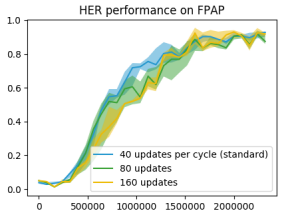
Firstly we would like to thank all of the reviewers for their insightful and thorough feedback. We really appreciate the time taken to go through the paper and will try our best here to respond to as many of the questions raised as possible.

To start with, three of the reviewers noted that we had not included any comparisons with other model-based approaches. Our initial reasoning for not including such a comparison was that because standard model-based methods generally do not outperform model-free methods in terms of their final performance, it seems implausible that a model-based approach designed for standard environments (i.e. dense reward, not goal-based) would perform better than, e.g., DDPG, which we do include as a comparison and which performs badly on the more challenging environments. Having thought about this more, however, it does seems natural that such a baseline should still be included to verify this intuition, and we will look to add such a baseline in the next draft. We also thank **reviewer 1** for drawing our attention to a number of papers that we missed out in the related work.

Reviewer 2 raises a point about the baselines potentially not being anchored with previous numbers. Whilst it's true that we ran all of the baselines ourselves, we nevertheless used the author's publicly released code for all of the baselines, except HER. For HER, we used the OpenAI Baselines implementation. If we compare with Plappert et al.[1], they show results for FetchPush and FetchPickAndPlace in their Figure 3 (note 1 epoch = 95000 environment interactions). Although perhaps not immediately obvious, if you carefully compare our results with theirs using this scaling you should see that they (at least approximately) match up. To address some of reviewer 2's other questions: we agree that it would be an interesting line of work to explore similar approaches for more general environments. The reasons we focused on sparse reward, goal-conditioned environments is that it felt natural to combine the training of the GANs with hindsight experience relabelling (which is really at the core of our approach), as well as the fact that typical model-based approaches are likely to be inappropriate for solving these kinds of task. Regarding the comment about l199, it's true that other strategies can be used here. In our initial experiments we did try simply using a single future goal for the whole unrolled trajectory, but found this did not perform quite as well. At the very least this should have been mentioned, and so we will include a comment and perhaps consider adding an ablation demonstrating this. We appreciate the comments re: l195, l217 and l225 and agree that your suggested terminology would be clearer. Re: the comment about l227, we think this is a slight misunderstanding and something we need to make clearer in the text. All GANs in the ensemble are trained from batches taken from the same replay buffer. For each step of each imaginary rollout we choose a random GAN in the ensemble, however these imaginary rollouts are not stored in the buffer (which only stores the actual rollouts with the final actions chosen by the planner, along with some initial random trajectories). This is related to a comment by **reviewer 3** who asks for clarification about this as well as asking how important "mixing" GANs in this way was. Essentially, this is the primary way in which the ensemble (rather than just using a single GAN) was made use of, so in that sense the ablation in Appendix B demonstrates that it does have a somewhat significant effect. However, it is true that there are other ways we could have made use of the ensemble, but we have not considered these so far. Re: the comment about l235, R is a typo (thank you for spotting this). We weren't 100% sure what you meant by "...(not an integer, depends on ordering...)", but just to clarify (and we will adopt this notation in the next draft): the planning process gives a score, n_i to each of the Q initial seed actions a_i (note the actions are continuous). We then define weights $w_i = e^{\alpha n_i}$ and return $a_t = \frac{\sum_{i=1}^{Q} w_i a_i}{\sum_{j=1}^{Q} w_i}$.

To address some of **reviewer 3's** other questions: the training curves do indeed account for the initial random trajectories (including those we discard and don't store in the replay buffer). M in figure 1 is supposed to represent the one-step predictive model, although we realise now there is an error in this (M should have both s_t and a_t as inputs) — we will fix this. We did consider removing the OSM regularization, but decided not to as we felt it was still an interesting experiment, despite the fairly small difference it made (having said this, the difference on FPAP was $\sim 5-10\%$, which looks small on the plot but is not entirely insignificant). Given that there was not much difference between noOSM and OSM, it's also the case that there is not any strong dependence on λ.



Addressing some of **reviewer 4's** comments: it's true that we use more updates than HER when training our model-based approach. However, previous work has shown increasing $num\_updates$ can actually degrade HER's performance slightly. We ran some experiments on FPAP to demonstrate this (and could add these to an appendix perhaps) — see the plot on the right. It's an interesting question to ask whether HER could also be improved by using an ensemble — however exactly *how* you would best make use of such an ensemble of policies/value functions in that context is not entirely trivial and an interesting research question in itself. We feel this is also true of employing HER with a forward model (again, a very interesting suggestion) — would you train DDPG+HER within a learned world model or use some other kind of model-based search to gather rollouts? How often would you update the model (or continuously update)? There are a number of considerations that, whilst interesting, go beyond a very straightforward baseline to compare with, in our opinion.

[1] Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research, 2018