We thank the reviewers for the positive feedback and valuable comments. Our response follows.

1) Computing smooth Wasserstein distance (SWD) [R3]: This is a great and important point. We will add a discussion about computation to the revision. As R3 mentions, the placeholder MC algorithm we used, in theory, has an exponential in d runtime (though no major issues arose in practice). Our submission studies statistical aspects of SWD, but we are actively working on efficient algorithms that compute it. To avoid sampling the kernel, we developed a method to compute $\mathsf{W}_1^{(\sigma)}(P_n, Q_m) = \sup_{f \in \mathsf{Lip}_1} P_n(f * \varphi_\sigma) - Q_m(f * \varphi_\sigma)$ under NN parameterization of f, with the high-dimensional convolution $f * \varphi_\sigma$ implemented in closed form. Accuracy and complexity of the alg. are being explored.

2) Choice of SD [R3,R4]: We will revise the presentation to thoroughly motivate SWD for inference. Briefly, SWD inherits much of its structure from classic W_1 [27], while enjoying faster empirical convergence. [27] shows: (i) SWD metrizes the same topology as W_1; (ii) it Γ-converges to W_1 as $\sigma \to 0$ (with convergence of optimizers); (iii) it is stable in σ, $|\mathsf{W}_1^{(\sigma)}(P, Q) - \mathsf{W}_1^{(\tau)}(P, Q)| \leq 2\sqrt{d|\sigma^2 - \tau^2|}$; (iv) $\mathsf{W}_1^{(\sigma)}$ is continuous and nonincreasing in σ. These suggest that SWD preserves the compatibility of W_1 for inference, while having superior statistical properties. Re generalization, Supp. A.1 states concentration inequalities for SWD, which imply generalization bounds for SMDE in the sense of [14] (lines 248-255). SWD achieves fast $n^{-1/2}$ generalization, as opposed to the $n^{-1/d}$ generalization gap of W_1.

3) Effect/choice of σ [R2,R3,R4]: Item 2 lists properties of $\mathsf{W}_1^{(\sigma)}$ as a function of σ. From a statistical standpoint, the expected empirical SWD scales as $\sigma^{-d/2} n^{-1/2}$ (see [26] and Thm. 1 herein). We are currently working on relaxing the prefactor under the manifold hypothesis, i.e., showing that it depends only on the manifold's (rather than the ambient) dimension. R4 is correct that characterizing the dependence on σ in limit distribution results is nontrivial, and we will comment on that in the revision, as suggested. In practice, having an excessively large σ may result in generative models that are slow to learn small-scale structures in the target distribution (though in theory they should be recovered eventually). That said, SWD is remarkably stable over moderate changes in σ and is thus not hard to tune.

4) Sliced W_1 (SW_1) vs. $\mathsf{W}_1^{(\sigma)}$ [R4]: We will gladly expand the SW_1 discussion as follows. SW_1 and SWD are different approaches for alleviating the curse of dimensionality (CoD) of W_1. SW_1 uses 1D projections, while SWD levels out local irregularities in the distributions by convolving with $\mathcal{N}_\sigma$. SW_1 and SWD share many similarities, e.g., both are metrics and (topologically) equivalent to W_1. SW_1 is easily computed using the 1D formula, while computational aspects of SWD are still in the works (Item 1). SWD might be preferable when comparing to regular W_1, as it is within an additive $2\sigma\sqrt{d}$ gap from W_1 [27] (see Item 2). Comparison results for SW_1 seem weaker, assuming compact support and involving implicit dimension-dependent constants (cf., e.g., Lem. 5.1.4. in `http://cvgmt.sns.it/paper/2341/`).

5) Entropic OT (EOT) vs. $\mathsf{W}_1^{(\sigma)}$ [R3]: Generally, we do not view SWD and EOT as competing techniques. It is beneficial for the community to have several methods for dealing with the CoD, especially since tradeoffs often emerge. We agree that a thorough literature review of EOT is appropriate, and will cover the following in the revision: (i) EOT is not a metric, while SWD retains the metric structure of W_1; (ii) $n^{-1/2}$ rate for EOT is proven only for smooth costs (excluding W_1) with compactly supported distributions (arXiv:1810.02733) or squared cost with subgassian distributions (arXiv:1905.11882), while for SWD the $n^{-1/2}$ rate holds under mild polynomial moment conditions; (iii) EOT CLT (arXiv:1905.11882) is similar to (arXiv:1705.01299) but markedly different from ours. Notably, they derive the result for two-sample populations with the unknown centering constants $\mathbb{E}[S_\epsilon(P_n, Q)]$ or $\mathbb{E}[S_\epsilon(P_n, Q_m)]$, which differ from $S_\epsilon(P, Q)$. (iv) A major virtue of EOT lies in fast algs., and matching those for SWD is a central goal going forward.

6) $\mathsf{W}_1^{(\sigma)}$ vs. MMD and $\sigma \to \infty$ limit [R2,R3]: MMDs do not suffer from the CoD, due to low complexity of RKHS function class. A main motivator for studying SWD is to alleviate the Wasserstein CoD, while preserving the metric structure. In light of Item 2, we are thus interested in $\sigma < \infty$. Also, [Cor. 2.4, arXiv:2005.00738] shows that $\lim_{\sigma \to \infty} \mathsf{W}_1^{(\sigma)}(P, Q) = |\mathbb{E}_P[X] - \mathbb{E}_Q[Y]|$, which is not informative as a discrepancy measure between distributions.

7) Weed-Berthet [R3]: Their focus is different from ours: they study density estimation under W_p and do not deal with limit distributions or MDE. Re computation, please see Item 1. Still, we are glad to cite this paper in the revision.

8) Applying analysis to EOT [R3]: Our proof technique to derive the asymptotic distribution relies on expressing SWD *exactly* as the supremum of an empirical process indexed by smoothed Lip_1 functions. As EOT cannot be written in this manner (for some popper function class), our proof technique does not directly extend to EOT.

9) Fig. 1 [R2]: The purpose of the point clouds is to show that the point clouds closely overlap. This implies that indeed a limiting distribution is emerging. We will add additional description to make this much more clear in the revision.

10) GAN method [R3], parametric experiments [R4], and appeal to ML audience [R6]: When referring to GAN-type methods, does R3 mean the generator's parameterization or the adversarial training? The former is covered by the MDE theory by viewing the weights/biases of the generator NN as θ (adversarial training aspects are indeed beyond the scope of this work). Our last experiment in fact implements a SWD GAN: while the data comes from a parametric distribution (GMM with $\exp(d)$ modes) we match it with a (generator) NN transformation of a latent variable. Per R6's suggestion, we will revise the presentation of this experiment to better communicate it to ML audiences. In general we will put more emphasis on NN-based ML applications, in addition to our focus on parametric setups.