R2: "The main shortcoming of your algorithm though is from the practical side: you need to synch the normalization factor, which means that your algorithm is two-rounds."

Yes, our algorithm needs an extra round. It is because sparsity can be caused by either dimension unbalance (a few dimensions have much larger values than others) or data unbalance (a few clients have vectors with much larger norms than others). To handle data unbalance and obtain a bound depending on the global sparsity, some form of coordination is necessary, otherwise clients would have no idea whether their vectors are heavy or light compared to others.

However, we would like to point out that, our algorithm can also be run without the synchronization round. For this setting, we can derive a communication bound for each client by simply setting $n = 1$ in Corollary 2.4, although s in the bound will become the local sparsity of the client when doing so. Local sparsity bound is worse than global sparsity when there is data unbalance, but the bound is still better than prior work as long as there is dimension unbalance. This is also verified in the experiments: In Fig 1(a)(c)(d), the data sets used do not have data unbalance (meaning the coordination round is effectively useless), and the results are still better than previous methods.

R3: "The lower bound seems to only consider SMPs and doesn't allow for the downlink communication by the centre. In this sense, I feel that the optimality of the proposed scheme has not been established in a strict sense."

Major factual error: The lower bound is proved in the broadcast model (see Sec 1.3, as well as Theorem 3.1). The broadcast model allows multi-round protocols and free downlink communication (i.e., each message is public). Thus, lower bounds proved in the this model also hold for other communication models, such as SMP or the message-passing model.

Our upper bound is derived under the message passing model only using private communication. Therefore, the optimality of the proposed scheme has been established in a strong sense: no protocol can beat the proposed scheme (up to a constant factor), even if that protocol is given free downlink or broadcast communications.

R3: "The resulting lower bound is obtained by Fano's inequality."

Factual error: We have not used or even mentioned Fano's inequality in the paper. Our proof is based on the multiparty information complexity framework, which is different from techniques for proving minimax lower bounds, e.g. Fano's. This is also the reason why our lower bound hold for the broadcast model.

R3: "some of which the authors have not cited (such as the ATOMO algorithm which provides a method for handling sparse vectors)"

Factual error: We have cited this paper, which is [22] in the reference; and we explained the difference in section 2.

R3: "Authors have a rather simple scheme, but it still requires coordination between the parties by the centre."

Please see our response to R2 above.

R3: "I feel that random rotations such as randomized Hadamard transform should have been used to avoid this."

(1) Sparsity can be caused by either dimension unbalance (a few dimensions have much larger values than others) and data unbalance (a few clients have vectors with much larger norms than others). Our scheme handles both types of sparsity. Random rotations cannot handle data unbalance, and even worse, it *destroys input sparsity*. (2) Even if there is no input sparsity, the cost of the random rotation method is sub-optimal by a log factor. It is worse than the variable length coding method from the same paper both theoretically and empirically (see [20]), so we didn't say much about it in our paper. (3) The random rotation method needs public randomness, since all clients and the server need to use the same random matrix. This in practice also requires coordination between the parties. See our response to R2 above for more comments on this issue.

R3: "There is not much innovation in the scheme or the lower bound. Both seem to follow the template of [20]."

Major factual error: Our lower bound does not follow the template of [20] at all. We adopt the information complexity framework and use Yao's minimax principle; the main technical part is a proof of the lemma which roughly says if a combinatorial rectangle contain too much entropy then a random input in it must have large variance. On the other hand, the lower bound in [20] relies on the statistical lower bound from [27]. To get the lower bound, the authors of [27] apply classical techniques for proving minimax lower bounds, which are quite different from ours. We have emphasized in the paper that we cannot simply use the same idea as in [20], because current results on statistical estimation is insufficient to obtain the desired lower bounds. As a result, our lower bound is able to exploits sparsity and is better than [20]. And it holds in the broadcast model as opposed to just independent protocols as in [20].

We agree that our protocol uses a similar framework as prior work (randomized quantization and encoding). However, the main contribution is its sparsity-sensitive analysis and the resulting tight bounds. In addition, the particular quantization and encoding methods are new in this framework, which are important for obtaining the tight bounds.