T}_\theta$
  Estimate the endogenous model uncertainty by the model disagreement
  Obtain the penalized reward $\tilde{r}$ with estimated uncertainty $\tilde{u}$ via Equation (2)
  Train the policy π_θ and value model V_θ on the data $(s^+_{1:H-1}, a_{1:H-1}, \tilde{r}_{1:H-1})$
end for

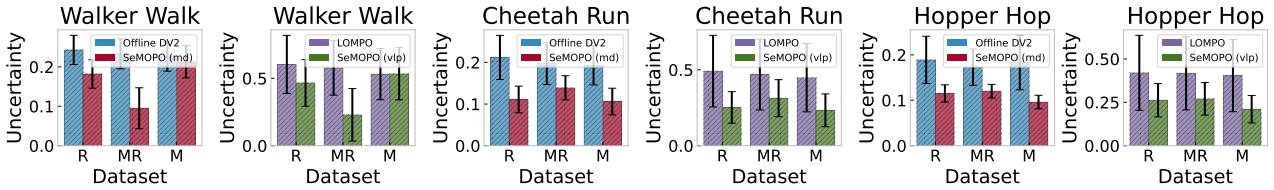

Figure 7. The model uncertainty estimation of SeMOPO, Offline DV2, and LOMPO on the LQV-D4RL dataset. We randomly select 1000 states and report the mean and standard deviation of uncertainty on them. "md" and "vlp" are the short names for the mean-disagreement and variance of logarithm prediction, denoting the calculation of uncertainty used in Offline DV2 and LOMPO, respectively.

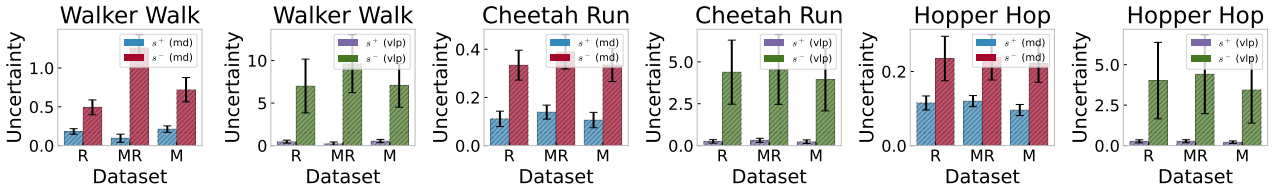

Figure 8. The model uncertainty estimation of SeMOPO on the LQV-D4RL dataset. We randomly select 1000 states and infer the endogenous states s^+ and exogenous states s^- by SeMOPO. We report the mean and standard deviation of uncertainty on these inferred states. "md" and "vlp" follow the same definition as in Figure 7.

F. Additional Results

F.1. The Model Uncertainty Estimation

To validate that the accuracy of model uncertainty estimation is due to the separation of endogenous and exogenous states, rather than a specific computational approach, we employ two different uncertainty estimation methods: the mean-disagreement of the ensemble (md) from Offline DV2, and the variance of logarithm prediction (vlp) from LOMPO. The task-related model uncertainty across various environments and datasets is shown in Figure 7. The model uncertainty of SeMOPO is superior under both computational methods compared to other techniques, suggesting that learning the model in the endogenous state space can reduce the estimated uncertainty.

Figure 8 presents the estimates of model uncertainty in both endogenous and exogenous state spaces. Notably, the model uncertainty in the exogenous state space is significantly higher than in the endogenous state space. This further implies that the overestimation of model uncertainty in task-related components in previous work is attributable to the unfiltered noise in the latent states.

F.2. Evaluation on the V-D4RL Benchmark

To assess the performance of our method on datasets without distractors, we compared SeMOPO with other approaches on the V-D4RL dataset. The results in Table 4 show that SeMOPO outperforms other methods on the random datasets of two environments, and achieves comparable performance on several other datasets. This indicates that our method is more suitable for datasets collected using non-expert policies, and can also address certain offline visual reinforcement learning problems without distractors.

F.3. Training Stability

The gradient norms during model training for Offline DV2 and SeMOPO are shown in Figure 9.

Table 4. The performance of different methods on the V-D4RL benchmark. We report the mean and standard deviation of test returns of SeMOPO over 4 seeds. The results for methods other than SeMOPO are sourced from Table 1 in (Lu et al., 2022).

Dataset		SeMOPO	Offline DV2	LOMPO	DrQ+BC	DrQ+CQL	BC
Walker Walk	random	**305** $\pm$ **8**	287 $\pm$ 130	219 $\pm$ 81	55 $\pm$ 9	144 $\pm$ 124	20 $\pm$ 2
	medrep	218 $\pm$ 21	**565** $\pm$ **181**	347 $\pm$ 197	287 $\pm$ 69	114 $\pm$ 124	165 $\pm$ 43
	medium	406 $\pm$ 30	434 $\pm$ 111	341 $\pm$ 197	**468** $\pm$ **23**	148 $\pm$ 161	409 $\pm$ 31
Cheetah Run	random	**330** $\pm$ **2**	329 $\pm$ 2	114 $\pm$ 51	58 $\pm$ 6	59 $\pm$ 84	0 $\pm$ 0
	medrep	410 $\pm$ 13	**616** $\pm$ **10**	363 $\pm$ 136	448 $\pm$ 36	107 $\pm$ 128	250 $\pm$ 36
	medium	190 $\pm$ 22	172 $\pm$ 35	164 $\pm$ 83	**530** $\pm$ **30**	409 $\pm$ 51	516 $\pm$ 14

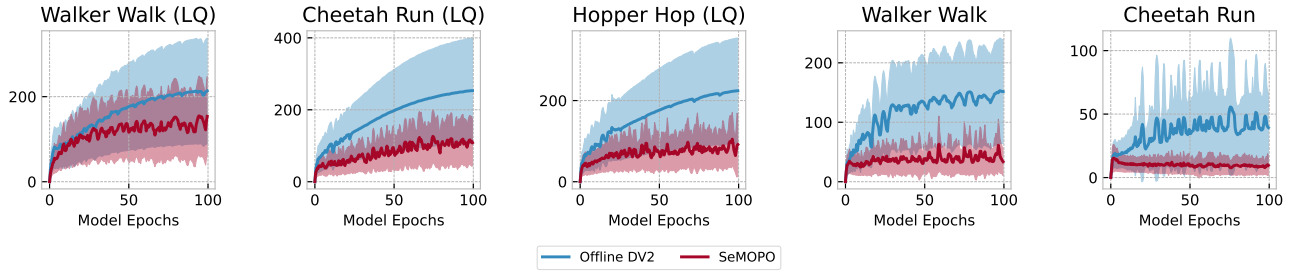

Figure 9. Comparison of gradient norms during model training for Offline DV2 and SeMOPO across five tasks: the first three from LQV-D4RL and the latter two from V-D4RL. Each curve represents aggregated data from three levels of datasets (random, medium_replay, medium), illustrating the mean (solid line) and standard deviation (shaded region) over four seeds. SeMOPO exhibits lower model gradient norms than Offline DV2, regardless of distractors in observations, indicating that the separation of task-relevant information from observations contributes to more stable model training.

F.4. The results on the LQV-D4RL benchmark

We record the original unnormalized returns for each method in Table 5.

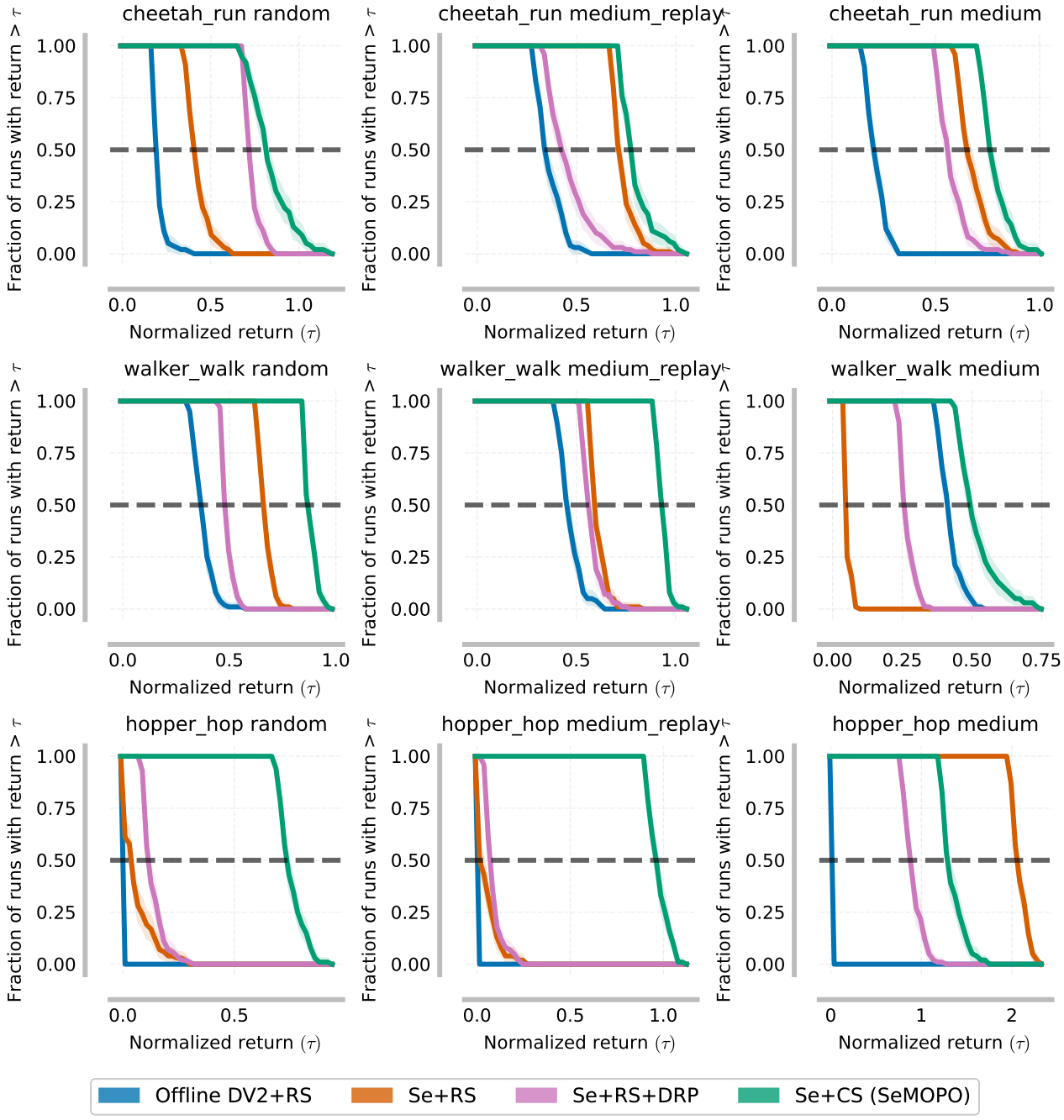

Figure 10. Performance evaluation results of ablated methods of SeMOPO on the LQV-D4RL benchmark for 200 test episodes. Shaded regions represent pointwise 95% confidence bands based on percentile bootstrap with stratified sampling (Agarwal et al., 2021). Removing any component of SeMOPO leads to a performance drop.

Table 5. The unnormalized test returns of different methods on the LQV-D4RL benchmark. Mean scores (higher is better) with standard deviation are recorded across 4 seeds for each task.

Dataset		SeMOPO	Offline DV2	LOMPO	DrQ+BC	DrQ+CQL	BC	InfoGating
Walker Walk	random	**459** ± **41**	167 ± 36	133 ± 40	27 ± 1	26 ± 2	25 ± 2	49 ± 7
	medrep	**521** ± **42**	175 ± 31	218 ± 69	27 ± 3	25 ± 3	28 ± 7	58 ± 22
	medium	273 ± 47	68 ± 26	63 ± 15	390 ± 36	25 ± 3	**413** ± **30**	98 ± 41
Cheetah Run	random	**254** ± **30**	40 ± 15	66 ± 18	94 ± 30	0 ± 0	22 ± 20	58 ± 16
	medrep	258 ± 28	66 ± 31	78 ± 32	166 ± 93	1 ± 0	20 ± 18	**267** ± **52**
	medium	**293** ± **32**	81 ± 57	52 ± 37	260 ± 28	0 ± 0	252 ± 40	288 ± 39
Hopper Hop	random	58 ± 5	0 ± 0	0 ± 0	7 ± 8	0 ± 0	5 ± 6	**67** ± **11**
	medrep	**77** ± **6**	0 ± 0	0 ± 0	21 ± 15	0 ± 0	3 ± 1	45 ± 14
	medium	**105** ± **14**	2 ± 4	1 ± 3	68 ± 16	0 ± 0	35 ± 5	49 ± 8
Humanoid Walk	random	**6** ± **3**	3 ± 1	2 ± 2	1 ± 1	3 ± 2	1 ± 1	3 ± 2
	medrep	7 ± 4	2 ± 1	5 ± 2	2 ± 1	2 ± 2	**10** ± **7**	2 ± 2
	medium	6 ± 5	4 ± 2	3 ± 2	**14** ± **7**	2 ± 1	4 ± 2	4 ± 3
Car Racing	random	**418** ± **79**	233 ± 44	387 ± 89	-10 ± 32	-92 ± 1	-59 ± 13	-79 ± 5
	medrep	**362** ± **87**	181 ± 88	325 ± 184	-68 ± 8	-93 ± 1	-80 ± 3	-76 ± 4
	medium	**408** ± **158**	285 ± 153	313 ± 108	180 ± 101	-83 ± 1	-67 ± 9	-67 ± 6