==========Reply to R#1==========

Thanks for pointing out our paper is significantly novel with competitive performance. We address your concerns below.

Why not SOTA: *We would like to clarify on the originality and goal of our contribution. In this paper, we want to show the benefits of using adaptive stepsize for learning a ConvNet-based EBM where the energy landscape is highly nonconvex, not only via experiments but with a rigorous non-asymptotic theoretical analysis. As our method aims at accelerating the convergence of the model in the first iterations, we argue that our paper does not aim at proving an additional SOTA in terms of generated outputs. Rather, we tackle this problem from an optimization point of view with the motivation of improving how the latent samples find rapidly a good enough maxima in the target conditional distribution. Our numerical experiments and theorem do provide insights on that regards as the first epochs show.*

Anistropic step size choice: *The simplicity of our stepsize comes in the fact that it is based on the gradient of the distribution at each iteration. The only tuning parameter is the threshold which can be found empirically.*

Benefits of STANLEY over existing methods?: *See first comment. We mostly focus on the convergence of the sampling scheme in the first epochs. This contribution is focussing on the transition regime and do not tackle the asymptotic convergence nor the final accuracy of the model.*

Compare with NUTS: *A5: NUTS sampler is based on HMC sampler. We refer the reviewer to check our results for HMC in the numerical experiments.*

Typos: *A6: We will correct the typos. Thanks.*

==========Reply to R#3==========

Assumptions of Allassonniere and Kuhn and EBM: *Our main contribution theory-wise is to extend their proof to the case of EBM, i.e. in the nonconvex case. Hence their assumptions do not suffice as they are stronger that our case.*

Clarity of Figure 1: *We will improve the resolution of Figure 6 for the sake of clarity.*

How robust are the curves to different model initializations?: *This is an interesting point that we believe all papers in that realm should tackle in the future. Most of our runs are single runs and studying the robustness to the initialization by averaging multiple runs could be interesting.*

Could STANLEY be leading only at the beginning in Figure 5? Behavior at convergence?: *As we tackle early convergence speed, we do not focus on the heavy tail of the convergence and certainly, our contribution does not reside in the resulting model accuracy but more on the ability to improve the ability of the MCMC to obtain good samples in the first epochs, i.e. when the EBM model parameters are still far from the target parameters.*

Why use FID rather than PSNR to evaluate image inpainting?: *Jianwen can you reply to that please?*

Using the same image in Figure 6: *we will follow suggestion in our revision.*

Relation between subscript k and subscript t? Random initial states?: *The two subscripts are independent. k monitors the MCMC chain and t monitors the EBM training. The initialization of the chain is random at each new parameter hence at each t.*

What is the drift function V_θ? What does χ stand for? Is there supposed to be $\pi_{theta}(\cdot)$ instead of $\pi_{theta}u(\cdot)$ in corollary 1?: *We define the drift function in the proof in the appendix. We will fix the typo raised by the reviewer. Thank you.*

Geometric ergodicity seems to follow the proof Allassonniere and Kuhn: *Our main contribution theory-wise is to extend their proof to the case of EBM, i.e. in the nonconvex case. Hence their assumptions do not suffice as they are stronger that our case.*

==========Reply to R#4==========

The authors do not do a good job of distinguishing their own contribution from prior art: *We will clarify that in the revision. Drastically reducing the number of MCMC transitions would have a great impact on the energy consumption and speed of the whole training process. Besides, we stress on the important theoretical contribution that is presented along our algorithm compared to prior art. To the best of our knowledge, EBM methods are presented mainly using empirical insights on their respective contribution. In this paper, we wanted to show the benefits of using adaptive stepsize for learning a convent-based EBM where the energy landscape is highly nonconvex, not only via experiments but with a rigorous non-asymptotic convergence analysis.*

ULA in the algorithm and MALA in the proof: *For the sake of the proofs we use the MALA algorithm. From a convergence standpoint, there is no interest in considering the ULA method.*

Dependence of the results on the anisotropic learning rate. Relation with existing literature, specifically Atchade (2006)/Roberts and Tweedie (1996): *Our main contribution theory-wise is to extend the proof of Allassonniere and Kuhn to the case of EBM, i.e. in the nonconvex case. Hence our results are intrinsically related to our introduced scheme. As for Roberts and Tweedie, their works is seminal and has paved the way to a long series of theoretical MCMC papers using their proof techniques. Indeed ours is also built upon those tools introduced in their work.*

Equation 4 - dependence the max norm? per chain?
The stepsize is anisotropic and depends on the max norm. Since it also depends on the norm of the gradient, this stepsize depends on the initialization of the chain and the iteration index k.

Section 5.1 - Langevin and STANLEY learning rates
The learning rates for Vanilla Langevin is the classical constant stepsize used in the Langevin Dynamics and is fine tuned over a grid. Coupling the quantitative FID curves and the qualitative synthetic images is important to be able