We thank all reviewers for their input on our paper. All reviewers recognized the novelty and appeal of the repulsive term in our proposed loss, which is indeed the main technical contribution of this work. Since many previous works use spectrogram-based losses without such a term, and since adding the term greatly boosts performance, we also expect this contribution to have the greatest impact on future research.

The main improvement asked for by the reviewers is the addition of a set of ablation experiments to more clearly tease apart the contributions of the different loss terms and modeling choices. We're currently hard at work running and MOS testing all mentioned combinations of losses and models, and we report currently obtained results below. All results for all experiments proposed by the reviewers will be added to the final version of our paper.

New ablations (R3, R4) Partial results for requested new ablations. Missing MOS scores were not yet finished and will be added later.

model→	**r+m+G**	**r+m**	**r+256+G**	**r+512+G**	**r+256**	**r+512**	
cFDSD→	0.033	0.033	still running	still running	0.035	0.034	
MOS→	4.25 ± 0.06	4.10 ± 0.07			3.44 ± 0.07	2.89 ± 0.09	
model→	**m+G**	**m**	256+**G**	512+**G**	256	512	**G**
cFDSD→	0.039	0.039	0.200	0.047	0.040	0.038	0.075
MOS→		3.00 ± 0.07					4.16 ± 0.06

Table 1: Showing results for all combinations of (1) repulsive term (**r**) yes/no, (2) multi-scale (**m**) or single window size (256/512) or no spectrogram loss, (3) GAN loss (**G**) yes/no. Note that these ablations sampled the cFDSD validation set uniformly, where we used length-weighted sampling for the submitted paper and the table below.

model→	GED + full GAN-TTS	GED + uncond. GAN	GAN-TTS only
cFDSD→	still running	0.040	0.077
MOS→		4.25 ± 0.06	4.16 ± 0.06

Table 2: Results for combining our proposed GED loss with full GAN-TTS, including the conditional discriminators (results pending), and comparing against GED + unconditional GAN, and GAN-TTS, as requested by R4.

Small flow-based models do exist (R3, R4) Our related work section now acknowledges that WaveFlow has made great progress in reducing the size of flow-based models while maintaining generation quality. A remaining advantage of our model is that it is still more parallel / has fewer sequential steps, as our models are fully convolutional while WaveFlow is a hybrid between autoregressive and parallel flow-based models.

Limited performance gains against baseline (R4) Our model with only the GED loss performs about the same as the SOTA GAN-TTS model from Binkowski et al. (MOS scores are not statistically different). However, our method is easier to train and converges faster. GED (which includes the repulsive term) also dramatically improves upon the baseline without repulsive term: Since using spectrogram-based losses without repulsive term is standard practice, we feel that comparison against this baseline is most informative in forecasting how useful the proposed techniques will be for the wider community.

Repulsive term not highlighted or studied enough (R4) Reviewer 4 notes that the repulsive term in our proposed loss is the main technical novelty in the paper, but that it should be highlighted and studied more. We now put even more emphasis on this contribution, and we add new ablation studies to better understand its impact (see above).

GED+iSTFT benefits from directly generating spectrograms (R4) Our iSTFT generator indeed has an inductive bias that might make it easier to do well on the cFDSD metric, however it also does very well on MOS. To avoid any possible confusion: the model still generates raw waveform audio, not just spectrograms.

Relation to *Improved Techniques for Training GANs* (R1) We now discuss the relationship between this paper and our proposed technique in the related work section. In short: the attractive term is indeed similar to the feature matching term in this paper if one would replace the discriminator activations of the feature matching loss with spectogram representations. Our repulsive term directly maximizes the distance in feature space between samples, whereas mini batch discrimination injects sample dissimilarity within a mini batch as side information into the discriminator which is trained with the usual discriminator loss of a GAN.

Add definition of $p(x), q(y)$ after equation 1. (R3) Done.

New citation *A new framework for CNN-based speech enhancement in the time domain*. (R4) Added to related work section.

More information on how to produce the linguistic features (R3) Unfortunately the linguistic features that are used as conditioning input for our speech generation model, and which are generated from the source text by a separate model, are indeed not reproducible using publically available code. We now more fully describe these features in the appendix and, for comparison, we reference many papers that have used these features before.