To Reviewer 1:

Clarity and miscellaneous comments. i) Your understanding of Alg 1 is correct. We will add more explanations on this for clarity.

ii) \( \mu = \{u_i\} \) represents the “empirical probability vector” over the support of \( \{x_i\} \). They are set to be uniform as shown in the Sinkhorn and IPOT algorithms in supplementary file, in the absence of prior knowledge. We have used this representation to make the presentation general enough to incorporate prior knowledge when available. We will make this clear and mention the Sinkhorn and IPOT sooner in revision.

iii) Label smoothing parameter 0.5 is set without hyper-parameter tuning. The investigation on smoothing parameter was done afterwards. We didn’t aim to achieve the best performance by hyper-parameter search but instead to show the general applicability of our approach under a broad range of hyper-parameter settings.

iv) The gap between PGD and the CW-variant is smaller under stronger attacks (§2) and we attribute the remaining gap to the nature of our model, where a one step unsupervised adversary is used for training, different from the multi-step supervised adversary typically used in Madry model.

v) “Emphasizing 1 attack iteration earlier” is a great suggestion. We will update this to avoid confusions as happened to Reviewer 2.

Disentangling of distance and coupling. Thanks for the great suggestion. We have preliminarily investigated the disentangling of distance and inter-sample coupling in our main paper as you have already noticed in Sec. 5.2 using the identity matching.

Further investigation on it (esp. the coupling) as suggested by the reviewer is interesting and we plan to work on it as our next steps.

To Reviewer 2:

Computational concern. The number of iterations \( T = 1 \) is used for our model as mentioned in line 225. We apologize for the confusion and will make it more clear as also suggested by Reviewer 1. Given that \( T \) is typically set to 7 in conventional PGD adversarial training (e.g. Madry), our approach does not take advantage of extra computation compared to conventional PGD training.

Random targeted baseline. We have experimented with random-targeted adversarial training as suggested by the reviewer. It achieves accuracy of 49.9/48.5 under PGD100/CW100 (min over 5 runs), outperformed by our model with a large margin (§2).

Understanding of feature-scattering. Conventional adversarial examples are decision boundary oriented (Fig. 2), making the effective manifold for training deviate from the original due to tilting and shrinking, hindering performance (line 56-52), with label leaking as one manifesting phenomenon. Feature scattering is inter-sample structure oriented and promotes data diversity without drastically altering the structure of manifold. We plan to conduct rigorous theoretical analysis of the proposed model as next step.

To Reviewer 3:

Batch size. As shown in the table on the right, larger batch size leads to better performance as it facilitates feature matching. Batch size of 60 is used for our model in the paper. Batch sizes larger than 60 lead to similar results. This observation is similar to other applications with embedded OT matching such as OT-GAN [48].

Label smoothing. Label smoothing is necessary for our model. Our model (with 1-step adversary) achieves compromised results without it, compared to standard PGD adversarial training (e.g. Madry) with 7-steps adversaries. This is an expected result as feature scattering makes the feature distributions more diffused (see Fig 1 in supplementary file), thus the corresponding label should ideally be “diffused” as well, in a spirit similar to mixup [70], which is achieved with label smoothing approximately in this work. Better schemes for joint treatment of feature and label scattering is an interesting topic and is left as our future work.

Choice of distance. Using cosine distance avoids introducing additional tuning parameter as the features are normalized before computing the distance. This and the usage of logits are just design choices. Other distance measures and intermediate features can be used together with our framework as well. We will explain this in the updated paper. As suggested by the reviewer, we will also introduce label leaking earlier in the introduction for clarity. We will release our trained model together with code as suggested.