We thank all the reviewers for the time and expertise invested in these reviews. We have corrected the typos and grammar mistakes, and rephrased some points/sentences to improve the overall readability of our paper. In the following sections, we respond to the major concerns of each reviewer respectively to, hopefully, clarify our methods and visions.

Responses to Review # 1

Q: What is the meaning of every notation? A: We are sorry that some abuse of notations in the paper hinders the understanding of our method. We have checked the notations to ensure that they are consistent throughout the whole paper. We would further explain that, generally, we use K to denote the number of ensemble models and H to denote the maximum horizon. Their corresponding lowercase letter refer to one instance in the set, e.g. s_{t+h} means a state in $\{s_{t+1}, \ldots, s_{t+H}\}$. Specially, s_g means a "good" state in $s_{t+1}, \ldots, s_{t+H}$ that can be regarded as a goal state (to be consistent with notations in other goal-conditioned work).

Q: What is the relationship to other Transfer Learning/Imitation Learning method? A: This work aims to tackle a problem that lies in the intersection of Imitation Learning and Transfer Learning (more specifically Sim2Real).We have included the most relevant works that tackle a similar problem and demonstrated our difference and novelty. We would like to conduct a complete literature review later to cover the recent works in Imitation Learning and Transfer Learning and add them to the related work section.

Since there are no major flaws pointed out in the review, could the reviewer please raise the overall score?

Responses to Review # 2

Q: Can this method work in real-time control requiring a reset-able simulator? A: Generally, the computation cost of a reset-able simulator is comparable to a model-based method and is thus acceptable. We will explore how to relax such a constraint in future work from both methodological and engineering perspectives.

Q: Can this method work considering the complexity of real dynamics mismatch? A: Our empirical results on various modifications to MuJoCo environments (3 types $\times$ 3 magnitude) can prove that our method is robust to different dynamics mismatch, so we believe it shows the potential to work in sophisticated real-world problems. We have plans to apply the method proposed in this work on a real quadruped robot in the future.

Responses to Review # 3

Q: What is the difference compared with others using Goal-conditioned Policy(GCP)/Hindsight Inverse Dynamics(HID)? A: HID adopted in our work contributes to the overall purpose to alleviate dynamics mismatch problem and augment the limited expert data. Other works use GCP/HID for different purposes. PCHID (arXiv 1910.14055) solves goal-oriented tasks in a supervised manner. Play-GCBC (arXiv 1903.01973) trains a goal-conditioned policy to address a multi-modal problem. Relay Policy Learning (RPL) mentioned in (arXiv 1910.11956) solves long-term robotic tasks in a hierarchical manner with the help of a goal-conditioned policy. We will discuss these studies in the related work section.

Q: Can this method work considering the complexity of real dynamics mismatch? A: Our empirical results on various modifications to MuJoCo environments (3 types $\times$ 3 magnitude) can prove that our method is robust to different dynamics mismatch. So we believe it shows the potential to work in sophisticated real-world problems. We have plans to apply the method proposed in this work on a real quadruped robot in the future.

Responses to Review # 6

Q: What is the meaning of "partial alignment" A: We realize that the word "partially" in Line 9 is a little bit confusing and have clarified it. Instead of referring to environment being "partially observable", it means that not every (s_t, s_{t+h}) pair needs to be matched, and a subset (which "partial" actually means) matching is enough. Please refer to Figure 3 in the paper for a visual illustration. All the experiments are done in fully observable MuJoCo environments.

Q: What is the rationality behind "partial alignment" assumption? A: Such an assumption comes from an empirical observation that in robotics control problems, some key poses in different dynamics are still alike. Aligning such key poses would make the long-horizon learning much easier. Besides, a weighted GAIL(GAILfO) optimization in our method is a relaxation to the "exact partial alignment". It can be viewed as an occupancy matching problem that automatically matches the most "similar" state. We would like to explore further in this area to see if this method will work under some formalization of dynamics differences.

Q: Do we use the simulator reward in the training phase? A: No, we do not use simulator reward functions in any training phases, i.e., both GAILfO and Goal-conditioned BC (which accords with the usual Imitation Learning setting).