We thank all reviewers for the insightful feedback. Below we address all questions raised in the reviews. We will add the related discussions and further experiment results in the new version, shall our paper be accepted.

[Reviewer 1] • **Intuition.** More intuition can be added in Section 3. For example, the set $\mathcal{H}(\mu)$ mentioned by the reviewer is a class of test functions for causal transports. Intuitively: causality is a concept of conditional independence (y_t independent of $x_{>t}$, given $x_{\leq t}$), that can be expressed in terms of conditional expectations, which in turn naturally leads to a formulation that involves martingales. • **Stability.** We can add more detail regarding stability of training COT-GAN. Empirically, we do observe that COT-GAN is relatively stable, in the sense that the loss tends to converge and that small changes in the hyperparameters do not obviously worsen the results especially in the lower dimensional settings. For high dimensional datasets, we indeed observe that sample quality fluctuates during training, which is a shortcoming suffered by all GANs. • **Figure 4.** We can show fewer frames.

[Reviewer 2] • **Novelty.** While most approaches rely on improving model architecture, compositional losses or domain-specific techniques, COT-GAN is a principled way of targeting generic sequential generation. Importantly, *causal* optimal transport was not present in the Machine Learning literature, and the proposed method is definitely not a mere marriage of two existing theories (COT and WGAN). First, it was not obvious to see that COT, formulated as a min-max optimization problem, naturally falls into the GAN framework. And even after this bridge was created, the development of the algorithm required substantial effort. Given our positive results, we believe the new theory of COT could greatly benefit sequential learning. • **Justification for mixed-Sinkhorn.** Our choice of mixed-Sinkhorn is inspired by the idea of taking into account the variation within the distributions μ and ν, as a way to correct the bias originating by mini-batched training. To support our intuition, we provide two arguments in Appendix A.3: the triangular inequality and the convergence to an unbiased estimator. Furthermore, this is confirmed empirically via the experiments in the paper as well as the additional results in the Figure below.

[Reviewer 3] • **Discussion on mixed-Sinkhorn.** We are happy to move some discussion to the main body of the paper. For the justification, please see our response to Reviewer 2. • **Comparison to other baselines.** We can add more details in the paper. In the Figure below, we provide an extra comparison between COT-GAN and TimeGAN, WaveGAN (trained with WGAN-GP loss) and COT-GAN without the mixing trick. Combined with the results in the paper, we see that the mixing trick is critical for the success of training and that COT-GAN achieves the best results among all.

[Reviewer 4] • **Weak Experiments.** We respectfully disagree with the reviewer on this comment. We thoroughly demonstrated the results for low and high dimensional applications using a variety of well-established performance metrics. Related work mostly focuses on either low or high dimensional datasets but not both, and often lacks reports on basic statistical features of the generated samples. For example, we achieved good results on EEG data without any domain-specific modifications, which outperform similar work specifically targeting EEG. As for the efficacy of our method on "rich-information datasets", we do not have reason to believe (either in theory or empirically) that our method would not be successful given sufficient computational resources. • **Comparison to other baselines.** Please see our response to Reviewer 3. • **Intermediate experiments.** It is unclear to us what is meant by "intermediate" experiments. If the reviewer is referring to experiments which investigate the specific contribution of each component of the model, we can include that in a later version of the paper. For example, isolating the impact of the martingale penalization p_M can be achieved by modifying the value of λ in (3.10), see Figure 8. We have run additional experiments omitting p_M and the mixing trick for the AR dataset, see the Figure below. • **Connection of the improved OT loss with COT.** In our model, the class of loss functions (parametrized in (3.9)) over which D optimizes is the one emerging from (3.4), which is the representation of the (regularized) COT problem as optimization over (regularized) classical OT problems. Therefore, D is de facto calculating the *causal* distance between the original data and the generated one. For concerns about differentiation from previous attempts, please see our response regarding novelty to Reviewer 2.