We would like to thank all reviewers for taking the time to review our work and for providing thoughtful suggestions.

R1: Avoid misattribution by using "original" space? Real-world problems often have no obvious or easily engineered "original space", as the success of deep feature learning attests. Very natural-seeming state spaces may still include nuisance factors—as in our dashboard light setting (Fig 1), where causal misattribution occurs when using the full image from the camera (scenario A). Our testbeds (Sec 3.2) do indeed introduce more deliberate nuisance variables for ease of evaluation, but evidence suggests that misattribution is pervasive in common imitation learning settings. For example, history would seem a natural part of the original state space for real-world driving, yet as shown in Pg 4, recurrent/history-based imitation has been observed repeatedly in prior work to hurt performance.

R3: better intervention methods? Policy execution and expert queries are intervention modes that are close to reinforcement learning and DAgger-style behavior cloning respectively, and inherit their strengths and weaknesses. Specific settings might indeed warrant other types of interventions for safety/practicality. As an example, the learner could solicit preferences (see Wirth et al, "A survey ...", 2017) that indicate that the interventional episodes of some hypotheses G are to be preferred over others, so that the expert need not be parachuted or placed in dangerous states. In all cases however, an intervention *must* involve some action by a suboptimal policy in the environment, which naturally incurs some risk—after all, causal misidentification can only be detected under distributional shift from the demos.

R3: What if state isn't disentangled? Then, individual dimensions in the state might capture both causes as well as nuisance variables. The problem of discovering true causes is no longer reducible to searching over 2^m graphs. To test this empirically, we create a variant of our MountainCar testbed, where the 3-D past action-augmented state vector is rotated by a fixed, random rotation. After training the graph-conditioned policies and applying 30 episodes of policy execution intervention or 20 expert queries, we get -145 and -165 reward respectively. This is significantly lower than in the disentangled (non-rotated) setting, indicating disentanglement is important for the effectiveness of our approach.

R1: Why disagreement? Choosing new samples to label based on disagreement among a committee (Seung et al 1992, "Query by Committee") is a widely used heuristic in active learning, with this simple intuition: to disambiguate among competing hypotheses G, the most informative states are those that induce maximum disagreement among the hypotheses. Our empirical results, in agreement with many prior active learning approaches, suggests that this heuristic works well in practice. Still, it is an imperfect heuristic, and it is certainly possible to construct cases, as R1 suggests, where useful states may be discarded because of high agreement among bad hypotheses. However, Alg 1 works because it need not identify every single useful expert query intervention — it suffices to identify a small subset of good interventions. It might be possible to do better than Alg 1 by updating the distribution over graphs after each query and recomputing the disagreement based on this rebalanced distribution. This would iteratively bias the distribution towards better-performing graphs, so that for later queries, the disagreement score would more effectively measure mismatch between good and bad graphs. We will experiment with this for future versions.

'R1: What exactly is π_{mix} in Alg 1?' π_{mix} samples a random graph G per episode and then executes π_G, with G concatenated to the state. We will clarify this. **R1: how many G's?** $|G| = 2^m$, with m being $2 + 1$ (state+action dimensions) for MountainCar, $11 + 3$ for Hopper and 30 for Atari (dimensionality of VAE latent). We observe that the number of interventions required, N, increases as $|G|$ increases (see Fig 7 + Ln323-339). **R1: DISC-INTERVENTION in Tab 2?** DISC-INTERVENTION, which employs a variational approach to causal discovery, gets progressively harder to train as the state space increases. Already on 14-D Hopper states, it does no better than UNIF-INTERVENTION. On Atari envs in Tab 2, DISC-INTERVENTION ran into optimization difficulties and yielded poor results. We will add a note.
R3: Typos & appendices. We will fix the typo $Z_t \rightarrow Z^t$. The reference to Eq 3 in page 7 should be Eq 1, we will fix this. We will attempt to move more appendices into the main paper, and we will release source code to remove any ambiguity. **R3: update time cost?** Yes, the model is updated after each episode, but this is very fast online linear regression, taking negligible time compared to executing an episode from the neural net policy. **R3: size of disentangled state space?** For Atari, we set the VAE latent size heuristically, to be as small as possible, but still produce good reconstructions, as assessed visually. **R3: DAgger sparse evaluation?** As shown in Fig 11, DAgger takes 20x for MountainCar and 600x for Hopper more queries to match the performance of our method — to keep the number of experiments and computation costs manageable, its performance is evaluated more sparsely. **R3: High variance in Hopper?** This is down to Hopper being inherenty unstable, where some random seeds for all methods result in the Hopper falling over, producing very poor rewards. **R3: More equivalent baselines than GAIL?** GAIL is a strong baseline for imitation, and we do not know of other methods that would be more equivalent/competitive with our approach in these settings. **R3: DISC-INTERVENTION good?** DISC-INTERVENTION does indeed perform well on low-dimensional state spaces (e.g. MountainCar), but runs into optimization difficulties as the state space grows. Also see our response to R1 above.
R2: Does performance degrade with more interventions? In all our experiments, performance was approximately monotonic: with more data from interventions, performance either stabilizes or significantly improves. We will evaluate UNIF-INTERVENTION at more expert queries in future revisions for a more complete version of Appendix Fig 11.