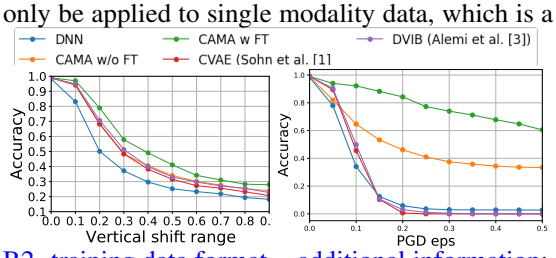
We thank the reviewers for their time and insightful comments on our paper. We respond to each reviewer(R) below. In addition, we will make the code publicly available, together with the paper.

R1. The proposed model is a form of conditional variational autoencoder (CVAE) : We believe this is a misunderstanding, and our contribution goes far beyond that. Our contributions are in three-fold: (1) a causal view on the robustness of neural networks and the discussion on valid artificial manipulations; (2) the CAMA model as an instance of causal consistent generative models, including versions for both single modality data and generic causal graphs; (3) fine-tuning approaches to improving model robustness to unseen manipulations. R1's comments cover only part of contribution (2), i.e., the CAMA model for single modality data. R3 also pointed out "The proposed fine-tuning phase to learn unseen M intervention is novel and sets this paper apart from previous work like Narayanaswamy et al." We discussed in detail how our work differs from CVAE-type of models in appendix B, and will clarify on this further in revision.

R1.decomposition necessity and CVAE comparisons: We emphasise that CVAE and other mentioned work can only be applied to single modality data, which is a special case discussed in our paper. For empirical comparisons, we present on the left of below figure the robustness of CVAE (Sohn et al. [1]) & DVIB (Alemi et al. [3]) to vertical shifts and against PGD attacks, which clearly shows the advantage of CAMA, especially the fine-tuning ability. Similar results have been reported in [25] of our paper's reference, and our work provides extra advantages due to the use of a causally consistent model and the fine-tuning method motivated by causal reasoning.



R2. training data format... additional information: The training procedure only requires knowledge on whether the training data is clean (in such case we set $m = 0$) or manipulated (potentially with unknown manipulation). In the manipulated case we do not require explicit label for M and perform inference on it instead. For test data, the labels for both Y and M are not available, only the input X is given, similar to the standard test setting.

R2. Comparisons to IRM: First, IRM only considers single modality data. Moreover, IRM aims to find representations that are invariant to environmental changes, which does not necessarily provide robustness to *unseen* manipulations. Imagine training a model with IRM using clean MNIST and MNIST shifted 50% to the left, then the model is very likely to fail when tested on MNIST shifted 50% to the right. Will clarify in revision.

R3. (adversarial) data augmentation: Deep CAMA also benefits from adversarial training (Figure 10). Both CAMA and discriminative model are robust to *known* adversarial attacks that are observed in adversarial training. However, the key advantage of CAMA over discriminative models is in its robustness to *unseen* manipulation. First, Figure 1 shows that training discriminative DNNs with one manipulation can hurt its robustness to another unseen manipulation. This is not the case for CAMA (see the overlapping green & orange curves in Figure 7(b)). Second, with fine-tuning (which is another main contribution of our work), CAMA can be significantly more robust to unseen manipulations.

R3. concerns on line 52 statement: In causality literature, causal factorizations correspond to modular/independent mechanisms [34,38] of our paper's reference; building ML models consistent with these causal factorizations also improves robustness [19] of our paper's reference. Will clarify further by expanding the arguments.

R3. Bayes' rule prediction clarity: As shown on the RHS of eq. (6), we sample m from $q(m|x)$ and sample z from $q(z|x, y, m)$ with the previously sampled m. In the paper, we use 1 sample for m and K samples of z associated with each m sample. Given the sampled m and z instances, we can compute the terms inside the log for each $y = c$ (as an approximation to $p(x, y = c)$), and apply softmax to obtain the predictive distribution. Will clarify further on it.

R3. parameter size: The network size detail is presented in appendix D for all experimental settings. In appendix C (Figure 16), we present the performance of discriminative models with larger network sizes and the result is similar to the one with smaller network sizes.

R4. computation cost: The time complexity of CAMA is in the same order of a regular VAE. For predictions, fine-tuning requires a small amount of additional time, as only a small fraction of data is needed for fine-tuning (Figures 8 & 20).

R4. "Limitations of Adversarial Robustness": We do not intend to claim CAMA's robustness to all possible manipulations *straightaway after training*. So the theory of the mentioned work (and many others) is applicable to CAMA before fine-tuning. However, this line of theoretical work evaluates the robustness of a classifier trained on clean data only. Rather, CAMA brings extra advantages by fine-tuning, which enables adaptation to the unknown manipulation in test time.This adaptation is required whenever a new manipulation is present, allowing CAMA to "learn" to be robust to a growing number of manipulations.

R4. Network architecture bias: Indeed RCNN-like networks can be more robust to shifts by design. However, Fig. 15 & 16 shows that the use of CNN (instead of MLP) does not fully address the over-fitting issue to seen manipulations. Moreover, architecture design typically provides robustness to specific manipulations. Instead, our goal is to make the model robust to (infinity number of) unseen manipulations, which cannot be archived by architecture design only.

R4. limit mis-specification: In practice, the causal graphs are either provided by domain experts or obtained by running causal discovery algorithms. For the former, it requires working closely with domain experts to refine the causal hypotheses. For the latter, choosing a suitable causal discovery algorithm for the application at hand would be critical, and there exist approaches to evaluating the performance of causal discovery algorithms.