We would like to thank all three reviewers for acknowledging our contributions and providing valuable feedback. Please find our responses to your comments below.

Reviewer #1:

Thank you for the positive comments on the novelty of our idea and insightful questions for further improvement.

We first characterize the solutions of DEAN. Let p_x and $p_{x'}$ be the distributions of real and fake data; p_e denotes the energy-based distribution. In DEAN, p_e is a bridge connecting p_x and $p_{x'}$. Now we provide two theorems for the characterization. For the IGN, the network is trained to have $p_{x'}$ equal to p_e. Please refer to Theorem 1, which is proved based on Theorem 1 of [JXS+17]. For the EGN, p_e is learned to estimate p_x. Please see Theorem 2, which is proved according to Theorem 1 of [ZML17] and Theorem 1 of [GPAM+14]. At present, Theorem 2 is proved with $\Lambda(\theta_e)$. Other choices for the energy objective will be left to future works. Detailed proofs of the following theorems will be given in the Supplement of the final version. Different from GANs, which are implicit generative models (IGMs), DEAN can explicitly estimate the underlying distribution of the real data after estimating θ_e and θ_g.

Theorem 1 *We assume that $\mathcal{D}_{x'}$ is drawn from $p_{x'}$. If the following conditions are satisfied: κ is a universal and analytic kernel; $\mathbf{E}_{a\sim p_{x'}}\mathbf{E}_{b\sim p_e}\left[s^{\mathrm{T}}(a)s(b)\kappa(a,b) + s^{\mathrm{T}}(b)\nabla_a\kappa(a,b) + s^{\mathrm{T}}(a)\nabla_b\kappa(a,b) + \sum_{i=1}^{d}\frac{\partial^2\kappa(a,b)}{\partial a_i \partial b_i}\right] < \infty$ with $s(a) = \nabla_a \log p_e(a)$; $\mathbf{E}_{a\sim p_{x'}}\|\nabla_a \log p_e(a) - \nabla_a \log p_{x'}(a)\|^2 < \infty$; $\lim_{\|a\|\to\infty} p_e(a)g(a) = 0$, where $g(\cdot)$ is given in Eq. (2) in Section 4.2; for any $J \geq 1$, almost surely $\mathrm{FSSD}[p_e, \mathcal{D}_{x'}] = 0$ if and only if $p_{x'} = p_e$.*

Theorem 2 *Let $\Lambda(\theta_e) = \mathcal{E}(x;\theta_e) + \left[\gamma - \mathcal{E}\left(G(z;\theta_g^*);\theta_e\right)\right]^+$ (please refer to Eq. (1) in Section 4.1 for details). The minimum of $\Lambda(\theta_e)$ is achieved if and only if $p_e = p_x$. With the optimized θ_e^*, $\int_{x,z}\Lambda(\theta_e^*)p_x(x)p_z(z)\mathrm{d}x\mathrm{d}z = \gamma$.*

Following your suggestion, we compare the powers (successful rejection rates) of MMD, linear-time MMD [GBR+12] and FSSD on toy problems, where MMD is a two-sample test statistic and FSSD is used for the GOF test. We adopt the distributions Gaussian $p(x) = \mathcal{N}(x|0, \mathbf{I}_d)$ and Laplacian $q(x) = \prod_{i=1}^{d}\mathrm{Laplace}(x_i|0, 1/\sqrt{2})$ for $d = 1, 3$, in which the parameters are set to make p and q have the same mean and variance so that the difference between p and q is subtle. F-SSD shows a higher power to discriminate the subtle difference (Figure 1). For larger sample sizes, the power of MMD is close to that of FSSD. However, in the GAN-type training, the batch size is usually less than 512. As the adversarial training continues, the distribution of the generated data gets closer to the energy-based distribution, and hence the difference becomes subtle. At this time, the power (discriminability) of FSSD for the subtle difference becomes important for generating high-quality images. Hopefully we have cleared up your main concerns with these theoretical and experimental discussions. We believe that the DEAN paradigm is promising, being versatile to yield specific training algorithms for different architectures of deep networks in different domains.

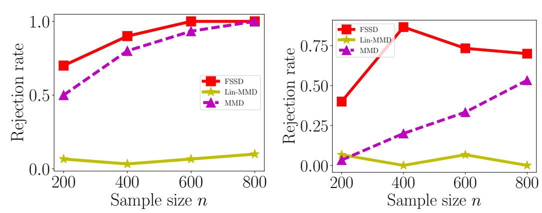

Figure 1: Rejection rates for $d = 1$ (left) and $d = 3$.

Reviewer #2:

Thank you very much for the encouraging comments and valuable suggestions.

Following your recommendation, we will add more discussions in the experimental part to provide takeaways and insights about DEAN. We adopted RBM as the energy function at the initial stage. However, the performance of DEAN with RBM is not comparable to that with autoencoder, so we discarded the results. We will add clarity on this in the final manuscript.

Reviewer #3:

Thank you very much for the positive comments and reasonable doubt.

In recent years, there are two emerging families for generative model learning, generative adversarial networks (GANs) and autoencoders (AEs) or variational AEs (VAEs), which are two distinct paradigms and have both received extensive studies. Goodness-of-fit (GOF) tests are a fundamental tool in statistical analysis, dating back to the Kolmogorov test in 1933. Our manuscript and [PDB18] both introduce GOF tests into deep generative modeling, but fall into different paradigms: [PDB18] is an AE-based method without adversarial learning while our paper is a GAN-type approach. The HTAE (hypothesis testing AE) in [PDB18] minimized the reconstruction error, but no adversarial learning (min-max adversarial optimization) was involved. The statistic in our manuscript is a kernel-based *nonparametric* GOF statistic. The Shapiro-Wilk test in [PDB18] is a traditional *parametric* GOF statistic for testing normality. Our paper is quite different from [PDB18]. The proposed DEAN with two generators is a pioneering work in the adversarial learning setting. Following your comment, we will cite [PDB18] in the final version.

References

[PDB18] Aaron Palmer, Dipak Dey, and Jinbo Bi. Reforming generative autoencoders via goodness-of-fit hypothesis testing. In *UAI*, pages 1009–1019, 2018.