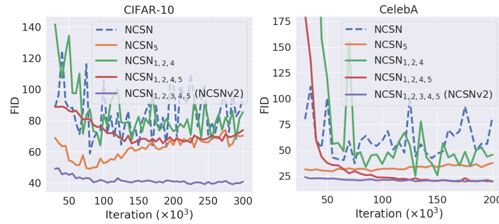
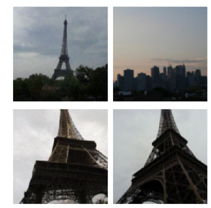
We thank all the reviewers for providing valuable feedback in this time of stress. Below we first discuss a new discovery on evaluation metrics, and then answer specific questions of the reviewers.



A surprising fact of FID computation. Because of annealed Langevin dynamics, the samples contain small Gaussian noise that is imperceptible to human eyes. After our paper submission, we discovered that this small Gaussian noise—though hard to detect by humans—can greatly hurt the FID scores. Therefore, we **denoise the samples by running one step of $\mathbf{x} \leftarrow \mathbf{x} + \sigma_L^2 \mathbf{s}_\theta(\mathbf{x}, \sigma_L)$ and compute the FID scores again.** We provide the new ablation results in the right figure, with the extra configuration $\text{NCSN}_{1,2,4}$ suggested by **R3**. We follow the same checkpoint selection method in Table 5 and provide full FID scores below. **The NCSNv2 model now obtains much lower FID scores than NCSN**, which aligns better with our visual inspection of samples. We are surprised by how the FID scores improve for both NCSN and NCSNv2 though **samples before and after this additional denoising step are the same to naked eyes**. We will include these new results in the revision.

	NCSN (CIFAR-10)	NCSNv2 (CIFAR-10)	NCSN (CelebA)	NCSNv2 (CelebA)
FID	27.44	**10.31**	17.57	**9.69**



[R1] Is the model memorizing data (like the Eiffel towers in Figure 1)? In the paper, we argued from several perspectives that the model is not memorizing data: (i) The test loss and training loss are comparable to each other (see Figure 12); (ii) Nearest neighbors in the training dataset do not look the same as samples from the model (see Section C.4.2); and (iii) The model can generate samples that smoothly interpolate from one to another (see Figure 7 and Section C.4.3). In the right figure, we additionally provide nearest neighbors (the right column) in ℓ_2 distance to the two Eiffel towers (the left column) which appeared in Figure 1.

[R1][R2] Whether EMA has a negative impact on performance? As **R1** and **R2** noted, EMA stabilizes training but sometimes may have a slightly worse peak FID score. Because a larger variance gives rise to larger extreme values, unstable methods naturally lead to a better peak FID score. However, we believe this is an imperfection of the peak FID metric, rather than an indicator that unstable methods perform better. In fact, as shown in Figure 4 and 11, EMA yields lower FIDs most of the time and samples with EMA look much more visually appealing than those without EMA.

[R2] Are all techniques needed for scaling to higher resolution? From the new ablation results above, we observe that using all techniques leads to the best performance. We agree that for a specific dataset like CelebA it may not be necessary to use all 5 techniques to get reasonable results, but one key point of our paper is that using all techniques make the model **work out of the box** for **a large number of different datasets**, which we demonstrate on many datasets of different resolutions, including 32^2, 64^2, 96^2, 128^2 and 256^2.

[R2] When does the RefineNet architecture change? We hope to clarify one confusion: in the ablation study, only NCSNv2 uses the new architecture and the others use the old one. The new architecture is necessary for using Technique 3 because it assumes an unconditional score network. We can view the impact of this architecture change by comparing $\text{NCSN}_{1,2,4,5}$ and $\text{NCSN}_{1,2,3,4,5}$ (*i.e.*, NCSNv2) in the ablation results.

[R2] Writing issues. Thanks for pointing them out! We will incorporate your suggestions in the revision.

[R2][R3] Evaluation metrics. There are many known issues with existing metrics of sample quality, and finding the right one is still an open problem. We choose FID and HYPE_∞ as an approximation to the real sample quality. The discrepancy of FIDs in Figure 5 and Table 5 is because FIDs in Table 5 **are the peak FIDs and are computed on more samples**. The HYPE_∞ scores are computed on more than 2000 **uncurated** samples, and are better when closer to 50. "Fakes Error" is the proportion of fake images perceived as real, and "Reals Error" is the opposite.

[R3] Oversimplified assumptions. Despite using simplified assumptions, our theory predicts parameters that perform very well across a large number of complicated real datasets. It is proved by our experiments to be useful and valuable.

[R3] KDE and Technique 1, 2, 4. When applied to multi-scale KDE as in the setting of Figure 2, annealed Langevin dynamics will converge to samples that **exist in the training dataset**. Training an NCSN makes it possible to generalize to novel samples. We provide the ablation study of Technique 1, 2, 4 in the above figure (see $\text{NCSN}_{1,2,4}$), showing that they can improve FID over NCSN even without EMA (Technique 5).

Dataset	Device	Sampling time	Training time
CIFAR-10	2x V100	2 min	22 h
CelebA	4x V100	7 min	29 h
Church	8x V100	17 min	52 h
Bedroom	8x V100	19 min	52 h
Tower	8x V100	19 min	52 h
FFHQ	8x V100	50 min	41 h

[R2][R4] How long the model trains on what hardware, and the sampling speed. We provide the statistics in the above table, and will add it to the paper in the next revision. The sampling time is for one mini-batch.