We thank the reviewers for the overwhelmingly positive feedback (scores: 7 7 7). We are encouraged that all reviewers appreciated our paper for the following: (i) the study on the robustness of FL is timely and important (**R2**), (ii) the proposed edge-case backdoor attack is effective and realistic (**R1**,**R2**), (iii) our work provides theoretical understanding of the vulnerability of FL (**R1**, **R4**), (iv) the experiments are extensive and solid (**R1**,**R2**,**R4**). Each reviewer provided helpful suggestions to improve our manuscript that we address below while providing additional experimental results.

R1:*"Baselines are limited to* KRUM *and RFA.(suggested additional baselines: coordinate-wise trimmed mean/median)."*
We first note that our submission indeed considered a total of four defence methods: KRUM, RFA, "*norm difference clipping*" (NDC) and "*weak differential privacy*" (weak-DP) proposed in [27] (shown in Figure 5(d), 6(c)). As per your suggestion, we have added the evaluation of our edge-case backdoor attacks (*blackbox*, *PGD with*, and *without replacement*) against coordinate-wise trimmed mean on the "Southwest airlines" example. The results indicate a stronger robustness of coordinate-wise trimmed mean compared to KRUM and RFA. Our attacks still manage to inject the backdoor, although they take longer (approximately 3× FL rounds *i.e.*, 1500 rounds to reach a target task accuracy of 35%). We have added a detailed description of the additional experiments in the revision.

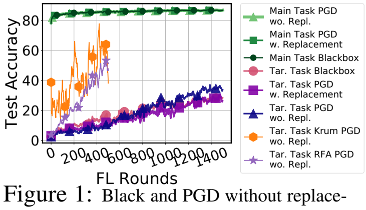

Figure 1: Black and PGD without replacement edge-case attacks under coordinate-wise trimmed mean (with fraction 10%) on "Southwest" example with CIFAR-10 dataset.

R1:*"Figures [...] too small [...] References are out-of-date."* Fixed!

R2: *Concerns related to the backdoors injected via $\mathcal{D}_{edge}$*
(i) *Novelty:* The semantic backdoor in [13] consists of a special slice of the dataset (*e.g.*, cars with racing stripes). However, that slice may not necessarily be rare in the underlying distribution. For instance, "images of airplanes in the blue sky" are common (semantic backdoor), however "images of WOW airplanes" (edge-case backdoor) are likely to be underrepresented. Our edge-case backdoor emphasizes that the slice of the dataset has to be rare to make the attack effective. Plausible evidence to this is the poor performance of blackbox (i.e., data poisoning) attack in [13]. It is likely that racing striped cars were not "edge-case enough" for the attack to go through despite being semantically sound. We note that the choice of "Southwest airplanes", however, is sufficient for a successful attack. (ii) *Construction of p-edge case attack:* Yes, choosing a dataset that is entirely out of distribution *e.g.*, CIFAR in the MNIST example would work. However, people are unlikely to use a model trained on MNIST for prediction on CIFAR examples. Thus such a "fully out-of-distribution backdoor" is not as practical. We agree that constructing a *good* p-edge-case dataset is non-trivial. We are actively working towards this direction. (iii) *Setup of $\mathcal{D}_{edge}$:* Yes, the default setting is that only the adversary has access to D_{edge}. The absolute number of edge-case examples varies across experiments *e.g.*, for Task 1, we randomly sample 66 of the ARDIS images and mix them with 100 randomly sampled EMNIST images. More details are included in Appendix A. (iv) *Evaluation:* Our evaluation did not include a comparison with BadNets since the latter requires both training-time and inference-time access to the data for inserting and triggering the pixel-pattern backdoor, while in the general FL scenario, the attacker has (at most) only training time access. We have added a more comprehensive discussion on trigger-based backdoors including BadNets in the revision for completeness.

R2: *"Concerns related to Proposition 1 and 2."*
We have added more intuition in the main text as per your suggestion. Proposition 2 uses a simple construction to demonstrate that there exist backdoors which are hard to detect. For scenarios having non-uniform data distribution, the hardness of backdoor detection can be confirmed by Proposition 1.

R2: *"Clarification on the experimental results."*
(i) *Questions on Fig 4:* We partition the correctly labeled instances among a set of sampled honest clients while having the attacker hold **all** incorrectly labeled instances. For Fig 4(a), we have 200 correctly labeled "Southwest" images. We assign 100 to the attacker and 100 to 5 out of the 200 honest clients (each client holds 20). The goal is to create an "almost edge-case" scenario where it is unlikely that the backdoor gets erased by clients who hold correctly labeled edge examples. (ii) *Questions on the DP defense:* We agree that the NDC defense minimizes the influence of a particular user's data. We did observe that NDC with a stricter norm bound leads to poor target and main task accuracy (sometimes preventing full model convergence). Thus in our experiment, we tune NDC (and other defenses) such that the main task accuracy reaches around 90% in 500 FL rounds (for the EMNIST experiment in Fig 5(d)). When studying the weak DP defense (built on top of NDC) under a wide range of noise levels, we observe that while high noise is effective in defending against the attack, it hurts the main task accuracy too much (Fig 6(c)).

R4: *"Thm 1: $XX^\top$ being invertible – I am not sure if this assumption holds in practice."*
We verified this assumption by considering a FC ReLU network of width 2000 trained on MNIST. Randomly selecting 1000 data points gives us $\mathbf{X} \in \mathcal{R}^{1000\times784}$ of rank 574. However, the activation matrix of the first layer is $\mathbf{X}_1 \in \mathcal{R}^{1000\times2000}$ which has rank 1000 $\Rightarrow \mathbf{X}_1\mathbf{X}_1^T$ is invertible. We have included the experimental results in the revision.

R4: *"[...] $\mathcal{D}_{edge}$ sets [...] actually having "exponentially small measure"? [...] heuristic [...] creating defenses?"*
The log density plot in Fig2 shows that ARDIS is of very small measure with respect to MNIST. We also included this metric for other datasets in the revision, however the success of the attack further supports the claim. For practitioners, a good heuristic would be the main-task accuracy of the model. Assuming that the test set is reasonably representative of the underlying distribution, having uncompromised test accuracy proves that the backdoor is sufficiently small.