We sincerely thank the five reviewers for their valuable feedback.

Reviewer 6: We thank the reviewer for valuable comments. Our point-to-point response is as follows:
Notations: In Lemma 2, the notation $\bar{\nabla} f(.)$ corresponds to the the average of $\nabla f(\theta_{r,i})$ over all clients $i \in [n]$. This is defined in the appendix (page 6) but will be clarified in the main. The subscript t will be changed to r, thank you for noticing that.

Convergence Bounds: We would like to highlight the usage of such suboptimality condition (LHS) in for instance["On the Convergence of Decentralized Adaptive Gradient Methods", Chen et. al, 2021]. The denominator being always bounded from above by construction (see adam like update), our bounds are acceptable. Of course, we do not claim to have established a tight convergence bound. Deriving a matching lower bound would be the only solution to prove so, but is not our focus and is in general a very challenging problem.

Assumptions: We decided to assume, in H2 and H3, that the variance and the gradient are uniformly bounded across all devices. That is the reason that we omitted the subscript i.

Reviewer 13: We thank the reviewer for the thorough analysis. Our remarks are listed below:
Notations: We will revise the necessity of the hyperparameters. The function $\Phi(.)$ has been set to the identity function for the purpose of illustration. More sophisticated function could be considered but is not the focus of our paper. Please refer to the original LAMB paper for more related discussion on this aspect.

Proofs: We will include in the appendix an extension of our proof for Lemma 1 with $T > 1$. We omitted this case given that the bound is constructed using uniform bounds on m_t and v_t.

Experiments: The grid search has been done thoroughly for each baseline method. The number of clients and the learning rate schedule have been carefully chosen. Extensive number of runs have been performed for all baselines and our method in order to present the best performing runs for each.

Reviewer 14: We thank the reviewer for valuable comments. We have fixed the typos. Our response is as follows:
Smoothness assumption: Theorem 1 does display a dependence on the smoothness L which is equal to the aggregated layered smoothness $\sum_{\ell=1}^{L} L_\ell$. The first term on the RHS of Theorem 1 also shows a dependence on h which is the total number of layers in the model. See the notations paragraph page 2 for its definition.

Experiments: The three datasets we use are common benchmarks in federated learning literature, and we tested three different architectures with both iid and non-iid data client distribution. Thus, we believe that our results are sufficient and convincing to demonstrate the advantage of our method. Please find below some additional results on a new dataset FMNIST (which is also a popular benchmark in FL) and another baseline method, the adaptive federated method (Adp-Fed,) proposed in [Adaptive Federated Optimization, Reddi et. al., 2020]. Our method performs the best on these tasks among all baselines. We are happy to add these results in the paper.

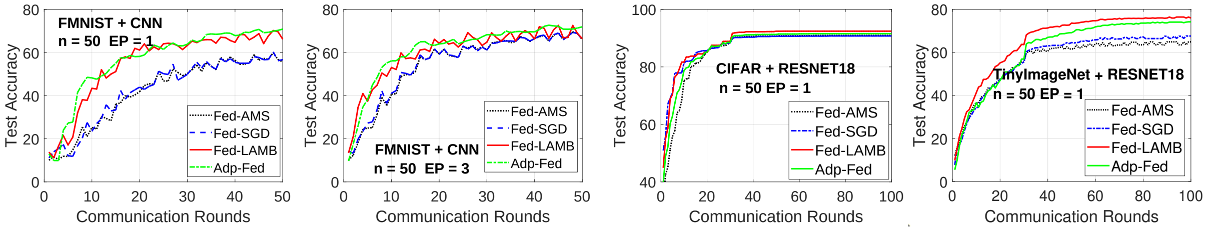

Figure 1: **non-i.i.d. data setting.** Test accuracy on FMNIST, CIFAR-10 and TinyImagenet with 50 clients.

Reviewer 15: We thank the reviewer for valuable comments:
Discussions on the assumptions: The smoothness assumption, the bounded variance and gradient norm are very classical in any optimization contributions (either central or federated). Our contribution is most importantly focusing on a novel method for federated learning combined with deep neural network than on the originality of the assumptions.