Reviewer #1

Q1: ...the claim that the algorithm really manages to align the latent distributions of real and simulated data...

A1: The goal of model adaptation is to align the feature distributions, but in the meantime we need to control the trade-off between model training and model adaptation to ensure the representations to be invariant and also discriminative. We will revise the inappropriate statements in the final version.

Q2: In the model adaptation phase, are state-action pairs simply sampled randomly from their respective buffers?

A2: State-action pairs are randomly sampled since model adaptation is distribution level. Actually this objective doesn't minimize the feature distance of arbitrary (s, a) pairs. Instead it minimizes the distance between feature distributions of two data sets. Constraining nearby (s, a) pairs to have similar features is more related to the Lipschitz continuity of NN.

Q3: How important is the division in feature extractor and decoder? Do you have results for a single, monolithic model?

A3: AMPO uses the same model architecture as MBPO, which can be regarded as a single monolithic model. The model is *conceptually* divided as feature extractor and decoder and one can regard it as a monolithic model. We propose to add a model adaptation loss over the output of feature extractor, which encourages such a conceptual division.

Q4: Did you investigate the reasons for the slow learning in the 500 steps on InvertedPendulum compared to PETS?

A4: The reason may be that MPC performs well in the environments with low action dimensions (1 in InvertedPendulum), which also holds in the experiments in the PETS paper, since it is easy to find good actions with limited action proposals.

Reviewer #2

Q1: The experiments shown in Figure 2 do not outperform MBPO beyond the confidence bounds.

A1: AMPO does outperform MBPO according to the results since the shaded area corresponds to standard deviation. For example, if five trials are $[1, 3, 3, 3, 3]$, then the mean is 2.6 and the standard deviation is 0.8. But the maximum value of shaded area as shown in our plots is 2.6+0.8=3.4, which surpasses the maximum value in the five trials, i.e., 3.

Q2: Exploration has no meaning if these are not samples from the real world, only samples from the model?

A2: We also need exploration when sampling data with the model. Imagine that the model is extremely accurate and we use the policy to sample data only with the model, then exploration is also needed to find a good policy.

Q3: Can you elaborate more why you choose the asymmetric feature mapping strategy?

A3: Asymmetric feature mapping (unshared weights) has been shown to outperform the weight-sharing variant in domain adaptation, due to more flexible feature mappings. This also holds in our experiments as shown in appendix.

Q4: Other medium and small points: ...better policy optimization in MBRL...both buffers...consistent color...etc.

A4: Thanks for your suggestions. We will fix these problems accordingly.

Reviewer #3

Q1: The proofs look very similar to the MBPO paper. The contribution to the theoretical part is quite incremental to me.

A1: Our analysis is based on occupancy measure, while MBPO decomposes to each timestep. Moreover, our analysis directly enhances the model training process while MBPO focuses on model usage rather than model training.

Q2: ...it is unclear to me how one can incorporate the third term in the policy optimization...

A2: We can use imitation learning to optimize this occupancy measure matching term over π, such as GAIL, where the collected real samples are viewed as the expert and the policy is run on the model to sample data. However, for the alternative training scheme of policy and model, optimizing this term over π is not necessary, which may further reduce the efficiency of the whole training process. For example, when the model is sufficiently accurate, one does not need to further optimize π using this term but just focuses on the $\hat{\eta}[\pi]$ term. Thus like we omit the model optimization in $\hat{\eta}[\pi]$, we also omit the policy optimization in this term. We are happy to provide more discussions on this in the final version.

Q3: Can a different distribution matching metric (other than Wasserstein-1) improve the performance?

A3: We have experimented with the MMD variant and observed good results. We will include this in the final version.

Reviewer #4

Q1: The explanation of model adaptation needs improvements. Are there prior works that use ... for MBRL?

A1: We will polish the corresponding writing in the final version. As far as we know, there is no such prior work.

Q2: Is it hard to tune the hyperparameters and architecture for the model adaptation?

A2: The main hyperparameter needed to tune is adaptation iterations, and it won't cost much to find a good one as shown in Fig.4(b). We use the same model architecture as MBPO, and choose first several layers as feature extractor.