We thank all of the reviewers for their valuable feedback and detailed comments. According to the reviewers' suggestions, we want to clarify the main idea of the paper. That is "improvement and justification of any implicit sampler". We know that in practice, even state-of-the-art generative models yield "unrealistic" samples, hence, are biased. From the MCMC perspective, we could treat these (already learned) models as proposals for the approximate MH-algorithm (Algorithm 3). Based on our theoretical analysis, we derive different losses for the discriminator (Table 1 in the paper).

R1: "irreducibility and aperiodicity does not imply the existence of a stationary distribution ..."

Thanks for pointing out this mistake! Indeed, we also need the minorization condition on the transition kernel.

R1: "the TV metric does not seem like a good metric"

We assume that the proposed algorithm could be applied not only for images. Hence, we provide analysis in the most general setting and use the TV-distance as a standard metric in MCMC analysis (Roberts, 2004, general MCMC).

R1: "assumption in Section 3.1 are terribly restrictive"

Although the minorization condition for the proposal distribution is indeed restricting, it automatically holds for an independent proposal (as we note on lines 134-135), which is the most common scenario for GANs. Moreover, if the support of the target distribution is a compact and the density of the proposal is continuous and positive on this compact, then we can lower bound the proposal density by a positive constant, hence, satisfy the minorization condition (as we note on lines 195-197). We can define a distribution of images on the support $[0,1]^d$ by adding a low-variance Gaussian noise (truncated to $[0,1]^d$) to the observations, thus defining positive target and proposal distributions on the compact.

R1: "empirically test the resulting methodology on tractable high-dimensional toy problems"

As you suggested, we provide additional experiments for high-dimensional tractable toy problem. As in (Roberts, 2001, optimal scaling), for the target, we take factorized distribution $p(x) = p(x_1)\prod_{i=2}^{d} p(x_i)$, where $p(x_1)$ is the mixture of two Gaussians and the rest $d-1$ components are standard normal $p(x_i) = \mathcal{N}(0,1)$. For the Markov proposal, we take homogeneous random-walk $q(x|y) = \mathcal{N}(x|y,\sigma I)$ and scale σ with dimensions as proposed in (Roberts, 2001, optimal scaling) to keep the acceptance rate about 20%. For the independent proposal, we take homogeneous Gaussian $q(x) = \mathcal{N}(0,\sigma I), \sigma = 1.2$. Empirical results are in Fig. 1.

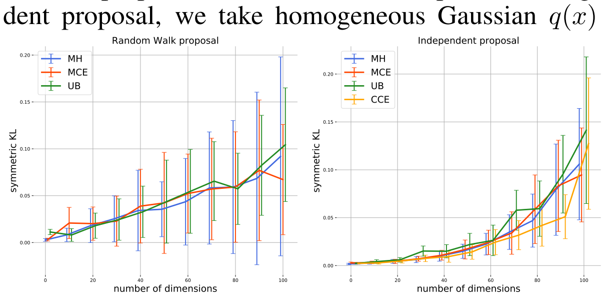

Figure 1: We evaluate the symmetric KL along the first dimension (as the most difficult) for a chain of length 20000, averaging across 100 independent runs. We compare the performance of our algorithm for different losses with the exact MH algorithm (MH on the plots).

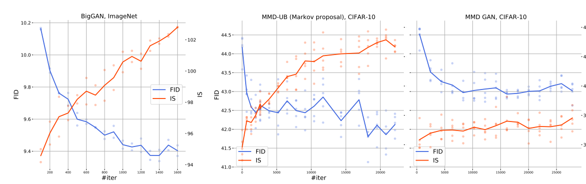

Figure 2: Monotonous improvements in terms of FID and IS for BigGAN and MMD-GAN (both for Markov and independent proposals). We learn a discriminator for the Markov proposal by optimization of the Upper Bound. Performance for the original model (baseline) corresponds to 0-th iteration of a discriminator.

R2: "Baselines for comparisons to all the cases are also needed"

The baselines for our algorithm are the initial implicit models that we improve. In Fig. 1 of the paper, the performance of the initial model corresponds to the very beginnings of the plots (0-th iteration of learning a discriminator for the MH-test). Further, we extend the algorithm to the case of Markov proposal. In Fig. 2 of the paper, we demonstrate that MH with a Markov proposal not only improves the initial models/baselines (the first points on the plots) but also improves over the independent MH with the same generator network.

R2: "comparisons with state-of-the-art models are needed"

Note that we use PyTorch implementation of the InceptionV3 network; hence, the metrics could be different from the TensorFlow implementation. For instance, the IS for our WGAN varies as 3.6 (PyTorch), 4.7 (TF). As you suggested, we provide additional experiments for the state-of-the-art models (see Fig. 2).

R3: "In section 4.1, are the models pretrained with their original objective?"

Yes, all the models are already trained with their original objective, and we filter them by running Algorithm 3.

R3: "generate “correct” samples from VAE with AIS. Would these samples have a 100% accept rate?"

To perform the AIS, one needs the densities of target and proposal. In the case of VAE, we can estimate the density of the proposal as $q(x) = \mathbb{E}_{p(z)}\text{decoder}(x|z)$, but we still need the density of the target. If we use the same discriminator for its estimation in AIS, then yes, we will obtain approximately 100% acceptance rate.

R3: "the experiment setup could be explained in more detail. Do you train an encoder for the VAE?"

Yes, we train the VAE as in the original paper, then we use only the decoder as a generator by sampling the latent variables from the prior. We will clarify the experimental setup in the final version.