All: We thank the reviewers for their insightful feedback! We feel encouraged that they (R1, R3, R4) appreciate the fact that our method achieves **SOTA** in En-De, En-Fr WMT'14 translation tasks and also outperforms the baselines (*e.g.,* Scaled Transformer) significantly (1-2 BLEU) in 8 other translation tasks in IWSLT and Flores (low-resource) despite being simple and universally applicable (R1, R2, R4) across different NMT architectures. While **all** reviewers appreciated our thorough analysis, some raised concerns regarding the *translationese effect* (R2, R4). In the following, we address this concern along with other specific comments and mischaracterizations.

1. Translationese effect (R2, R4): Our method is **NOT** affected by this. To verify, we performed an experiment with the En-De testset provided by Edunov et al., (ACL'20), and compare our method with the Scaled Transformer baseline where both systems were trained on the standard WMT dataset from Vaswani et al., (2017). The table below shows that our method consistently outperforms the baseline in **all 3 scenarios** tested in Edunov et al. Meanwhile, Edunov et al. show that BT outperforms only in $X^* \rightarrow Y$ (source translationese to natural target), thus suffering from the translationese effect, while **our method does not**.

En-De WMT'14	$X \rightarrow Y^*$ (natural src $\rightarrow$ translationese tgt)	$X^* \rightarrow Y$ (translationese src $\rightarrow$ natural tgt)	$X^{**} \rightarrow Y^*$ (translationese src $\rightarrow$ translationese tgt)
Baseline	31.35	28.47	38.59
Our method	33.47	30.38	41.03

Our explanation is simple: the BT method of their paper is a **semi-supervised** setup that uses **extra natural monolingual data in the target**. In our method, back-translation is conducted on the translationese part (target side) of the parallel data, and does not enjoy the introduction of natural text which benefits only $X^* \rightarrow Y$. Simply put, back-translation **with** and **without** monolingual data are two different concepts that should not be confused.

In the paper, we sufficiently conducted **9** different analyses (**3** in Appendix) to understand our method better and to distinguish it from BT. It is therefore natural for us to miss the translationese effect. We urge you to allow us to include the translationese analysis and reconsider your decisions.

Reviewer #1 (R1) **(1)** The scaled Transformer is the same as vanilla Transformer, only with more GPUs. The vanilla Transformer is no longer used much for WMT experiments, so we don't think it is necessary to include it, but we can include it if the reviewers disagree with us. **(2)** Perplexity is calculated in the standard way, *i.e.,* the *exponential of cross-entropy*. The perplexity is reversely proportional to BLEU during training (as expected). However, **the best valid set perplexity** (for selecting the model) does not always correlate with the **testset BLEU.** Vaswani et al., (2013) points out that sometimes we get higher BLEU by sacrificing perplexity using methods like temperature-controlled softmax. **(3)** Table 9 is only to show that our method complements BT, not to compete for SOTA (which would require huge compute). The choice of News Crawl 2009 was just random in that aspect. We do not claim our method performs similarly as BT, we claim it complements BT. Introducing **extra monolingual data** gives BT an unfair advantage in comparing with ours. **(4)** Yes, we believe our method has a regularization effect. The vocabulary is joint BPE.

Review #2 (R2) **(1)** We had to put the Diversification details in the Appendix (C) due to page limit. We can bring it back to the main content, if reviewers want. **(2)** There is no extra inference cost compared to standard Transformer. In case you talk about data generation, it costs about 7hrs, which is 30% time to train the baseline. **(3)** About BLEU gains, we compared with SOTA and reported about 1-2 points improvements across the datasets (WMT, IWSLT), which everyone would agree is quite decent if not huge. For your reference, most of the recently published papers (*e.g.,* Shaw et al., 2018; Edunov et al., 2018; Wu et al., 2019) report only 0.1-0.5 BLEU improvements in the standard WMT testset. Table 5 is to show that our approach has an ensemble effect but not to compare with Ensemble as ensembling requires **7x more memory and computations** (thus not fair to compare). And yes, we did **perform checkpoint averaging** (mentioned in Appendix B), so the gains are **not spurious**. **(4)** About Sec. 5.1, we do **not** assume the models have *perfect fit* on the data. In fact, ensembling addresses and regularizes the issue that the models do not have a perfect fit.

Review #3 (R3) **(1) We humbly disagree with this definition of "novelty". Presuming simple methods as non-novel and biasing "novelty" only towards complex architectures can be a hindrance to scientific progress.** NeurIPS should value efforts that are robust, effective and make high impact, yet have not been tried before. Our work is novel as no one has tried using multiple forward & backward models to augment data, and makes high impact as it pushes SOTA by a decent margin. **(2)** We tried bidirectional models but didn't work. Hassan et al. has a semi-supervised setup with one model inducing confidence to the loss of another, while ours are trained independently.

Review #4 (R4) **(1)** The BLEU evaluation is **exactly the same** as previous SOTA papers (Vaswani et al.,2017;Shaw et al., 2018;Edunov et al.,2018; Wu et al.,2019). We use their BPE code and measure tokenized BLEU (Appendix B). Most previous work didn't report SacreBLEU. In case, you want to know SacreBLEU on WMT, here are the numbers with the default detokenized SacreBLEU (*sacremoses* and *sacrebleu* scripts): Scale Transformer gets **28.5** in En-De, **40.9** in En-Fr; our method gets **30.0** in En-De and **41.9** in En-Fr. **(2)** Out-of-domain generalization could be a good extra analysis, where our method has the potential to do better as it is trained on more diversified data. However, notice that we had to leave out many details in the Appendix due to the page limit. We can include it in a later version.