We thank reviewers for their valuable feedback.

General Response: ► **(R1/R3) Contributions and Difference from Related Work.** Our first contribution is presenting PD estimation methods that avoid optimizing MI (neural) variational bounds. The probabilistic-classification method defines a binary-cross-entropy loss to differentiate samples from joint distribution or samples from the product of marginal distributions. GAN instead distinguishes between samples from true data distribution or samples from generator distribution. The density-ratio fitting method aims at estimating density-ratio between the joint distribution and the product of marginal distributions. Prior work studied kernel-based methods, while we take advantage of high-capacity neural networks.

Our second contribution is leveraging PD into 1) MI estimation, 2) self-supervised learning, and 3) cross-modal learning. For MI estimation, we find plugging-in estimated PD results in lower bias/variance than optimizing from MI's lower bound. We find the loss inspired by the density-ratio fitting method consistently outperforms the SOTA baseline for self-supervised learning. For cross-modal learning, we showcase the usage of PD for cross-modal retrieval (and cross-modal adversarial samples debugging provided in Supplementary) task(s).

► **(R1/R3) Advantages over MI Variational Bounds Methods.** We discussed the reason why we prefer the presented approach over the MI variational bounds methods in lines 125-131. It might seem that our idea is straightforward in retrospect, but we argue that its effectiveness in reducing the variance in MI variational bounds is important in many real-world applications. For example, prior work *[Song et al., Understanding the Limitations of Variational Mutual Information Estimators, 2019]* has pointed out that these MI variational bounds methods often have a large variance, and the large variance leads to the numerical issues in practice. On the contrary, our presented method does not estimate mutual information directly. We estimate the point-wise mutual information. Our approach either 1) utilizes the binary cross-entropy loss that has the benefit of numerical stability form the recent optimization package (e.g., PyTorch or TensorFlow) or 2) contains no logarithm or exponentiation, which is the cause of the numerical instability. In the final version of the paper, we will include more motivations for our presented approaches in Section 3.2.

► **(R1/R4) MI Estimation.** The comparisons with the SMILE method can be better understood by providing quantitative comparisons. For the quantitative analysis based on the bias-variance trade-off (e.g., Bias^2 + Variance), the Probalistic Classifier has the best performance, and the SMILE method is runner-up. The detailed quantitative numbers will be provided in the final version of the paper.

We would like to emphasize the that main takeaway from Figure 1 is an overall trend: estimating mutual information directly from its lower bound has a larger variance (with SMILE as an exception) compared to approximating mutual information by plugging-in the estimated PD. SMILE achieves superior performance because it clips the model's outputs to prevent abrupt large or small numbers. The remaining approaches in Figure 1 do not post-process the model's outputs. We will also include this discussion in the final version of the paper.

► **(R2/R4) Remark on Crossmodal Retrieval.** For this experiment, our purpose is to showcase the usage of PD estimation. Note that we have performed analysis using Density-Ratio Fitting method (92.26% top-1 retrieval accuracy) in line 280. In the future work, we will elaborate more on the usage of PD estimation for cross-modal retrieval and compare it with more baselines. Besides this experiment, in Supplementary, we also study cross-modal adversarial samples debugging using PD estimation.

Reviewer #1: ► **Connection between Section 3.1 and 3.2.** We thank R1 for suggesting 1) better motivating why we need the proposed method in Section 3.2 and 2) absorbing and abstracting some part of Section 3.1, which is not the main contribution of the paper. To address the concerns, we will slightly shorten Section 3.1 and expand Section 3.2 in the final version of the paper.

Reviewer #2: ► **Figure 1 and 2.** Figure 1 presents the results for MI estimation. Figure 2 shows the results for the self-supervised representation learning. Prior work *[Tschannen et al., On Mutual Information Maximization for Representation Learning, 2019.]* suggests a higher MI does not result in a better representation, which shares a similar observation as ours. A better understanding of the relationship between good MI estimation and a good representation learning objective is still an open research problem.

Reviewer #3: ► **Datasets are Toy Data.** For experiment 1, correlated Gaussians have a closed-form mutual information expression, and hence it is viewed as the benchmark experiment for mutual information estimation *[Belghazi et al., MINE: Mutual Information Neural Estimation, 2018].* For experiment 2, we are happy to provide experiment on ImageNet. We use ResNet-50 as the backbone model, where we obtain the test accuracy 74.0% when using Density-Ratio Fitting method. The baseline is contrastive predictive coding, which gives us test accuracy 73.70%. We will include this result in the final version of the paper. For experiment 3, our purpose is to showcase the usage of PD estimation. It is our future plan to elaborate more on the usage of PD estimation for cross-modal retrieval and compare it with more baselines.

Reviewer #4: ► **Similar Pairs in MNIST/ CIFAR10.** For an input image, we perform two different data augmentations on this image, viewing these two augmented variants as a similar pair.