General response

We thank all reviewers for their valuable feedback and thoughtfull suggestions.

> Comparison with Gu et al., Stern et al, Welleck et al

- > Direct comparison with Gu et al.

  To the best of our knowledge, there is no official implementation for the paper by Gu et al. (no link to the code in the paper, no repository stating that it's an implementation for the paper). Thus, direct comparison using the original implementation was not possible.

  However, in Section 5.1 we compare the lower bound on the objective we use with the one of Gu et al.(2019), when applied to our model. This is a more fair comparison of the two approaches since both model architectures can be used with these lower bounds and should be compared separately.

- > Comparison with Stern et al, Welleck et al

  These works do not report significant improvements in BLEU scores against the autoregressive baselines. Stern et al.(2019) focus on parallel decoding (with the final result matching the vanilla Transformer). Welleck et al.(2019) find left-to-right models superior to their approach.

While we took into account these papers for the presentation of our work, we want to highlight that this was concurrent work.

> More experiments on standard data sets, e.g., WMT EN-DE

We agree that this is important for future research on this topic and would allow for easier comparison with our and previous work. While we can not provide these results during the author response period (due to large training time of NMT models for high-resource language pairs), we will add them should the paper get accepted.

Response to reviewer 1

> Why not consider tasks without language output? ... For example, you could consider music generation tasks, ...

Note that we consider not only natural language output, but also Image-to-Latex, where output is LaTex formulas.

We believe applying our approach to music generation task could be an interesting direction for future work.

> Notation issue: ... Do you mean that you also compute an expectation over sampling from $T^*(Y)$ first, and then use that to sample τ?

By $\tau \sim p(\tau|X, T^*(Y), \theta)$ we actually mean $\tau \sim p(\tau|X, \tau \in T^*(Y), \theta)$. Thus, we sample the trajectory from our model, but only allow the correct insertion operations (i.e. the ones that lead to producing Y).

We will make it more clear and expand the discussion.

> The footnote 8 on page 6 is important. You should have introduced [28] earlier and more explicitly.

> In Section 3, you write ... I would suggest moving some of this discussion up earlier; it will make the presentation of the algorithm clearer.

We agree this would improve the presentation. We will revise the paper accordingly.

Response to reviewer 3

> how to decide the output length

We describe this in Section 2.2. Specifically, we introduce the EOS element, which indicates the end of generation process. The model can produce this element instead of a pair of position and a token.