SUPPLEMENTARY APPENDIX FOR *EBM Life Cycle: MCMC Strategies for Synthesis, Defense, and Density Modeling*

Anonymous authors
Paper under double-blind review

This supplementary appendix presents additional experiments to support our responses to reviewers. These experiments will be incorporated into future revisions of our paper.

A UPDATED DEFENSE RESULTS

Since the time of our original submission, we have verified to the best of our abilities significantly stronger CIFAR-10 defense results than reported in our original paper. The framework remains completely consistent with our original results, and the only difference comes from a more effective natural classifier and better selection of EBM checkpoints. We were aware at the time of submission that our method could produce such results, but a last minute uncertainty with our implementation caused us to report a more conservative result that had our full confidence. Since the time of submission, we have re-verified our best CIFAR-10 defense and report our updated defense results in Table 1. Our new results show that the EBM defense can surpass the defense for SOTA adversarial training.

Table 1: Defense vs. whitebox attacks with l_∞ perturbation $\varepsilon = 8/255$ for CIFAR-10.

Defense	$f(x)$ Train Ims.	$T(x)$ Method	Attack	Nat.	Adv.
Ours	Natural	Langevin	BPDA+EOT	0.8664	**0.6760**
(Hill et al., 2021)	Natural	Langevin	BPDA+EOT	0.8412	0.5490
(Song et al., 2018)	Natural	Gibbs Update	BPDA	0.95	0.09
(Srinivasan et al., 2019)	Natural	Langevin	PGD	–	0.0048
(Yang et al., 2019)	Transformed	Mask + Recon.	BPDA+EOT	0.94	0.15
(Carmon et al., 2019)	Adversarial	–	PGD	0.897	0.625
(Zhang et al., 2019)	Adversarial	–	PGD	0.849	0.5643
(Shafahi et al., 2019)	Adversarial	–	PGD	0.859	0.4633
(Madry et al., 2018)	Adversarial	–	PGD	0.873	0.458

B UPDATED LONGRUN RESULTS

We have updated our longrun sampling experiments with new results on CIFAR-10 that exhibit significantly more stable trajectories. The original CIFAR-10 results were trained using a midrun EBM as the prior distribution rather than a shortrun EBM as used for the Celeb-A and ImageNet experiments. We trained the CIFAR-10 model using a shortrun EBM as the prior distribution and got much better results. We believe this is because the shortrun prior EBM will oversaturate quickly so that the new EBM can immediately learn to patch the defects of the prior EBM. This will be discussed in more detail in future paper versions. The new result for CIFAR-10 longrun sampling is shown in Figure 1.

We also computed FID scores for 5000 samples at both 100K steps (as in the original submission) and at 1 million steps. Due to the computational cost of longrun sampling, we were unable to use a large sample set and the scores are significantly higher (worse) than they would be with a full FID evaluation of 50K samples. The FID score remains relatively stable between 100K and 1 million steps, which is evidence that the samples have approximately reached the steady-state. This is consistent with the expected behavior of a well-formed probability density. Our CIFAR-10 results are from

100K Steps 1M steps

Figure 1: Longrun samples from our new CIFAR-10 longrun model. The new results have significantly better appearance at the extremely long trajectory of 1 million steps. Image realism is quite consistent from 100K to 1M steps, indicating that the samples have approximately converged to the steady-state.

Table 2: FID for 5K samples after 100K Langevin and 1M Langevin steps.

		FID	
Data	Resolution	100K Steps	1M Steps
CIFAR-10	32×32	49.2	51.7
Celeb-A	64×64	37.4	45.9
ImageNet	64×64	82.3	77.8

a new model, while the Celeb-A and ImageNet result use the models from the original paper. The results are reported in Table 2. Our new CIFAR-10 results are slightly worse at 100K steps but much more stable at 1M steps.

C Longrun Samples of Normalizing Flow and Diffusion Models

To underscore our claims about the difficulty of calibrating the probability mass of a density model, we investigate longrun MCMC samples from a normalizing flow and diffusion model. We observed in Appendix G of our original submission that the density implied by the gradients of a score model have a misaligned steady-state, similar to a misaligned shortrun EBM. A recent work shows that the problem of steady-state misalignment has gone mostly unnoticed even for the Restricted Boltzmann Machines (Decelle et al., 2021). We further find that the normalizing flow from the GLOW model (Kingma & Dhariwal, 2018) and the recovery likelihood diffusion model (Gao et al., 2020) have misaligned steady-states as well. This shows that the problem of improper density estimation extends well beyond the EBM. Tractable density modeling with a normalizing flow does not prevent steady-state misalignment. These experiments corroborate our claim that log likelihood experiments in previous works are not able to detect the misaligned probability mass of many prior models. We believe that the calibration of the model steady-state is currently best diagnosed with longrun MCMC sampling because the distribution of longrun MCMC samples represents the probability mass of the model.

Figure 2 (*left*) displays initial and final states from a GLOW model density after 100K sampling steps. Despite the fact that the GLOW model has a fully tractable density, it is unable to learn a valid distribution of probability mass. Figure 2 (*right*) shows initial and final samples from the Recovery Likelihood T6 model after 100K steps of the conditional model at the lowest noise value. We observe the same oversaturation for the conditional density as for the unconditional density of

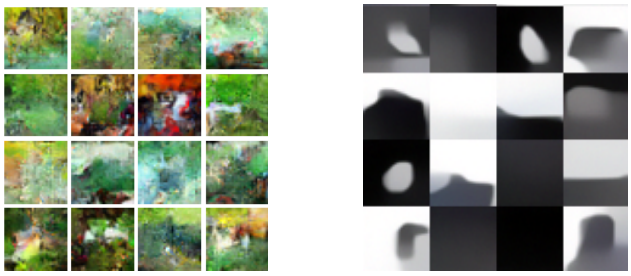

Figure 2: *Left:* MCMC samples after 100K steps using a GLOW model (Kingma & Dhariwal, 2018) trained on CIFAR-10. *Right:* MCMC samples after 100K steps using a conditional recovery likelihood model (Gao et al., 2020) trained on CIFAR-10. MCMC samples were initialized from data samples. Neither model can correctly approximate the distribution of probability mass for the data density. The problem of steady-state misalignment extends beyond EBMs to many other generative density models. We tried several different temperatures close to 1 for the GLOW model and found equivalent results.

a standard EBM. Code for the T1K model that the authors evaluate in their longrun experiments is not released so we have not yet been able to directly test their results. The T1K model is equivalent to an EBM version of the score model that we test in Appendix G which we have shown has a misaligned steady-state. Further, we note the longrun experiments with the T1K model are very misleading because the experiments use 100 steps with 1000 distinct conditional models and claim this is a longrun evaluation of 100K steps. The correct evaluation is to use 100K steps on a single conditional model. We strongly believe that the Recovery Likelihood model as originally presented has a misaligned steady-state like many other methods. We hope that the observations in our work can lead to efforts to stabilize the sampling trajectories of many existing models.

D Shortrun Synthesis without Langevin Gradient Clipping

Beyond what is discussed in the initial submission, we use gradient clipping on Langevin chain gradients and network update gradients for shortrun experiments only. No gradient clipping is used for midrun or longrun experiments. While we find that gradient clipping for network update gradients improve learning stability, we find that removing the Langevin gradient clipping has negligible effect. Nearly equivalent results can be obtained with no Langevin gradient clipping, as shown in Table 3. In the original implementation, the Langevin gradient clipping was set to a high value so that it rarely interfered with the dynamics. In future revisions we will report the scores without Langevin gradient clipping to remove this unnecessary hyperparameter.

Table 3: Comparison of FID Scores for Shortrun Synthesis for learning with and without Langevin gradient clipping.

		FID	
Dataset	Resolution	Lang. Clip	No Lang. Clip
CIFAR-10	32×32	22.9	22.12
Celeb-A	64×64	15.3	16.3
ImageNet	128×128	40.6	38.9

E Importance of Learning Rate Annealing

This section demonstrates the importance of learning rate annealing for learning a robust energy landscape. We repeat the midrun and longrun learning experiments for CIFAR-10 except that we never anneal the learning rate. We then sample with the models for 1500 steps for the model trained with the midrun method and 100K steps for the model trained with the longrun method. The results

in Figure 3 show that learning rate annealing is essential for stabilizing both midrun and longrun trajectories.

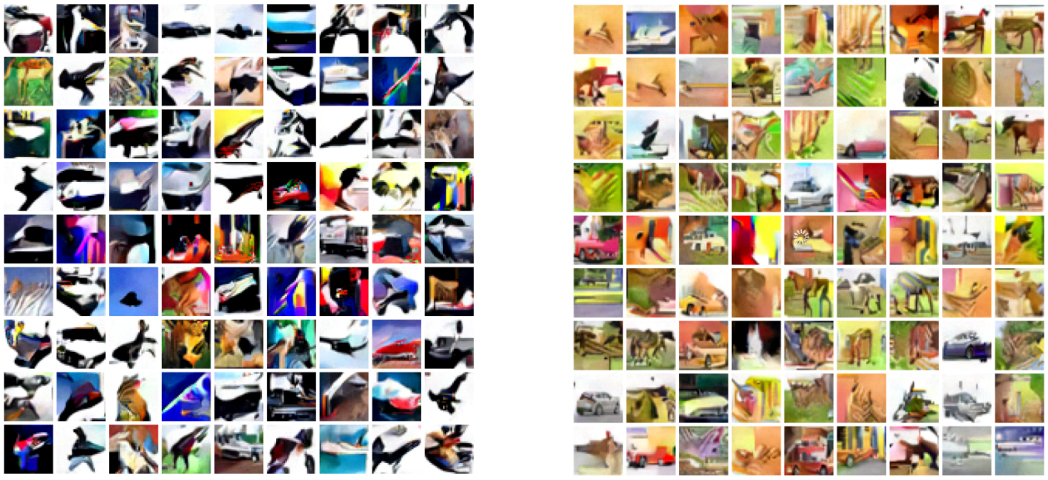

Figure 3: Ablation study showing the importance of annealing. *Left:* Samples from a non-annealed model trained with the midrun method after 1500 MCMC steps. *Right:* Samples from a non-annealed model trained with the longrun method ater 100K MCMC steps. MCMC samples were initialized from data. This shows that rejuvenation of the midrun trajectories from data and the separation of longrun samples into burn-in and update banks alone is not enough. Annealing ensures that samples from past EBMs function as approximate samples from the current EBM, since the weights are changing very slowly.

The importance of annealing can be understood as follows. If the EBM is being updated with a very low learning rate, then samples from recent EBM snapshots can function as samples from the current EBM. In the case of midrun trajectory, annealing allows the model to robustify trajectories that are approximately as long as the lifetime of a persistent sample between rejuvenation. In the case of longrun learning, annealing allows the burnin samples to approximately reach the model steady-state before they are included in the update bank. This allows the persistent samples in the update bank to function as approximate steady-state samples from the current EBM, leading to proper modeling of probability mass.

REFERENCES

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aurelien Decelle, Cyril Furtlehner, and Beatriz Seoane. Equilibrium and non-equilibrium regimes in the learning of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, 2021.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125*, 2020.

Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=gwFTuzxJW0`.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free!, 2019.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

Vignesh Srinivasan, Arturo Marban, Klaus-Robert Muller, Wojciech Samek, and Shinichi Nakajima. Defense against adversarial attacks by langevin dynamics. *arxiv preprint arXiv:1805.12017*, 2019.

Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7025–7034, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7472–7482, 2019.