We thank all reviewers for insightful comments. *All added experiments are tabulated in revised manuscript/appendices*.

[Reviewer 1] • **Saito et al.**: We will rename our framework to TimeGAN to minimize confusion with [Saito, ICCV 2017], which operate *within* the standard GAN framework, proposing a special 2-stage generator (detail in revised appendices). By contrast, we propose a different GAN *framework* altogether, where adversarial learning occurs in the (jointly-optimized) latent space itself. • **Number of variates in experiment data**: While TimeGAN components can indeed be instantiated with various architectures, we focus on the time series setting, using RNNs to illustrate consistent improvement across a variety of data. Media-specific domain applications (e.g. video) are beyond the paper's scope. However, we agree an even higher-dimensional validation is beneficial. We have conducted additional experiments on UCI Human Activity Recognition ($dim = 561$). In short, TimeGAN achieves 0.062 (21.2%) & 0.012 (12.7%) discriminative & predictive gains relative to the best benchmark (RCGAN). • **Static vs. temporal features**: We will clarify the following before lines 106-107: "Consider the general data setting where each instance consists of two elements: static features (that do not change over time, e.g. ethnicity, gender), and temporal features (that occur over time, e.g. vital signs, clinical events)." Accommodating static features gives the most general framework, since they often accompany temporal data (e.g. patient data). However, static features are *not* required (we can simply drop the non-recurrent parts of e, r, g, d); the novelties of TimeGAN are in how it handles temporal aspects. • **Further details on architecture**: We will publish full source code for TimeGAN with the camera-ready manuscript; this will contain all specifications, training settings, and parameters necessary for reproducibility. Furthermore, in addition to lines 37-59 in the appendices, we will tabulate all technical information with the same granular level of architectural detail (down to dimensions of individual variables) as Appendix B in [Lucic, ICML 2019]. Moreover, for additional sensitivities on hyperparameters λ, η, see response Hyperparameter sensitivities for Reviewer 3. • **Discriminative metric**: To quantify the fidelity of synthetic samples (among other desiderata; see response Evaluation metrics... for Reviewer 2), we use the discriminative metric to gauge how indistinguishable samples are from actual data. First, actual sequences are labeled "real", and sampled sequences "not real". Then, an off-the-shelf (RNN) classifier is trained to distinguish between the two (a standard supervised task). We are not doing any *pairwise* testing for differences between individual sequences.

[Reviewer 2] • **Evaluation metrics, datasets, and benchmarks**: In the familiar application of GANs to images, the vast majority of evaluation relies on inception scores and variants, as well as visual fidelity by inspection; importantly, observe that the former is based on a separately trained model [Salimans, NIPS 2016]. This approach is virtually universal [Lucic, ICML 2019; Brock, ICLR 2019; Wang, ECCV 2018]; furthermore, using post-hoc classifiers for the evaluation of generative models is well-established [Isola, CVPR 2017; Zhang, ECCV 2016]. In the context of time-series GANs, we observe *three* comprehensive desiderata: (1) *fidelity*—samples should be indistinguishable from real data; (2) *diversity*—samples should be distributed to cover the real; and (3) *usefulness*—samples should be just as useful as real data when used for the same predictive purposes (i.e. train-on-synthetic, test-on-real). In our evaluation, the discriminative score, t-SNE/PCA analyses, as well as predictive score respectively give measures of (1), (2), and (3).

Our approach to evaluation is *much more comprehensive* than prior works on GANs for time series. First, they do not address (1) and (2) directly. C-RNN-GAN uses a single dataset, relying on hand-crafted measures of audio fidelity. RCGAN uses a post-hoc classifier to evaluate usefulness of samples (i.e. (3)), tested on MNIST and a single real dataset (very low $dim = 4$); other metrics are only applied to synthetic data. By contrast, we focus on all 3 desiderata. Second, we provide experimental results across 6 competing benchmarks over all metrics; RCGAN and C-RNN-GAN provide zero. Third, our 5 datasets are specifically picked to vary with respect to dimensions, correlations, periodicity, discreteness, etc. (see lines 239-254), including a massive ($n = 150k, dim = 54$) real-world medical dataset (Events). For these reasons, we submit that our approach to evaluation is *more* comprehensive—especially w.r.t. RCGAN as the reviewer mentions. Furthermore, see response Number of variates in experiment data for Reviewer 1 for additional results on even higher-dimensional ($dim = 561$) data. • **Static vs. temporal features**: Kindly refer to response Static vs. temporal features for Reviewer 1. • **Discrete data**: We already use discrete data. Our largest real-world dataset ($dim = 54$) is discrete, with $\sim 150k$ sequences (see lines 252-254, and Table 2 in appendices). TimeGAN significantly outperforms all benchmarks on both discriminative/predictive scores (as with all datasets. See Table 2 in manuscript).

[Reviewer 3] • **Difficulty of training**: Although GANs in general are not the easiest to train, we did not discover any additional complications in our experiments. The embedding task serves to regularize adversarial learning—which now occurs in a lower-dimensional latent space; similarly, the supervised loss has a constraining effect on the stepwise dynamics of the generator. For both reasons, we do not expect TimeGAN to be *more* challenging to train; standard techniques for improving GAN training still apply. Here, we use covariance feature matching across all models to improve the diversity of generation. Kindly refer to response Number of variates in experiment data for Reviewer 1 for equally favorable results on even higher-dimensional data. See also the following response. • **Hyperparameter sensitivities**: We find empirically that TGAN is not very sensitive to λ and η. While we set $\lambda = 1, \eta = 10$ for all experiments, we agree that showing additional sensitivities may be beneficial. We have conducted experiments across a range of hyperparameters, tabulated in the revised appendices. For example (for Stocks), across $\lambda = \{1, 5, 10, 20\}$ and $\eta = \{0.1, 0.5, 1, 2, 5\}$, the min and max discriminative scores are 0.097 and 0.108, with variance 0.004—showing that performance is by no means brittle—thereby providing further reassurance that TimeGAN is not more difficult to train.