Method	HC Rand Vel	Walker Returns	Walker Custom Metric	Ant Lin Class Success Rate
Meta-DAgger	**0.082 ± 0.059**	2404.1 ± 91.7	0.275 ± 0.089	**0.67 ± 0.00**
SMILe (state-action)	**0.118 ± 0.078**	**3106.6 ± 67.6**	**0.710 ± 0.036**	**0.68 ± 0.01**

Table 1: Meta-Dagger results compared to SMILe (SMILe results taken from the paper)

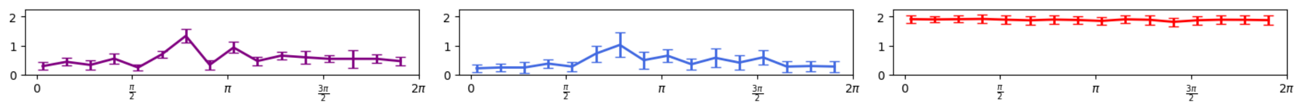

Figure 1: Ant Random Goal, Final Distance to Goal: Meta-DAgger (Left), SMILe (Middle), Fully Observed RL (Right)

We thank the reviewers for their careful reading and constructive comments, and have already updated the manuscript.

New Experiments: **R1, R4 (Stronger Meta-Imitation Baseline)**: As an additional baseline for Meta Imitation Learning (Meta-IL) we implemented a meta variant of DAgger and tuned it carefully. The implementation of Meta-DAgger is similar to Meta Behavior Cloning with the difference that in each epoch expert policies are used to label an additional set of policy rollouts. Importantly, Meta-DAgger operates under the much stronger assumption than SMILe that expert policies are always available to give the correct actions for states visited by the student policy. Meta-Dagger therefore serves as one of the strongest Meta-IL baselines. As shown in Table 1 and Figure 1, SMILe's performance is comparable to Meta-DAgger in all tasks, except for Walker Random Dynamics where SMILe achieves significantly better performance.

R1, R4 (Meta-RL Baseline): One of the fundamental advantages of Imitation Learning compared to standard RL is in sparse reward tasks. As such, in the Ant 2D Goal task we trained "fully observed" policies (using SAC) which observe task parameters as part of the state; such policies are strictly stronger than any Meta-RL baseline such as MAML or PEARL. The reward function is 1.0 if the agent is within 0.5 radius of the target and 0 otherwise. As shown in the Figure above, fully observable policies were not able to solve this sparse reward task. We will incorporate these as well as Meta-DAgger results in our manuscript.

Addressing Other Comments R1,R3,R4 (Significance of Contribution): To our knowledge, the only prior Meta-IRL method that scales to the function approximator setting is the work of Wang et al. The relative contributions of our method are not limited to computational gains. Importantly, **SMILe does not require learning a generative model over trajectories**. Wang et al. train a VAE on expert trajectories, where a decoder must learn to both imitate expert policies and learn a good generative model of environment dynamics. Learning good dynamics models is a difficult, active problem studied by the model-based RL research community. This problem will become even harder as the community seeks to scale meta-IRL methods to domains with richer observation spaces (e.g. image observations).

R3 (Off-policy as Contribution): We do not consider off-policy training a core contribution of our work. It is merely used to make training of our models significantly faster in terms of wall-clock time. Lastly, we agree that our method does have a number of similarities to the Meta-*RL* work of Rakelly et al as discussed in Related Work. We note that their work is concurrent (first posted on arxiv 2 months prior to Neurips deadline), and they are addressing the Meta-*RL*, not the Meta-IRL (Meta Imitation Learning) problem.

R4 (Testing Harder Generalization): In the Walker Random Dynamics task, during training models observed 50 random settings of dynamics variables. Generalizing to the 25 testing dynamics is highly non-trivial. SMILe's respectable performance gives us confidence about the quality of adaptive policies learned, given our new result that Meta-DAgger performs substantially worse than SMILe on this task. We believe these results demonstrate that SMILe is an important contribution to Meta Imitation Learning. **R4 (Clarifying Ant Linear Classification)**: R4 is correct in their interpretation of our task setup. As part of their state, agents observe two 4-dimensional vectors. If the first point is from the positive class the agent should navigate to the first target, otherwise it should navigate to the second target. When generating expert demonstrations we use our knowledge of the correct label to choose an appropriate expert policy from the Ant 2D Goal task, which ignores the two 4-dimensional points in the state. We have updated Section 4.4 in our manuscript with these clarifications. **R4 (Connecting E to C in Graphical Model)**: The graphical models in our manuscript describe the generating process for states and actions under a policy π. If we know the task identity (T observed), knowing the environment (E observed) the policy is being rolled out in does not provide additional information about what context (C) the policy conditioned on. In contrast, if T was not observed, E would tell us that the policy conditioned some context coming from a task that is compatible with E. Hence we do not believe there should be an arrow from E to C. **R4 (Parameterizations)**: All models (encoder, discriminator, policy, Q, and V) architectures are provided in the Appendix G.2 of our manuscript. We used the Adam optimizer throughout. **R4 (Entropy in Policy Objective)**: In contrast to GAIL, the AIRL algorithm does not include a causal entropy term. **R4 (exp_base Notation)**: We used exp_base to denote the expert policy for the Walker environment with unmodified dynamics. We will remove this notation from subsequent revisions as there is no need to name this policy explicitly.