We thank all reviewers for insightful comments. *All existing references are numbered per the bibliography in appendices*.

[Reviewer 1] • **Language**: As suggested, we have now standardized the paper for formality and removed colloquialisms.

[Reviewer 2] • **Language**: Kindly refer to response for Reviewer 1. • **What information is lost in behavior cloning**: We have now included (after Line 55) an example pertinent to us: the state visitation distribution of the demonstrator—which results from dynamics—is information in the data, but which BC discounts by focusing only on action conditionals. • **Line 112, "directly parameterizing a policy"**: You are correct; we are not referring to the parametric/ non-parametric distinction. Instead, we are distinguishing: ① methods that first learn a *reward function* [12–19, 32–37], thus indirectly inducing a policy (for that reward), vs. ② methods that directly learn a *policy mapping* [11, 20–25, 41]. We have now clarified this (after Line 113). • **Line 116, "intrinsically batch"**: This simply refers to an algorithm that operates offline *without* recourse to off-policy evaluation (vs. off-policy adaptations of online algorithms). We have now clarified this (after Line 117), and also explicitly reference Table 1 for its elaboration. • **Energy-based learning**: You are correct; we rely on a standard technique from statistical physics, and we agree it is more informative—and responsible—to clearly state this from the get-go. We have now updated the manuscript to properly introduce EBMs after defining the objective, importantly invoking the joint EBM technique [44–46]—with particular credit to Eq. 7 in [44] for popularizing the "discriminative+generative" model that is our Eq. 10. This introduction now (appropriately) replaces the admittedly decorative (and potentially confusing) presentation of Lemma 1. • **Gradient estimator**: Yes, the choice of step size does affect stability. However, the $\nabla_\theta \hat{\mathcal{L}}_\rho$ update in Algorithm 1 (Line 9) is analogous to that in standard contrastive divergence, and we inherit any practical implementation details from prior work [44]. • **Online setting**: The online setting is very different: there is no need to approximate state distributions as we do, since they can just sample from the true distribution directly. We have now included a brief note of this in the Discussion section. • **Stochastic policies**: Policies per Eq. 6 can readily capture stochasticity. We have now included additional experiment results in the appendices to measure calibration, which EDM generally *improves* over others: e.g. for `MIMIC-III-2a`, the expected calibration error is 2.32% less than BC. • **Discussion section**: Agreed; this is now given its own section.

[Reviewer 3] Thank you for your thoughtful comments. For further detail, we will upload code per official guidelines.

[Reviewer 4] • **Proof of Proposition 2**: Thank you for pointing out the need for clearer justification for the transition in Line 504. To be clear, the proposition and algorithm are both correct as intended. We agree, however, that additional detail would benefit exposition. While you are correct that gradients and expectations cannot be freely exchanged in the *most general* case, here we can exploit regularity assumptions. To justify the exchange, we have now included a version of the following (known) result as auxiliary lemma: Let $\theta \in \Theta$, r.v. $s \in \mathcal{S}$, and fix $f\colon \mathcal{S} \times \Theta \to \mathbb{R}$, where $f(s, \theta)$ is continuously differentiable w.r.t. θ and integrable for all θ. Assume for some r.v. X with finite mean that $|\frac{\partial}{\partial\theta} f(s,\theta)| \leq X$ a.s. for all θ. Then $\frac{\partial}{\partial\theta}\mathbb{E}[f(s,\theta)] = \lim_{\delta\to 0}\frac{1}{\delta}(\mathbb{E}[f(s,\theta+\delta)] - \mathbb{E}[f(s,\theta)]) = \lim_{\delta\to 0}\mathbb{E}[\frac{1}{\delta}(f(s,\theta+\delta) - f(s,\theta))] = \lim_{\delta\to 0}\mathbb{E}[\frac{\partial}{\partial\theta} f(s,\tau(\delta))] = \mathbb{E}[\lim_{\delta\to 0}\frac{\partial}{\partial\theta} f(s,\tau(\delta))] = \mathbb{E}[\frac{\partial}{\partial\theta} f(s,\theta)]$, where for equality 3 the mean value theorem guarantees existence of $\tau(\delta) \in (\theta, \theta+\delta)$ and equality 4 uses the dominated convergence theorem where $|\frac{\partial}{\partial\theta} f(s,\tau(\delta))| \leq X$ by assumption [Weir, 1973]. Generalizing to gradients simply requires the bound be on $\max_i |\frac{\partial}{\partial\theta_i} f(s,\theta)|$ for elements i of θ. To be clear, most machine learning models (and energy-based models) meet/assume these regularity conditions or similar variants; we have also now included a brief note about their reasonableness. • **Relationship with Algorithm 1**: While Proposition 2 sets the stage for the analysis in Section 4, the (gradient-based) implementation of Algorithm 1 is *also* correct due to a simpler reason: the batched (empirical loss) $\nabla_\theta \hat{\mathcal{L}}_\rho$ portion of the update (Line 9) is analogous to that in standard contrastive divergence. In other words, simply to show it works as intended, we could have stopped at equality 3 in Line 504 and be done. (Of course, that would have been at the expense of the simplicity of subsequent derivations for Section 4). • **Energy-based learning**: We agree more background on EBMs is beneficial. Since we rely on an existing technique in statistical physics, we agree it is more informative—and responsible—to properly introduce them from the get-go. We have now updated the manuscript to properly introduce EBMs after defining the objective, importantly invoking the joint EBM technique [44–46]—with particular credit to Eq. 7 in [44] for popularizing the "discriminative+generative" model that is our Eq. 10. This introduction now (appropriately) replaces the admittedly decorative (and potentially confusing) presentation of Lemma 1. Moreover, we have also streamlined the paper by removing any redundancies/overselling as suggested. • **Approximation of true occupancy**: You are absolutely correct that the *true* state occupancy distribution depends on MDP dynamics, and that the proposed formulation simply involves an *approximation* of this distribution. With the updated presentation of joint EBMs (see previous point), this should now be clear. We now explicitly emphasize it is impossible to obtain the former without actually executing policies online; we agree the (prior) oversold presentation may be misconstrued to suggest that our offline approach achieves the impossible. (We also include a note of this contrast/relation to online distribution-matching). • **Environments+comparisons**: Using common control environments follows recent work in offline IL, and like most (if not all) IL work we use the same NN architecture for all benchmarks for standard comparison [see e.g. VDICE, DSFN, GAIL]; we also go beyond [32–37] by adding more complex (`BeamRider`) + realistic (`MIMIC-III`) examples. (Note: MuJoCo is not applicable to categorical actions). We have also added experiments with Linear BC ("LBC"); see above figure for an example on `LunarLander`.