We thank the reviewers for their insightful & constructive feedback, to which we have carefully responded below.

- ***(Shared by R2, R4) Clarify the claim on counterfactual cross-validation (CV).*** Counterfactual CV (CF-CV) means counterfactual term $\tau(x)$ is directly used in the CV target (our Robinson residual). In contrast, "normal" CV in ITE would evaluate factual residual $\epsilon = Y_t - \mu_t(z)$, which does not effectively assess counterfactual performance. See [51, 67] for formal discussions & empirical evidence on how CF-CV avoids biased or unreliable estimation suffered by normal CV. Also, there are settings with additional information on more precise propensity $e(z)$ (*e.g.*, randomized trials, missing / pending outcomes) to improve Robinson residual based CF-CV. We have made these clear in our revision.

Reviewer #1.

- ***Clarify "more robustness to unmeasured confounding."*** We agree the statement needs to be more precise. We advocate the view that assuming unmeasurable confounders is a more natural, robust alternative relative to assuming no unmeasured confounders, known to be invalid for many empirical settings. (*e.g.*, in targeted advertising true confounder should be the latent consumer profile (*e.g.*, buying power, personal flavor, etc.), not observed past purchase records (*i.e.*, proxies)) Note endogenous uncertainties arise in data-generation, and as noted in [20], encoding proper latent uncertainties in the model is a valid, possibly less restrictive alternative to *ad hoc* sensitivity analysis [21]. With little scholarly consensus on the matter, we advise modeling with uncertainty unless domain knowledge suggests otherwise, in the same vein to the fixed effects models that are commonly used in econometrics [Angrist & Pischke (2009)].
- ***Comments on causal identification with proxies.*** These are excellent points. The community has very different takes on "identification" (see positive comments from R3), and improving upon Miao *et al.* remains an open challenge. We will expand discussion on causal identifiability, listing settings & examples where it might be feasible. Informally, for individual-level identifiability, we first need latent identifiability & verify latent effects can be averaged out.
- ***More ablations to show when noise helps or hurts.*** Note the noise in the proxy model $p(x|z)$ diminishes information & leads to uncertainties in causal estimates, while the noise in the inference model $q(z|x)$ helps prevent overfitting causal estimates until it starts to hurt training stability. More results and analyses will be added to clarify the interplay.

Reviewer #2.

- ***Do you need values of $\tau(z)$ for training?*** No, we do not. We believe the reviewer misunderstood the R-learning framework, and we wish to clarify. Here $\tau(z)$ is directly trained using factual data (x, t, y) using the Robinson factorization defined in Eqn (3). The nuisance components, namely the mean outcome $m(z)$ and propensity $e(z)$, are learned using factual pairs (x, y) & (x, t), respectively, with standard regression. We do not model μ_0, μ_1 in BV-NICE.
- ***Novelty of the work.*** Our key insight is that existing generative causal models failed to account for covariate balance & counterfactual validation, resulting in compromised performance, and we provide a new method accounting for that.
- ***Why use KL instead of the IPM metrics (e.g., Wasserstein, MMD).*** IPM metrics are not without caveats. Its primal-form estimation suffers quadratic scaling, and its dual-form estimation requires intricate constrained optimization. Our KL estimation is simple, linear scaling, numerically stable, and it yields very strong performance.

Reviewer #3. We thank the reviewer for the very positive comments.

- ***Backing up "BV-NICE variant alleviates algorithmic bias for minorities."*** We appreciate this suggestion. Informally, distribution shifts are more likely for under-represented populations, thereby less generalizable with standard learning schemes, but can be better handled with BV-NICE variants. We will extend the supplemental material with clear definitions & setups, and present some results. A more formal, dedicated presentation is being prepared separately.

Reviewer #4.

- ***Explain bad performance for some baselines, and additional results in the SM.*** As documented in our experimental setup, we fix the representation dimension to two for all representation-based models. Some baseline models (*e.g.*, CFR) turned out to be very sensitive to the representation dimension, yielding bad results in low-dimensions. Limited by space, we present only key results in the main text upon submission. We will use the extra page offered by the final version to provide a more comprehensive presentation and analysis of the additional experimental results.
- ***Relation to the mentioned works on regularizing causal models.*** Thank you for pointing out these interesting refs on counterfactual risk minimization (CRM), which we have carefully read and added to our discussion. We note that their motivations, target objectives, and estimation procedures are very different from our work: these CRM models regularize KL divergence on the policy (model) distributions to upper bound excess risk, not to promote representation balance as in BV-NICE. Nevertheless, together they present a more holistic picture of how Bayesian formulation and information-theoretic regularization help to improve counterfactual reasoning.
- ***Does the paper intend to identify latent confounders?*** We are not explicitly pursuing that goal here. Discussions on the additional assumptions for latent confounder identification (see supplemental material) are provided for completeness, and to bridge our future work on associating causal interpretations. The main take-away is that representation balancing, incorporating uncertainty & direct modeling of causal effects are important factors to consider for (black-box) generative causal modeling, which we have demonstrated with the success of BV-NICE, without enforcing latent identification.
- ***Other minor comments.*** We can generalize to symmetrized KL. VI variants can help. σ is hyper-parameter for noise & justification added. Z models hidden confounding (related to identifiability). Alg 1 clarified as suggested. See replies above on counterfactual CV & robustness to unmeasured confounding for how we address points (i) and (iii).

We also enriched our discussions in the broader impact statement, and fixed the typos and references.