Rebuttal for Review 1:

Weak Supervised Dereverberation Models:

We will add to introduction

I noticed that some \*\*fricative sounds exhibit strong artifacts\*\* in the proposed USD-DPS method. I would appreciate it if the authors could comment on the origin of these artifacts and whether they are inherent to the method.

Thanks for pointing this out. We started to observe these artifacts from implementing the vanilla BUDDy in our test set. We believe this is caused by the alternative style of updating the RIR and diffusion step. We listened and observed that the initialization of BUDDy (10-step WPE) is not strong, such that the early stage of the speech is blurry and the RIR estimation is off. During the later optimization process, no matter for vanilla-BUDDy, MC-BUDDy, or USDSPS, the RIR will gradually converge, but make some small adjustments to let the RIR model converge to the “local minima” in terms of the current diffusion step. Then, the diffusion model will take the likelihood score from this “overfitted” RIR optimizer to produce a dereverbed speech, and this “overfit” scenario will happen again in the next iteration. Thus, we may observe some “ghost” artifacts after the whole diffusion process. This is even more obvious in our experiment of using FCP for all channels’ likelihood score estimation. Just because FCP is a relatively greedy method without much signal processing content, it will overfit in every iteration, and the dereverberation result will contain a strong artifact, just like there is a second voice coming from a different frequency range

Rebuttal for Review 2:  
Please discuss the possibility of the speech signal arriving at non-reference microphones before the reference microphone  
We enforce one reference microphone to respect and simulate the nature of time-arrival of microphone arrays. This could be the case. But even in beamforming tasks (e.g. MVDR), selecting a reference microphone doesn’t receive the signal first will only cost a constant time shift in result. And that’s why we introduced SI-SDR as one of our metric. On the other hand, the issue of time-alignment in the nature of diffusion based model is a problem that we won’t tackle in this paper. So thanks for pointing this out, yet we believe our assumption is within the reasonable range

Hyperparameter choices such as epsilon, zeta, and lambda.

Even though we’ve conducted the experiments of hyperparameters, with the page limitations, we couldn’t put it in the paper. We’ve include it in the demo page now.