We thank all the reviewers for their valuable comments.

R1 : We would like to clarify that, 'When the model was trained without the mel-spectrogram loss, the training process became unstable, and some pronunciations were synthesized differently from the ground truth audio.' only corresponds the result of the ablation studies using the V3 generator, which has the smallest expressive power among the three generator variations. The problem we mentioned is also present in MelGAN, and when referring to our additional experiment results[Table 1], it can be seen that the V3 generator performs better than MelGAN without L1 loss.(The details of additional experiments are described in the following section.) We understand that this unexplained part can be misleading, and we will make up for it in the final version. Additionally, MelGAN and GAN-TTS are studies that do not use L1/L2 loss, but there are several studies such as Isola et el.[1], Parallel WaveGAN, and MB-MelGAN to use L1/L2 loss. We also think that applying the L1/L2 loss gives no disadvantage in one-to-one mapping as our work. It is because the loss helps for the generated samples to be close to the target ground truths. All MOS numbers for WaveNet, WaveGlow, and MelGAN are from the publicly available implementation of the models, not sourced from the original papers. We will clarify the details of the experiments in Section 3.

R1 & R3 : To verify the effect of MPD in the settings of other GAN models, we introduced MPD in MelGAN and conducted MOS evaluations. Specifically, we trained MelGAN up to 500k steps and compared it with all samples used in ablation study as well as samples of V1. The summarized MOS evaluation results are shown in [Table 1]. MelGAN trained with MPD outperforms the original one by a gap of 0.50 MOS, which shows statistically significant improvement. Furthermore, considering that V1 and V3 generators outperform MelGAN, we claim that our overall architecture is well tuned. We will update the result of the extended ablation studies to our final version and add related audio samples at the bottom of the demo page.

Table 1: Mean Opinion Scores. All models were trained up to 500k steps.

Model	MOS	95% CI
Ground Truth	4.40	±0.07
MelGAN	3.15	±0.12
MelGAN + MPD	3.65	±0.09
HiFi-GAN V1	4.30	±0.07
HiFi-GAN V3	4.11	±0.07
HiFi-GAN V3 w/o L1	3.43	±0.10

R2 : As for MPD and RWD, leaving the differences such as existence of Markovian window or strided convolutions aside, there is resemblance in the initial convolutional layers as Reviewer 2 pointed out. When the first layer of RWD is the grouped 1d convolution, the 2d convolution with 1xk kernels of MPD becomes similar to RWD, but there is still a difference as RWD would not share weights across each group. In that regard, we argue that our 2d convolution operations in MPD help to increase parameter efficiency. We will add the similarity and difference of MPD and RWD in the paper.
As vocoders are only trained on ground truth mel-spectrograms, fine-tuning with predicted samples from TTS models helps to improve the overall mel-spectrogram-based end-to-end speech synthesis quality. Although Reviewer 2 commented scores for actual speech synthesis for unseen text look meagre, we believe we showed the strength of our model in that our model scored higher than the comparison model even before fine-tuning, and after fine-tuning, the quality improved further.

For MOS tests, we set WaveGlow, MelGAN and MoL WaveNet as our comparison group, because they use 1) commonly used mel-spectrograms as input conditions and 2) they are publicly available in the open source community. Since the mentioned models such as WaveRNN, ParallelWaveNet, and GAN-TTS are reported to use different input conditions called linguistic features, which makes them hard to reproduce, we do not set them as our comparison group.

R3 : Thanks Reviewer 3 for the suggestions and concerns. We will conduct additional experiments to verify the effectiveness and limitation of MPD for capturing periodic patterns and add the results to the final version.

R4 : Following the valuable suggestions, we will provide more discussion about the related works, which Reviewer 4 mentioned, in the new version of the paper.

[1] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.