We thank all the reviewers for their valuable comments and acknowledging the significance and timeliness of this work. The reviewers agree that MelGAN is the first GAN-based method for conditional raw waveform synthesis without distillation or domain specific loss terms. MelGAN has important qualities such as: 1.) fast inference speed (2500 KHz) 2.) trained from scratch and does not require KL-distillation from trained autoregressive models 3.) shows generalization to unseen speakers for the task of mel-spectrogram inversion, 4.) generalizes to at least three different tasks involving strongly conditional waveform synthesis. We believe that these important contributions warrant publication in the conference. To the best of our abilities, we address the following critical comments raised by the reviewers:

Datasets used for all the experiments: For experiment results in tables 3.1 and 3.2, we use the publicly available LJSpeech dataset. For section 3.3, we use a subset of the MusicNet dataset (Thickstun et al., 2016) similar to Mor et al. (2018). For the VQ-VAE experiment, we use the piano dataset provided by Dieleman et al. (2018).

State-of-the-art claims for spectrogram-to-waveform inversion: The authors would like to clarify that MelGAN is a state-of-the-art *non-autoregressive* method for spectrogram-to-waveform inversion *trained from scratch* (does not require KL-distillation from a teacher autoregressive model). Since this definition is quite narrow, we will clarify in the final version that MelGAN is a *high quality* (instead of state-of-the-art) spectrogram-to-waveform inversion method. Admittedly, autoregressive methods such as WaveNet and WaveRNN are slightly better at this task, but we believe future work along this direction will close the gap.

State-of-the-art claims for text-to-speech: Furthermore, we will remove the state-of-the-art TTS claim made in line 87 in the final version. This claim was initially made since MelGAN paired with text2mel shows the highest reported MOS (of 3.88) on the publicly available LJSpeech dataset, beating Tacotron2 paired with WaveGlow (at 3.71). The MOS of ground truth audio in this dataset is 4.72. We did not explicitly compare with Tacotron2 paired with WaveNet since Prenger et al. (2019) show that WaveGlow performs similar to WaveNet in ground truth mel-spectrogram reconstruction. However, we agree that a more direct comparison in the TTS setting is necessary to substantiate our claim. Note that the MOS scores reported in the original Tacotron2 paper cannot be reproduced / compared with due to the unavailability of the dataset or the original code.

Discrepancy in MOS scores between Table 3 and Table 2: The scores for the ablation study in Table 2 specifically compares the importance of different components of the final MelGAN model. For this purpose, we only trained each model for 400,000 iterations (1/6th the time required for the final converged model used in Table 3, which is trained for 2.4 million iterations). This is the reason for the discrepancy in MOS scores in the two tables.

Updates for the final version: The authors will add additional ground truth spectrogram-to-waveform inversion MOS results for MelGAN compared with WaveNet, WaveGlow and original audio, as well as a stronger Text2Mel + WaveNet baseline for TTS. We will refrain from claiming state-of-the-art unless substantiated by these tables.

R1 *: claiming "autoregressive models can be readily replaced with MelGAN decoder" (line 89, line 228) without necessary experiments*

We would like to clarify that this statement was not meant to convey that the perceptual quality of MelGAN decoder is equivalent or better than autoregressive decoders in general. This statement was only meant to express the fact that the MelGAN decoder is successfully shown to work in 3 different experimental setups that traditionally use autoregressive decoders, such as : 1.) inverting mel-spectrograms to the corresponding acoustic waveform, 2.) mapping discrete latents produced by a discrete variational auto-encoder to its corresponding observed waveform, 3.) mapping latent codes produced by a Universal Music Translation Network to the corresponding raw waveform. We believe that this evidence is sufficient to claim that MelGAN decoder is robust enough to replace autoregressive models for strongly conditional waveform synthesis. We will update the paper to better reflect our intention.

In addition, for quantitative analysis of the performance of MelGAN on unseen speakers (without finetuning), we report MOS scores on ground truth mel-spectrogram inversion on the VCTK dataset. We believe that this will serve as a good task to test generalization for future research along this direction. For the sake of brevity, the results are as follows: Original (4.19 ± 0.083), MelGAN (3.49 ± 0.098), Grifin Lim (1.72 ± 0.07). Note that Griffin Lim is rated poorly as there was no additional noisy baseline to anchor the scores causing a stark contrast between Griffin Lim and MelGAN.

R1 *: [...] it's interesting to know how much benefit GAN brings in this work. A baseline to compare with is to train only a Generator model with MSE loss (or other simple loss), without using Discriminator.*

This was an obvious first experiment that we tried. The model completely fails to capture the structure of the acoustic waveform resulting in pure silence.

R3 *: For the comparison in Table 1, it isn't clear at all whether the same hardware was used [...]*

Thanks for the feedback. Yes, the exact same hardware and computing specifications were used to compare all the models. We will clarify this in the footnote.