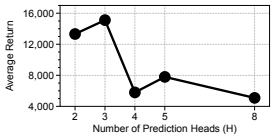
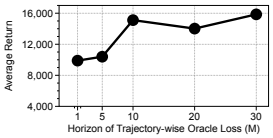
We thank all reviewers for carefully reading our paper and their valuable comments. We appreciate that our paper is recognized for several positive aspects: [R1, R2, R4] novel and well-motivated, [R2, R3, R4] clear write-up, and [**All**] extensive and strong experiments. Below are our responses to the reviewers, which we will incorporate in the final draft.



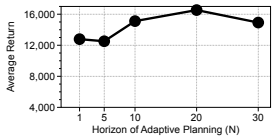
(a) Prediction heads

Mass	Head 1	Head 2	Head 3	Head 4	Head 5
0.25	0.0	0.0	0.0	100.0	0.0
0.50	0.1	0.0	0.0	0.0	99.9
1.50	56.3	43.7	0.0	0.0	0.0
2.50	0.0	44.3	55.7	0.0	0.0

(b) Trajectory assignment



(c) Trajectory-wise oracle loss



(d) Adaptive planning

[R2, R3, R4] Effects of major hyperparameters. We choose the number H of prediction heads and the horizon M of trajectory-wise oracle loss based on the trajectory assignments in training environments, i.e., how distinctively trajectories are assigned to prediction heads (see Figure 4 and 5 of the original draft and above Figure (b) for examples). Also, the horizon N of adaptive planning is set to the same value as M to match the horizon of trajectory-wise oracle loss and adaptive planning. We use the same hyperparameters H, M, and N for all environments. We will clarify the hyperparameter setups in the final draft.

Following the suggestions of R2, R3, and R4, we also perform ablation studies in HalfCheetah environments by varying the major hyperparameters of our method: $H \in \{2, 3, 4, 5, 8\}$, $M \in \{1, 5, 10, 20, 30\}$, and $N \in \{1, 5, 10, 20, 30\}$. Figure (a) shows that $H = 3$ achieves the best performance because three prediction heads are enough to capture the multi-modality of the training environments in our setting. When $H > 3$, the performance decreases because trajectories from similar dynamics are split into multiple heads as shown in Figure (b). However, as pointed out by R4, we expect that more heads would be effective in environments with more varying environmental factors. Figure (c) and (d) show that choosing the head per-trajectory is more effective and our method is robust to change in hyperparameters $M \geq 10$ and $N \geq 10$. We will include more comprehensive results for all environments in the final draft.

[R1] Limitation of our method. Thank you very much for your pointers. As we assume that MDPs with similar dynamics will behave similarly, the effectiveness of our method would be limited if the dynamics of unseen environments are significantly different from the dynamics of training environments. As you pointed out, gain from our method may decrease if the agent is not exposed to a variety of MDPs. We will clarify these limitations in the final draft.

[R1] Clarification on problem formulation. Our work addresses the dynamics generalization problem of model-based RL methods, where learned dynamics models fail to provide accurate predictions as the transition dynamics of environments change. Thank you for your suggestion, and we will clarify this in the final draft.

[R1] Clarification on multi-modal nature of environments. We assume that the transition dynamics distribution of MDP is multi-modal, which emerges as the environmental factors (e.g., mass and length of the agent) change. We remark that capturing this property is important as environmental factors are ever-changing in real-world environments. Prior model-based RL methods that do not consider this property fail to generalize as the future prediction from dynamics models becomes inaccurate. We will clarify this in the final draft.

[R2] Editorial comment. We will remove the "Due to space limitation" phrase in the final draft.

[R3] Novelty. As R1, R2, and R4 pointed out, we believe that we propose a novel and well-motivated combination of multiple choice learning and model-based RL, whose major components are (i) context-conditional multi-headed dynamics model, (ii) trajectory-wise oracle loss, and (iii) adaptive planning. We also verify the effectiveness of the proposed method for improving dynamics generalization via exhaustive ablation studies on various benchmarks.

[R3] Extension to other model-based RL methods. Applying our trajectory-wise multiple choice learning scheme to model-based policy optimization methods (e.g., MBPO and Dreamer) is an interesting direction, and we think our method is naturally applicable. For example, one can consider employing our method for learning specialized policies using specialized prediction heads. We leave this as future work, but will include related discussion in the final draft.

[R3] Broader Impact. We will add more discussion related to the broader impact of our work to the final draft.

[R4] Comparison to PEARL. We emphasize that our method is evaluated using the zero-shot generalization performance in test environments, while PEARL is evaluated using the few-shot adaptation performance, i.e., PEARL conducts adaptation to test environments before evaluation. Our method still achieves superior sample efficiency compared to PEARL in most environments and outperforms PEARL in HalfCheetah and CrippledAnt environments even in terms of asymptotic performance. We also remark that as in PEARL, it is possible to learn a context-conditional policy using context vectors from our method (e.g., see [1]) to further improve performance.

[R4] Qualitative analysis on control tasks. This is a very interesting question. Following your suggestion, we visualize the behavior of a cheetah agent using the prediction head specialized for low-mass environments on the HalfCheetah environment with a default body mass. Interestingly, we observe that the cheetah agent moves as *if* it has a lightweight body, i.e., moving its limbs about very fast. We will include videos in the final draft.

[1] K. Lee et al. Context-aware dynamics model for generalization in model-based RL. In *ICML*, 2020.