We sincerely thank the reviewers for their valuable feedback. We are glad to see that the reception of our paper has been mostly positive. R2 requested a conceptual discussion of the assumptions. R7 had clarification/improvement questions. We address these questions and the others below. All minor comments will be implemented in the camera-ready.

R2: *Philosophy of Low-Entropy Assumption.* There are two philosophical interpretations. First one is the statement that the true causal model has a small amount of randomness. This is the interpretation taken in [11]. Note that this requires a comparison between $H(X) + H(E)$ vs. $H(Y) + H(\tilde{E})$. Even though this incorporates $H(X)$ as you point out, it does not require $H(X)$ to be small, but the total entropy to be smaller. The second interpretation is that "the **additional** unobserved randomness is small", which can be seen as relaxing the determinism assumption. Our objective is to quantify how much we can relax this assumption and still retain identifiability. A different interpretation is that entropy of the model can be seen as a way to approximate its Kolmogorov complexity. Kolmogorov complexity of a causal model has been proposed by Janzing et al. in [8] as a way to identify the causal direction. We are going to point out this connection and hope to make this a more formal connection in the future. Thank you for pointing out to Janzing et al. We will discuss this counterexample in relation with our assumptions.

Nature has structure. We agree that if we know the structure, this could help. Assuming uniform distribution over the function space is our way of measuring how often the proposed method may fail, given that we do not know nature's structure. This assumption can be relaxed in specific ways, for example when there is one state y whose inverse image is largely supported, this is sufficient. We see this as an indication that uniform assumption is not necessary. We will flesh out these alternatives to uniform sampling of f in camera-ready.

Related work. We will add all the citations along with a detailed discussion on how they compare to our approach.

R3. *No confounder assumption.* We have provided experimental results to illustrate that the proposed method is robust to light confounding. Please see Section 4 lines 277-287.

What can be said about line graphs? Thank you for pointing this out. As long as exogenous entropy for each variable on the line graph is a constant and there are not too many such variables, our machinery can be applied. B and D in your example is a valid case. This is due to the fact that we can write $B \to D$ with exogenous entropy of at most $H(E_C) + H(E_D)$. We will explain this and other relevant settings in camera-ready.

Identifiability While used in multiple ways in literature, the term identifiability within the causal discovery settings typically refers to identifying the causal direction/graph. We will point to the related work that employs the same usage.

Bayesian interpretation. We agree that one can interpret the assumptions as being Bayesian. However, the posterior distributions for these quantities are very hard to compute. Therefore, we refrained from this terminology. Instead, we put a measure on the model space to be able to quantify what fraction of models can be identified using the proposed framework. If a prior on the function space is known, this can be incorporated in our proof to understand whether identifiability persists. Without any knowledge, we believe uniform prior is suitable.

Comments on quantization. We will try the ensemble approach and report in camera-ready. We chose to use the whole dataset rather than bootstrap, since the number of samples is small (100). Thank you for the other research directions.

R4: Thank you for your feedback. We will clarify that Renyi 0 case is resolved in [11].

R6: Please see Section O (line 820) of Appendix for the details about the synthetic data generation. We will move these details to the main text, making use of the extra page in camera-ready. Thank you.

R7: *Implications of the assumption* We provided several experimental results to assess the effect of our assumptions. In Section 4 lines $237 - 253$ we experimentally evaluate the implications of the low exogenous entropy assumption. If $H(E)$ is too small, this implies $H(X) > H(Y)$. If it is large enough, this implies $H(X) < H(Y)$. This creates three different regimes. We show in Fig. 2 that our method almost always identifies the causal direction in all three regimes.

What if more than one model can be fit to the data? We aim to find the model with the minimum exogenous entropy in both directions. Even if there are multiple such models, the minimum entropy will be unique. Also note that our bounds hold for *any E for which there exists an f s.t.* $Y = f(X, E)$, thus providing us a lower bound to the minimum entropy.

Comparing with existing methods. The performances of these methods on Tübingen are available in the literature (see [15] by Mooij et al.). We will include the accuracies reported by them. As far as we are aware, ANM provides around 64% accuracy. Information-geometric (IGCI) approach performs worse than ANMs. A nonlinear extension of LiNGAM gives $62 - 69\%$ as reported in Hyvarinen et al. "Pairwise Likelihood Ratios for Est. of Non-Gaussian SEMs". Also note that, except IGCI, all of these methods require ordinal variables, whereas we can handle categorical data.

We never know alphabet size is large. In practice, we fit the simplest model in both directions and pick the one with the smaller entropy, regardless of the alphabet size. In our experiments, we set a threshold $t < 1$ and make a decision only if $H(E) \leq t\log(n)$ or $H(\tilde{E}) \leq t\log(m)$. Please see Table 1. Thank you for your feedback.