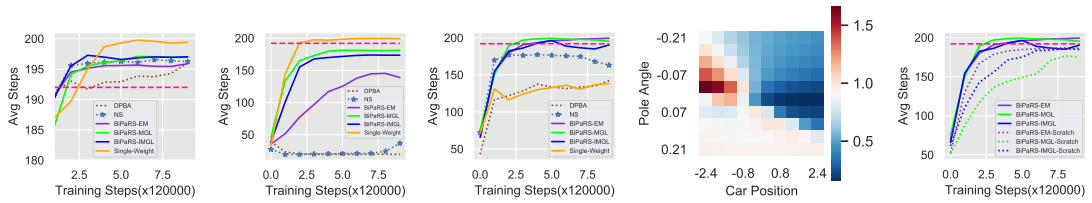

(a) (b) (c) (d) (e)

Figure 1: Results of the additional experiments. Figs 1(a) and 1(b) are the results of the single-weight baseline method in continuous carpole with beneficial and harmful shaping rewards, respectively. Fig 1(c) is the result of the second experiment and Fig 1(d) is the heat map of BiPaRS-EM's shaping weights in different states. Fig 1(e) shows the comparison between our methods with normal and zero initialization of the shaping weights

We thank all reviewers for the valuable comments and suggestions. Our responses to the main concerns are as follows.

1. **Limitation of the experiments** (proposed by reviewer 1): We conduct three additional experiments in the continuous cartpole task to answer the reviewer's questions.

(1) In the cartpole experiment in Section 5.1 and the first adaptability test in Section 5.3, the shaping weights learnt by our methods differ slightly across the state-action space. So learning a state-action-independent shaping weight in the two tests is truly sufficient. We implement the baseline method which replaces the shaping weight function z_ϕ with a single shaping weight and test it in the two experiments. The results are given in Figs 1(a) and 1(b) and it can be found that the single-weight method outperforms the other methods.

(2) As suggested by reviewer 1, we also conduct an experiment where half of the shaping rewards are helpful and the other half of the shaping rewards are harmful. It can be found from the results in Figure 1(c) that our methods perform the best and the single-weight baseline method cannot perform as well as in Figures 1(a) and 1(b). We plot the shaping weights learnt by the BiPaRS-EM method across a subset of the state space (containing 100 states) as a heat map in Figure 1(d) to show that our methods are able to learn state-dependent or state-action-dependent shaping weights.

(3) We conduct the third experiment to test the baselines methods which learn shaping rewards from scratch. According to reviewer 1's suggestion again, such baseline methods are simply implemented by our BiPaRS methods with zero initialization of the shaping weights. For convenience, the setting of the shaping rewards is the same as the second experiment and the results are shown in Figure 1(e). It can be found that when learning from scratch, all our methods fail to learn as well as their normal versions where the shaping weights are initialized to 1. The zero initialization of the shaping weights means that the prior knowledge incorporated in the shaping rewards is invisible to the algorithm in the beginning and this may lead to more effort of exploration.

2. **Saftey concern in MuJoCo** (proposed by reviewer 3): In fact, we adopt the MuJoCo setting of the RCPO paper in our experiment because it provides a good shaping reward function for the MuJoCo tasks. Although such shaping reward function is originally used as constraints, we only care about whether our methods can obtain higher true rewards and how they will do if the shaping rewards are in conflict with the true rewards.

3. **Comparison with PBRS method** (proposed by reviewer 4): The PBRS family mainly focuses on the guarantee of policy invariant, and our methods are proposed for solving the utilization problem of given shaping rewards. Although most PBRS methods have the policy invariant property, whether an optimal policy can be learnt by a learning algorithm (especially a DRL algorithm) does not totally depend on this. In the function approximation setting, how to utilize the shaping rewards to learn a good policy seems more important than just keeping policy invariant. Furthermore, BiPaRS also has policy invariant property because the solution of its objective is an optimal policy of the original MDP.

4. **Relation to the optimal reward framework** (proposed by reviewer 4): Our methods are essentially different from the optimal reward framework (ORF). Firstly, an ORF method such as the LIRPG algorithm (reference [25] in our paper) is similar to our BiPaRS-MGL method. But our first method BiPaRS-EM, and the third method BiPaRS-IMGL, cannot be directly derived from LIRPG. Secondly, all our methods are based on Theorem 1, which actually is more general than Eq. (5) in the LIRPG paper. It may be better to say that "our methods and ORF are special cases of meta-policy gradient methods" than to say that "our methods are special cases of ORF". The shaping weights learnt by our methods are evaluation and guide of the utilization of the shaping rewards. Perhaps we can treat $z_\phi(s,a)f(s,a)$ as an intrinsic reward, but $z_\phi(s,a)$ itself is not.

5. We will correct the typos and modify our paper according to the comments of the reviewers (e.g., improving the related work section, providing comprehensive comparison between our methods, and adding more experiments).