1. Dataset and SOTA benchmark comparisons [R1, R3, R4]: Several of the SOTA methods mentioned by the reviewers are concurrent to our work, appearing a few weeks before NeurIPS deadline. During the rebuttal period, we performed benchmark comparisons to these methods (suggested by R3 and R4), including TF-GAN (Soni et al. 2018), Metric-GAN (Fu et al. 2019), SDR-PESQ (Kim et al. 2019), T-GSA (Kim et al. 2020), Self-adaptation DNN (Koizumi et al. 2020), and RDL-Net (Nikzad et al. 2020). Trained on the same dataset (i.e., Valentini's DEMAND) as in those methods and tested in the same way, our method is highly competitive to the best of those methods under each metric (different existing methods perform the best under different metrics). For example, under PESQ, our method produces **3.16**, while existing methods reported the following scores: 2.53 (TF-GAN), 2.86 (Metric-GAN), 3.01 (SDR-PESQ), 3.06 (T-GSA), 2.15 (Self-Adaptation DNN), 3.02 (RDL-Net). Under CBAK, our method produces **3.54**, while others are: 3.12 (TF-GAN), 3.18 (Metric-GAN), 3.54 (SDR-PESQ), 3.59 (T-GSA), 2.82 (Self-Adaptation DNN), 3.43 (RDL-Net). We also evaluated speech intelligibility using STOI. Our method produces **0.98**, while RDL-Net reported 0.94 (other methods do not report this score). We are happy to include a full set of comparison scores in the revision of the paper.

Why creating our own dataset? We would also like to clarify that there are strong reasons for creating our own dataset. First, it is important to train/test denoising models under a wide range of noise levels, including strong noise (i.e., low SNRs). Unfortunately, existing datasets have limited ranges of SNR levels. For example, DEMAND consists of audios with SNR in [0dB, 15dB]. This is significantly narrower than what we wish to test against (e.g., with an SNR of -10dB), because many real-world recordings, as we show in supplementary material, have strong noise (much lower than 0 dB SNR). Secondly, a key challenge faced by any denosing model is to suppress *structured* noise. Existing datasets like DEMAND provide several kinds of environmental noise, but lack the richness of other types of structured noise (such as music). The lack of noise diversity makes the datasets less ideal for denoising real-world recordings.

To support our rationale, we performed an experiment to compare our dataset (AVSPEECH+Audioset) with DEMAND for real-world denoising. We train two versions of our model using our dataset and DEMAND, respectively, and use them to denoise real-world recordings (those shown in supplementary material). For the denoised real-world audios, we measure the noise level reduction in detected silent intervals. This measurement is doable, since it requires no knowledge of noise-free ground-truth audios (which are not available for real-world recordings). In every tested real-world recordings, the noise reduction from the model trained with our dataset is significantly higher than that from the model with DEMAND. On average, it produces **22.3 dB** noise reduction in comparison to **12.6 dB**, suggesting that our dataset allows the denoising model to better generalize to many real-world scenarios.

2. Test on real-world data [R1, R2, R3]: We believe that all methods must be examined against real-world data. Quantitative evaluation on real-world data, however, is not easy. To evaluate a denoised signal, all the metrics (like SSNR, PESQ, CBAK, STOI, etc.) require to know the clean, ground-truth signal, which is not available for real-world recordings. It is for this reason that almost all existing denoising models were evaluated on synthetic data, and we follow this common approach to evaluate our model quantitatively.

Nevertheless, we did one step further than previous works. In the supplementary webpage, we reported denoising results of various *real-world recordings*. These are recordings we downloaded online or recorded in daily environments. These are recordings in diverse scenarios: in a driving car, a café, a park, on the street, in multiple languages (Chinese, Japanese, Korean, German, French, etc.), with different genders and accents, and with singing songs. None of these recordings is cherry picked. And we also showed the results from other methods, including the widely used professional audio processing software, Adobe Audition. The point we wish to convey is: testing our method with real-world data and letting the reader to judge by themselves how well our method can generalize to different real-world scenarios. While these are qualitative results, to our knowledge no previous work has reported such diverse real-world results.

3. Generalization across datasets [R1, R3]: Our real-world results in the supplementary material serve for demonstrating the generalization ability of our model in various real-world scenarios. In addition, we performed *three* cross-dataset tests to evaluate the generalization ability of our model, as suggested by R1. **i)** We train our model on Audioset but evaluate it on DEMAND testset—the same testset as in other methods. The PESQ score is **3.00**. As comparisons, our method trained on DEMAND training set yield a PESQ score of 3.16, and the best score reported in previous methods (trained and tested on DEMAND) is 3.06 (T-GSA). **ii)** We train our model on AVSPEECH+Audioset and evaluate it on our second dataset in which the noise is from DEMAND. The PESQ score is **2.65**, and for comparison, our model trained and tested with DEMAND noise distribution yields a core of 2.80. **iii)** We train our model on our second dataset (in which the noise is from DEMAND) and test it on AVSPEECH+Audioset. The PESQ score is **2.12** in comparison to 2.30 from our model trained and tested on Audioset noise distribution. We also note that we couldn't compare the generalization ability of our model with existing methods, as no previous work reported cross-dataset evaluation results.

4. Metrics for the noisy signal [R3]: Thanks for your suggestion. We measured the noisy signals under the six metrics we used. Corresponding to what we reported in Fig. 5, these scores are: 2.72 (PESQ), 3.04 (SSNR), 0.85 (STOI), 3.55 (CSIG), 2.70 (CBAK), 3.05 (COVL) for DEMAND noise, and 1.91, 3.46, 0.71, 2.79, 2.34, 2.27 for Audioset.

5. Google WebRTC VAD [R3]: We understand that VAD is designed to work with low-noise signals. It is included in Table 1 merely as an alternative approach that can detect silent intervals. We will clarify this point in the revision.