We thank the reviewers for their feedback and time. We are happy to see that they found our article clear, well-written, original, innovative, and technically sound. The detailed responses are given below.

R1: We thank the reviewer for their valuable comments. We believe that we have now addressed all the raised issues. We hope this would help the reviewer to reconsider their score.

"... results are better for the linear distance, making me wonder whether the improvement ... might be an artifact..." We suspect that there has been a simple misunderstanding, which we first would like to clarify. As opposed to the claim, Fig 2 actually shows that linear SW is *always* outperformed by GSW with *polynomial* projections ('Poly 3' and 'Poly 5' in the legend) and also by all the variants of max-GSW (dashed lines). Besides, the figure also shows that *max-linear-SW* is consistently outperformed by *max-GSW-NN* as well. Then, if we consider Fig 3, we observe a clear continuation of this behavior, where max-GSW-NN outperforms max-SW on a real-data experiment.

Therefore, we would like to underline that the performance gain in Fig 3 is not due to choosing the best model setting or parameter tweaking. Besides, we would like to indicate that, we have provided our code (to obtain Fig 2) in the supplementary material, and the same code can be easily adapted to reproduce Fig 3, by simply loading MNIST in place of generating synthetic data. We invite the reviewer to use our code to verify the reproducibility.

Again regarding Fig 2, we would also like to clarify that, the only variant of GSW that is outperformed by linear SW is the GSW with neural network-based defining function (which is different from *max*-GSW-NN). This variant is unsurprisingly not performing well due to its inherent complexity of approximating the integral over a very large domain (Eq. 11) with a simple Monte Carlo average. On the contrary, max-GSW circumvents this issue by replacing sampling with optimization (note that for the same reason, [35] also preferred optimization over sampling). We agree that we could have emphasized more this point in the paper and we will add a short discussion to be more clear about it.

Experiments. Up on the reviewer's suggestion, we have contrasted our experiments with the ones of the three mentioned papers. All these papers first consider an experiment on synthetic data and then apply their approaches on real data. While we have the exact same structure in our experiments, we noticed that the main difference between our experiments and theirs is the application on a larger dataset. We have now conducted additional experiments on the CelebA dataset (Fig R1). The results are inline with our existing results: our approach finds a better solution in less number of iterations and the quality of the generated images is slightly better. We will add these new results to the paper. On the other hand, we have comparisons to SW and max-SW, which have already been compared to other baseline methods in [28,34]. Therefore, with the inclusion of these new results, we believe that our empirical validation is as comprehensive as the ones of the mentioned papers.

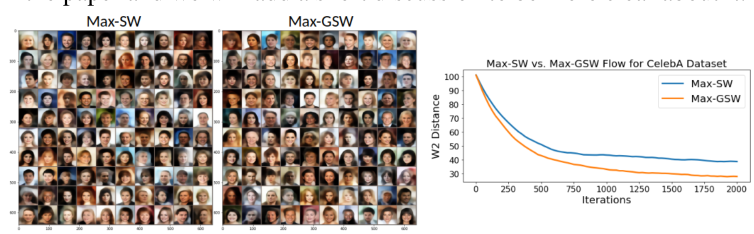

Fig. R1. Results on CelebA

Computational/statistical complexities. We will add a paragraph about the computational requirements as requested. Regarding the statistical complexity, we can consider the sample complexity of GSW. This has been established very recently for SW and max-SW only for Gaussian measures [34], and proving analogous results for GSW is out of our scope. However, to gain intuition about the sample complexity of GSW, we computed max-GSW and max-SW for varying number of samples drawn from two different Swiss Roll distributions (Fig R2). The distances exhibit the same behavior: we conjecture that GSW and SW have similar sample complexities. We will discuss this point in the supp. document.

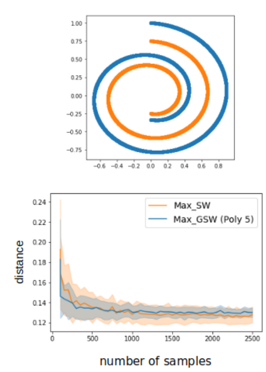

Fig. R2. Sample complexity

Preliminaries. We indeed deliberately allocated some space to introduce preliminary notions that might not be familiar to the ML community: the Radon transform and its generalization, which are the key constituents of GSW, are mostly used in tomography. Nevertheless, we will shorten Sec.2 as suggested and move some of the experiments of the supp. document as well as the ones presented here. In the end, our article will contain results on *several synthetic datasets, MNIST and CelebA, using GSW-flows and GSW-auto-encoders*. We believe this meets the NeurIPS standards.

R2: We thank the reviewer for the positive evaluation. We have actually provided the code to reproduce Fig 2: see the supplementary material of the submission. The code for the other experiments is very similar and will also be publicly released.

R3: We are grateful to the reviewer for their highly positive and encouraging feedback.