We would like to thank the four reviewers for their feedback. We first discuss a few common concerns shared by **Reviewer 2**, **Reviewer 3** and **Reviewer 5**.

●●● **Comparison with distributed SGD using quantization or sparsification**:

Reviewer 1: We thank the reviewer for valuable comments and references. We would like to make the following clarification:

Discussion on the assumptions:

Reviewer 2: We thank the reviewer for the useful comments and typos. Our point-to-point response is as follows:

Numerical Runs: We present in Section D of the Appendix, additional runs on CIFAR-10 showing similar performance of our method. The number of local updates τ has been set to 1 and 5 in the main text and we added runs with $\tau = 2$ in the Section D of the Appendix as well. Larger number of local updates τ tend to undermine the learning performance as we have observed empirically. In the heterogeneous setting, increasing τ can present a risk of learning bad local models. We acknowledge that there is a trade-off to be found here between speed of convergence and the quality of the local models (to obtain a good global one).

Reviewer 3: We thank the reviewer for valuable comments. We clarify the following point on the comparisons:

Comparison with other compressors:

Reviewer 5: We thank the reviewer for valuable comments. Below we address your concerns:

Additional Numerical Experiments: Additional runs on CIFAR-10 are presented in the Appendix (Section D). While runs with different ratio of active devices at each iteration is interesting, we reported results with a practical one (half of the devices) for illustrative purposes. We agree that rigorously comparing the number of bits transmitted between FedSGD and our methods is interesting. Yet, we give the important values of 12 and 75 compressing ratio yielding a good order of magnitude on this latter quantity. Our method being almost as fast as FedSGD, despite the high compressing ratio, shows its benefits.