We thank all reviewers for their thorough reviews and insightful feedback! We are encouraged that they found the proposed idea to be impactful (all reviewers), clear (all reviewers), novel (R1,R2), principled (R3,R4) and applicable to other areas (R2). Specifically, all reviewers recognized our contribution to training deeper models which is an important problem to tackle and the resulting improvement on three sequence modeling tasks over current state-of-the-art (SOTA) approaches. We are also encouraged that the benefit of harvesting a sparse model with reduced inference cost was appreciated (R1,R3). We appreciate that reviewers found our experiments comprehensive and sound (R1,R3,R4), especially acknowledging our comparison to prior work (R1,R3) on training deep models as well as with sufficient ablations and discussion (R2,R3,R4) on modeling choice and sensitivity of hyperparameters. Below we address specific questions which permit more discussions. We will incorporate all suggested improvements in the final version.

[R2] "The biggest weakness of this work lies in its lack of comparison with existing machine translation models." We appreciate you emphasizing the importance of comparing to SOTA models, which we could not agree with more and we'd like to emphasize that we **did** compare to several existing SOTAs. First, we compared the standard Transformer with increased width (Table 1,3,4) which has shown strong performance in bilingual and multilingual MT (Arivazhagan et al. (2019)). Second, we also compared to prior approaches for training deeper (>30 layers) Transformers (e.g. DLCL and LayerDrop in Table 1$\sim$ 5). We did not compare to Zhang et al. (2019) because (1) our method is independent of initialization and thus different angles (just as Zhang et al. (2019) did not compare to non initialization-based methods; (2) most results in Zhang et al. (2019) on deeper models are 12 layers which we found did not diverge and our focus was on >24 layers. We missed Zhang et al. (2020) since it was published at ACL'20 which is *one month after* our submission. But we will include both and relevant multilingual MT references within it in the final version.

[R2] "unspecified measure of variance" in Table 1. It is the standard error after running with different seeds.

[R1] "24/24 outperforms 12/100" This is a correct statement, however, it results from a combined effect of increased decoder depth and reduced encoder depth. In Table 4, we compared 12/100 (24.16 BLEU) to 12/24 (23.7 BLEU) so as to isolate the effect from increased encoder depths. We did found increased encoder depths brings improvement (especially for low resource) and balanced encoder/decoder depth brings consistent gains as is illustrated in Fig 1. We did not include it in the submission due to the space limit, but we will add it in our final version.

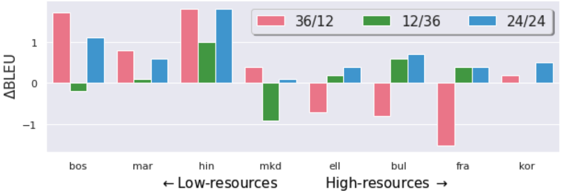

Figure 1: Quality improvement (over static depths 12/12) by allocating increased capacity to all-encoder (36/12), all-decoder (12/36), and balanced (24/24).

[R1] "how initialization techniques interacts with the proposed method": Our approach solved the same problem (i.e. gradient vanishing) as initialization-based approaches but from a different angle. In addition, it does not require specific initialization (which is usually architecture-dependent) but instead can work with different initializations thanks to the implicit gradient re-scaling effect from latent layers. Also, it enables faster training and inference by learning a shallower network.

[R1] Additional ablation studies. We had ablation studies on different loss terms as our approach has few bells and whistles. We provided analysis on the effect of different modeling choices and hyperparameters (in Section 5 and the Appendix) which R2, R3 and R4 found useful.

[R3] "Would it be possible to compare the performance of latent depth multi-lingual models to non-latent depth models of the same depth on single high-resource language pairs?" Yes. We have done the suggested comparison on the fra-eng (high-resource pair) from the TED corpus where we compared to the strongest bilingual models (12-layer since static depths diverges for deeper models). Our approach outperforms for both directions: fra-eng 40.25 (bilingual static depth 24/12), 40.0 (multilingual static depth 24/12), 41.2 (multilingual latent depth, 24/24); eng-fra 40.22 (bilingual static depth 24/12), 40.3 (multilingual static depth 24/12), 41.5 (multilingual latent depth 24/24).

[R4] "deeper models also increase the inference time." This is true for deeper models with static depth. However, our approach allows pruning to a shallower network (especially decoder depth which contributes to the majority of the inference time) and thus addresses the above challenge.

[R3] "It would be interesting to see a quantitative evaluation of to what extent different languages use the same layers" We have computed the Hamming distance of layer selections (among 36 layers) between related (9.33 ± 1.24) and unrelated languages (17 ± 2.34) respectively, where we can see the former has a higher degree of parameter sharing.

Depend on language embeddings [R1] or input [R3]: Good points! As we elaborated in L104-L108 about the trade-offs, we chose to obtain clear learnings without being confounded by the quality of these additional parameters. But our approach can be easily extended to use both and we fully agree with exploring them in future work.

References

N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

B. Zhang, I. Titov, and R. Sennrich. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 897–908, 2019.

B. Zhang, P. Williams, I. Titov, and R. Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*, 2020.