We thank the reviewers for their detailed feedback. We're glad that the reviewers thought our general identification theorem is sound and interesting (R1, R2, R3), found our work is creative (R2), and see VDE as an elegant solution to satisfy the independence conditions required in theorem 1 (R3).

To address clarity concerns, we have adopted the reorganization suggestions of section 2 from R2. This should move the "identification comparison with other control function methods" closer to theorem 1 which clarifies one of R1's main concerns by connecting our work to Guo and Small 2016. This would also surface the error analysis of GCFN from section 2.3 and appendix A.7 that R1 requests and provide the more intuitive structure of section 2 that R3 requests.

[R1 + R3 : limited experimental evidence | R3, toy experiments] We do not know of any established benchmarks for causal effect estimation with IVs. We briefly discuss our evaluation here. Section 3.1 shows GCFN can handle assumptions that existing methods like CFN, 2SLS, and DeepIV cannot. Then, as demonstrated below and in section 3.3 in the paper, GCFN performs on par or better than DeepIV and DeepGMM on *high-dimensional* simulated data from each paper [1, 2] respectively. Thus, GCFN is competitive on data that satisfies the assumptions required by DeepIV and DeepGMM *while* being applicable to data generated with non-additive outcome processes [1, 2].

[Additional evaluation of GCFN] Due to reviewers' concern about toy experiments, we present further evaluation of GCFN on simulated data with a high-dimensional IV. We use the data generating process given in DeepGMM [2], a recent state of art method. We ran GCFN with 10 different random seeds and report results for $\kappa = 0.3$, chosen based on mean outcome MSE. We report results for DeepGMM and DeepIV as reported in [2]. GCFN performs competitively with an effect MSE of $\mathbf{0.077 \pm 0.022}$ compared to DeepGMM's $\mathbf{0.07 \pm 0.02}$ and DeepIV's $\mathbf{0.11 \pm 0.00}$. Effect MSE for $\kappa \in \{0.2, 0.4\}$ were similarly within standard errors of DeepGMM's performance and better than DeepIV [1].

[R1 + R3, hard to gauge assumptions | realism of assumptions] Solving VDE exactly guarantees reconstruction (A1) and marginal independence $\hat{\mathbf{z}} \perp \boldsymbol{\epsilon}$. A strong IV (A2), intuitively, is one that is able to set treatment to any value given any fixed confounder value. Joint independence requires assumptions on the treatment process like additivity. Additivity is common in the economics literature and strong IV is an assumption that can be reasoned about using domain expertise: for e.g. can college proximity influence a student's decision to go to college regardless of skill?

[R1, Theorem 1 and traditional identification results] To clarify, theorem 1 characterizes control functions that guarantee identification. We compare GCFN's identification for an additive treatment process to traditional control function identification from Guo and Small (reference 15 in the paper) in the paragraph "New conditions for effect identification" in section 2.2. Briefly, a strong IV requires more than the exclusion restriction and relevance properties of IVs. For this added assumption, GCFN drops the additive outcome process and the additional conditions on noise that CFN requires. We have expanded this discussion in the paper.

[R1, Section 2.2 is opaque, no analysis] As noted above, we have clarified our presentation in section 2.2 and A.7. We briefly summarize GCFN here: 1) GCFN's first stage, VDE, has an objective where at optimum, control functions meet condition A1 in theorem 1. Similarly, VDE's decoder leverages treatment process assumptions where at optimum joint independence (A3) is satisfied (line 187). 2) GCFN then runs a flexible outcome regression on the treatment and VDE's control function to estimate effects. As in section 2.3, we analyse GCFN's effect error in appendix A.7. In A.7.1, we show how the two components of VDE, reconstruction error and dependence of $\hat{\mathbf{z}}$ on $\boldsymbol{\epsilon}$, influence effect error in data with additive treatment processes. We also give a general bound in A.7.2.

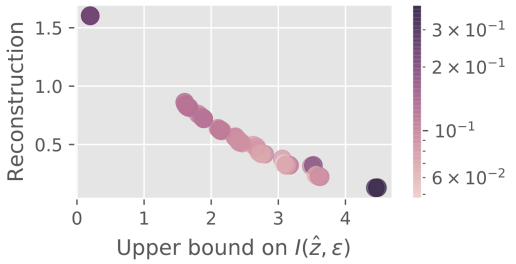

Figure 1: Effect estimation MSE of GCFN in the MNIST IV experiment above plotted for the corresponding VDE's reconstruction loss and upper bound on $I(\hat{\mathbf{z}}, \boldsymbol{\epsilon})$ (as in derivation in A.2); the lighter the color, the better the error. Effect MSE is good and very mildly sensitive to changes in reconstruction or information for values in $0.3 - 0.5$ and $2.5 - 3.5$ respectively. Outside those ranges, effect MSE is more sensitive and is bad when either quantity is large.

[R2, $\mathbf{I}(\hat{\mathbf{z}}, \boldsymbol{\epsilon}) > 0$, reconstruction error, GCFN failure cases] We thank the reviewer for this suggestion. We will add discussion about failures modes of GCFN. We plot the influence of non-zero information and reconstruction on effect MSE in fig. 1 using models trained on the MNIST IV data above for different κ.

[R3, Real experiment unconvincing] The reported effect in Nunn et al. (reference [34] in the paper) is a well-known result supported by modelling choices informed by domain knowledge. We believe that recovering this effect gives evidence that GCFN works well even without strong parametric assumptions used by Nunn et al.

[R3, Kernel IV, Multidimensional z] We thank the reviewer for the suggestion. We will include a Kernel IV baseline. If $\mathbf{t}$ is scalar but $\mathbf{z}$ is multidimensional, an additive treatment process with a scalar $f(\mathbf{z})$ would be $\mathbf{t} = g(\boldsymbol{\epsilon}) + f(\mathbf{z})$. In this case a scalar $\hat{\mathbf{z}}$ suffices to capture $f(\mathbf{z})$ and satisfy ignorability.

[1] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.

[2] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, 2019.