To All Reviewers We wholeheartedly thank all reviewers for all valuable and insightful feedback, along with taking the time to review our submission. We have provided detailed comments to each reviewer below.

To Reviewer 1 Thanks for the valuable and insightful feedback!

Regarding the suggestion about more experiments on more datasets, we full wholeheartedly agree that more experiments will definitely improve the comprehensiveness of the paper. We are running experiments on these tasks right now and include results on these datasets (SST, SNLI, WMT En-De etc.) by the camera ready version as supplementary material.

Pertaining to Q3, weights are extracted from a trained model (on NLI) on sampled $1K$ datapoints from the dev set. Similar graphs appear with repeated sampling. Regarding deletion being prevalent, our hypothesis is that only tokens that are particularly significant will be compositionally (add/subtracted). For most words, the sigmoid function provides flexibility of deleting tokens.

Regarding the value of α, β in equation 1 and equation 2, we set them to 1 in the experiments. The higher value of α, β is, the 'harder' the compositional pooling becomes. The hardness of CoDA required is possibly related and analogous to hard vs soft attention and can be domain dependent. For most language tasks, we find that not biasing CoDA towards being hard is quite sufficient. Not all tokens are important, so CoDA maintains the flexibility of standard attention while enabling arithmetic compositionality. We will include more supplementary distribution visualisations on different tasks in the revised version.

Regarding the form of Equation 9 used in the experiments, we used one layer non-linear projection network to compute the pairwise affinity. We thank the reviewer for pointing it out and we will mention it clearly in the revision.

For other comments such as references or typos, we will correct and add them in the revision. We will also be sure to include a discussion about sparsemax and softgen.

To Reviewer 2 Thanks for the insightful and valuable feedback!

Regarding the evaluation on Tensor2Tensor, IWSLT En-Vi was chosen because of the size and resource limitations. Please be assured that the tasks were not cherry-picked and we have not experienced any failure cases on T2T tasks yet. We also had success on the arithmetic T2T and subject-verb agreement tasks but did not report due to lack of space. We will prepare detailed supplementary material to cater to extra experiments. Moreover, we are currently running WMT En-De and WMT En-Ro. This may take sometime due to our limited hardware but will definitely be ready by the camera-ready version.

Regarding the form of CoDA, we use the following version our experiments, i.e. subtract the mean before applying the sigmoid or tanh, and they are scaled before applying self-attention. We will make this absolutely clear in the final version. We left several variations open in our technical exposition since certain hyperparameters (e.g., α) could allow practitioners to control properties (i.e., hardness) of CoDA. We apologize for any confusion. In the revision, we will also include more ablation studies of different alternations of CoDA form, providing better guidance for usage of CoDA configurations. We will also include supplementary visualisation as requested.

Regarding the confusion of using notions and references in the papers, as pointed out in the detailed comments, we thank the reviewer for pointing them out. We will correct and clean them in the revision.

To Reviewer 3 Thanks for the insightful and valuable feedback!

Our intuition is that tanh saturates at $\{-1, 1\}$, which doesn't allow the model the flexibility to *delete* (erase) tokens. Sigmoid provides this flexibility to our model. Some early ablation results on retrieval tasks are reported in Table 1. We will include more comprehensive ablations in the final version.

Thanks for the suggestion about adding a visualization of the learned attention weights on examples for interpretability. We will definitely include it in the revision. Pertaining aesthetic comments, thanks for pointing them out and we will be sure to correct them in the revision.

Method	TrecQA	WikiQA
tanh only	66.78 / 73.49	67.13 / 67.81
CoDA	79.84 / 84.78	68.07 / 68.28

Table 1: Dev set results of *tanh* only versus CoDA (*tanh* * *sigmoid*).