We sincerely thank the five reviewers for their valuable feedback. We list below our replies to your concerns.

Reviewer 1

*** Benefit of Anisotropic stepsize:** We develop STANLey in order to more efficiently sample from the Gibbs potential. H ence, our goal is not to reach an optimal and high resolution generated image, but rather to decrease the number of kernel transitions need at each EBM iteration in order to obtain relatively good samples. Drastically reducing this number 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. 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. This also echoes with R4's remark on "how the approach fits within the broader landscape of energy-based modeling approaches". Indeed, we specifically design STANLey update to take into account the curvature of the nonconvex energy landscape by embedding a dimension gradient informed stepsize.

The goal is to propose an efficient algorithm with low computational cost and with theoretical guaranties of relevance. These are particularly hard tasks when dealing with high dimensional data and nonconvex models such as CNNs. Hence, we propose an optimized Langevin update and prove that this new sampler leads to a geometrically uniformly ergodic Markov chain.

*** Visual Checks:** We will improve the resolution of Figure 6 for the sake of clarity.

Reviewer 2

*** Performance of HMC:** The slightly better performance of HMC in terms of FID can naturally be explained by its Hamiltonian dynamics that exploit first and second order moments of the target distribution to explore the sampling space. HMC has been for a long time the state of the art method for sampling high dimensional posterior distribution. Yet, the second-order information computation brings heavy computational burden exhibited Table 1. For that reason, reaching similar performances using our cheaper method is a good sign.

Reviewer 3

*** Quantitative results:** We would like to point out that Figure 2 presents a visual comparison between our method and the vanilla Langevin, Figure 3 displays how STANLEY and three other baselines compare when generating synthetic natural images and Figure 6 corresponds to the visual comparison of STANLEY and the vanilla Langevin on the celeb-A dataset. Those visual plots aim at getting a visual perspective of the benefits of our method. We add that Figure 4 and 5 compare all baselines in terms of FID which is also a visual comparison of the generated images.

*** Complexity of STANLEY:** Running times are now reported on Table 1. Our method is relatively comparable to the vanilla method in terms of computation complexity since line 3 of Algorithm 1 in the main paper uses the already computed gradient vector (no added computation unlike HMC method). In terms of memory, we shift from a constant scalar learning in the vanilla Langevin to a vector of stepsizes depending on the dimension index d. No memory issue were to be reported during our extensive experiments.

Reviewer 5

*** FID plots:** We would like to invite the reviewer to reassess this remark on the FID curves. Indeed an FID equal to 0 corresponds to two sets of exactly identical images, which would be absurd when dealing with generated synthetic images. As a result, plotting FID=0 is out of context here.

*** Complexity Analysis:** We provide the running times of our method and the baselines in Table 1 on CIFAR-10 and Celeb-A datasets with a batchsize of 100. We would like to stress on the similar computational complexity between the vanilla Langevin and our method STANLEY since our newly introduced stepsize uses the already computed gradient vector. On the contrary, the HMC method has recourse to both the gradient and the Hessian of the target distribution, resulting in longer computation time as reported on Table 1.

Table 1: Runtime (in s) for training our EBM during 1 epoch.

	Vanilla Langevin	HMC	GD	**STANLEY**
CIFAR-10 Dataset	232.5	698.4	211.3	265.2
Celeb-A dataset	376.3	640.1	345.2	414.8

We run each of the method, including ours, on a single TitanXx8 GPU for our experiments.

In terms of memory complexity, we acknowledge that STANLEY requires to store a gradient-informed stepsize larger than a constant one for the vanilla Langevin. Though, in none of our runs we encountered any memory issue. The benfits of using a gradient-informed in terms of convergence outweighs its memory constraint.

*** Reproducibility:** As the reviewer as evaluated the reproducibility of our work as fair, we would like to highlight that the proofs of our results can be found in the supplementary material, the data is open source (CIFAR, Flowers and celeb-A) and the code can be requested.

*** Conclusion:** Except the FID curves concern, which was out of context, we addressed the main concern of Reviewer 5. For that reason we would like the reviewer to consider increasing its score on our contribution.

Reviewer 6

*** Originality of our contribution:** See Reviewer 1 reply.

*** Comparison with GANs and Flow models:**

Comparing to flow models and GANs is out of the scope of our contribution. Indeed, we introduce a novel MCMC sampling method for the sole purpose of training an energy based model. Since EBM are based on the learning of an energy function through negative sampling (obtaining samples from the conditional distribution is the main challenge), we do not believe that including discriminator-based architecture is irrelevant.