We thank all the reviewers for their valuable comments, efforts, and time. In particular, we sincerely appreciate that all reviewers agree on novelty/originality of our method, which we believe a useful contribution to the literature on conditional GANs. Below are our responses to the reviewers, which we will incorporate in the final manuscript.

For Reviewer #1

Justification on the gap of log-densities. We emphasize that our goal is to measure the point-wise (or sample-wise) discrepancy, while the relative entropy (that you suggested) measures the discrepancy under the entire domain, *i.e.*, an expectation of the point-wise discrepancies.[1] To this end, as we did, it is natural to use the difference of the densities for the point-wise discrepancy. In particular, we suggest to measure the difference of *log*-densities because it leads to a computationally efficient estimator (as shown in our paper). We will clarify these in the final manuscript.

Properties of the gap of log-densities. The gap of log-densities is *not* a metric in a mathematical sense, *i.e.*, it is neither a function $X \times X \to [0, \infty)$ of a domain X, nor satisfies the three conditions you mentioned. Instead, it is a simple scalar value to measure the point-wise discrepancy. which is enough for our purpose. We will revise our manuscript to avoid such confusions in using the term 'metric'.

Why equation (6)? By using the generator's distribution p_g as a proposal distribution and p_{data}/Mp_g as an acceptance rate, it is easy to prove that the sample distribution theoretically converges to the target distribution p_{data}. As the GOLD estimator approximates $\log(p_{\text{data}}/p_g)$, the equation (6) is a natural choice for the acceptance rate.

For Reviewer #3

Projection discriminator. Our definition of type (a) includes the projection discriminator, as we cite the paper in line 61. As R3 mentioned, it decomposes the marginal and conditional terms in their architecture, which would result in another estimator form of the gap of log-densities. We will add a related discussion to the final draft.

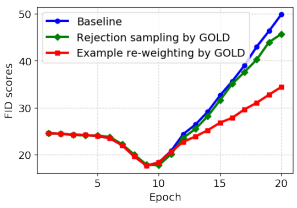
Metropolis-Hastings GAN (MH-GAN). As MH-GAN requires the density ratio information p_{data}/p_g to run, one can indeed apply the GOLD estimator to it. We will add a related discussion in the final manuscript.

Other comments. We will revise our manuscript by clarifying all of the following points.

- Line 62. "c)" refers to the codomain of the discriminator D, not a method group.
- Line 129-131. The re-weighted loss encourages the generator to strengthen under-generated regions while regularizing over-generates regions.
- Equation (7). As we do not know the true class c_x of data x, we estimate it by $c_x \sim D_C(c|x)$. Taking an expectation over the class probability leads to the entropy formula (7).
- Line 180. We choose different architectures for different types of inputs since it is more adequate to test a larger network for a more complex task. To this end, we follow the same choices in the related work, *e.g.*, InfoGAN is used for the MNIST dataset (1-channel images) whereas ACGAN is for the CIFAR dataset (3-channel images).
- Line 245. The discriminator is often overfitting and thus re-initializing parameters can help to find better local minima.
- Line 255. In column 2, the GOLD value is the highest for the leftmost region, which is uncovered by the generator. In column 3, the GOLD value is the highest for the upmost region, where samples are not obtained, *i.e.*, uncertain.

For Reviewer #4

FID scores and unstable training. Following R4's suggestion, we will add FID scores and corresponding discussions in the final manuscript. Here we show some of them. (a) When applied to rejection sampling, our method consistently improves the FID score as shown in the table below. (b) In order to address R4's question on the case when training is highly unstable, we also measure FID scores during training with only 10 labeled samples (and no additional unlabeled samples) from the MNIST dataset. In this case, we observe that mode collapsing occurs at around 10-th epoch, and we apply our method for the next 10 epochs. The figure on the right shows that our method can mitigate the instability issue, significantly improving the FID score during training.



	MNIST	FMNIST	SVHN	CIFAR-10	STL-10	LSUN
Baseline	10.78±0.04	12.38±0.03	8.28±0.07	9.46±0.04	14.47±0.04	14.38±0.03
GOLD	**10.70**±0.05	**12.32**±0.06	**8.12**±0.06	**9.44**±0.02	**14.44**±0.04	**14.35**±0.06

Realistic generation? Both our re-weighting scheme and the original one do not force the generator to "fool" a classifier, *i.e.*, the generator and discriminator are adversarially trained only to force to generate realistic samples and do not compete to produce the right class. The re-weighting scheme targets both class and reality of samples to improve.

GAN loss (equation (1)). We use the non-saturating GAN loss (proposed in the original paper of GAN by Goodfellow et al.) to improve the stability in training. We will clarify this in the final manuscript to be consistent with our code.

ImageNet experiments. Following R4's suggestion, we will add larger-scale experiments to the final manuscript.

[1] For R1's interest, we remark that one can obtain the KL divergence $D_{\text{KL}}(p_{\text{data}}\|p_g)$ and the reverse KL $-D_{\text{KL}}(p_g\|p_{\text{data}})$ by taking expectation of the gap of log-densities under the real and generated distributions, respectively.