We thank all the reviewers for their detailed and thoughtful comments!

R3

R3's two main requests are (1) a comparison against a variant of HER that relabels with future states and (2) a discussion of two prior works.

Future state relabeling: We have already compared against the variant of HER that relabels with future states ("Future State Relabelling" is the green diamonds in Fig 5). This baseline consistently performed worse than random relabeling. Note that this baseline is not applicable in the setting of more general task distributions (Fig 6).

Differences with prior work: Whereas [Zhao 20] and [Pitis 20] focus solely on goal-reaching tasks, our work is applicable to tasks beyond goal-reaching, such as discrete sets of tasks, linear reward functions, and more general task distribution (see Fig 6). While both our work and these works all mention "maximum entropy," the actual contributions are orthogonal and substantially different:

- [Zhao 20] propose a method for prioritizing experience in a replay buffer.
- [Pitis 20] propose a method for sampling goals for exploration.
- We propose a relabeling method for sharing experience across tasks.

The actual algorithms implemented by [Zhao 20] and [Pitis 20] use a combination of previously-introduced relabeling strategies: [Zhao 20] uses HER and [Pitis 20] uses a combination of "final state", "future state", and "no relabelling". We have already compared against each of these relabeling strategies in our goal-reaching experiments (Fig 5). Note that only the "no relabelling" baseline is applicable to tasks beyond goal-reaching (Fig 6). We'll include a discussion of both papers in the camera-ready version.

R1

Thanks for the writing suggestions! We will revise the paper to (1) discuss the concurrent work [Li 20], (2) add a discussion for why random relabeling works so well, (3) increase the brightness of error bars in Fig 5, (4) move the discussion limitations from the Broader Impact to the Discussion, and (5) fix the noted typos.

R2

How does HIPI do Exploration? Our focus in this paper is how to use previously-collected experience to solve multiple tasks. How that data is collected is largely orthogonal. In our experiments, we simply sample a task from the prior, $\psi \sim p(\psi)$ and then take actions using corresponding (stochastic) policy, $\pi(a \mid s, \psi)$.

Writing: Thanks for the suggestions! We'll (1) clarify how SAC fits into HIPI, (2) clarify how exploration is done, and (3) fix the figure ordering. We'll also include a discussion of your observation that approximate inverse RL will assume the best-seen trajectories for some task are optimal for that task.

R4

Effect of batch size: We ran an additional experiment varying the batch size used by HIPI-RL on the sparse 2D reacher. Fig. 1 (right) shows that increasing the batch size significantly improves performance, suggesting that better approximate inverse RL results in better performance. We used a batch size of 32 for the results in the paper, but this experiment suggests that we could have gotten stronger results by using a larger batch size.

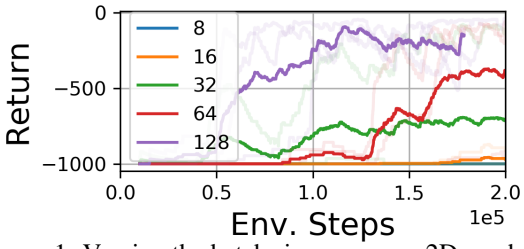

Figure 1: Varying the batch size on sparse 2D reacher

Visualizing the inferred goals: The figures below visualize the inferred goals on the gridworld example (Sec. 6.1). Each subplot corresponds to a different transition, denoted by the orange arrow. Dark blue cells denote likely goals, while white cells denote unlikely goals. When the dynamics are modified so the agent cannot move left (Fig. 3), states to the left of the agent are no longer inferred as likely goals.

Figure 2: Original gridworld.

Figure 3: Modified gridworld where agent cannot move left.