We thank the reviewers for the generous comments and invaluable feedback on the manuscript. Please find below our point-by-point response to reviewers' concerns.

Reviewer #1

- *"Analysis of wall-clock time breakdown":*
  We will move the wall-clock time breakdown analysis into the main text of the paper.
- *"Literature of architectures that are explicitly easy to distribute":*
  Thank you for pointing out the related work, which we will cite in the camera-ready version. We would also like to emphasize that although mixture-of-experts-type architectures are relatively easy for device placement, the dispatching of examples and gathering of experts outputs is often non-trivial, and comes at the cost of expert load balancing, expert utilization, batch-size tuning etc. We will add a detailed discussion regarding this in the camera-ready version.
- *"Overview the architectures for AmoebaNet and Transformer":*
  We will include a brief overview and figures for the architectures considered here in the appendix.
- *"Suggestions on improvements"*:
  Thank you very much for the suggestions. We have been running experiments with very deep transformer language models and are currently exploring approaches to stabilize training for these models, beyond those discussed in the paper. We will try to share our findings in the camera-ready version. At the same time, we will soon open source the code to train 128+ block transformers on Github to allow researchers to train these models with their own data.

Reviewer #2

- *"Batch size experiments feel a bit out of place":*
  Thank you for the suggestion. We added this discussion in order to provide more details for our machine translation experiments. We will move this set of results into the supplementary material and move the wall-clock time breakdown analysis into the main paper as suggested by reviewer #1.

Reviewer #3

- *"Elaborate the scheduling algorithm":*
  We agree with the reviewer that we should have done a better job at explaining our algorithms in greater details. Currently, we allow the user to specify the per-layer computation cost for each layer (c_i) as measured in flops, as briefly mentioned in Section 2.1. For distributing layers across accelerators we try to balance the per-accelerator computation cost by minimizing the variance in the estimated cost of all accelerators. We have open sourced our heuristic algorithm in our open-source repository. Since actual computational time might differ from the flops estimation c_i, depending on different accelerators and architectures, we also allow the user to manually specify layer placement. We are also planning to further improve our scheduling algorithm to balance both memory utilization and computational time in order to maximize batch sizes and training efficiency. We will describe the scheduling algorithm more clearly in Section 2 in the camera-ready version.
- *"How possible node failures are taken into account":*
  Our system uses synchronous training so any one of the machine failures would force the whole system to restart from the previous checkpoint, as with standard non-model parallelism training.
- *"The micro-batching algorithm details are not presented in the paper.":*
  The overview of the micro-batch algorithm is depicted in Figure 2(c). We first split the tensor of mini-batch input along the batch dimension into micro-batches and apply control-flow ops to loop over each micro-batch to compute the gradients. Accumulated gradients over all micro-batches are applied to variables at the end of each global step. We will describe the micro-batching algorithm more clearly in Section 2 of camera-ready.
- *"The lack of empirical benchmark system.":*
  We agree that providing a unified benchmark system for scaling giant neural network would be beneficial to the research community. However, we are afraid it is a bit beyond the scope of our paper to define such a benchmark system to evaluate all model parallelism libraries. For example, Mesh TensorFlow does not support convolution operations, making it infeasible for us to compare our image model performance with theirs.
- *"Missing reference in the supplementary material":*
  Thank you for pointing this out. We will fix this in the camera-ready version.