Model Ablation	SOURCE			TARGET	
	Accuracy	Median	Mean	Median	Mean
Position Features Only	50.0%	0.0	2.47	3.03	3.41
Last Hidden State Attention	47.9%	2.28	2.88	3.15	3.48
Dual Attention	51.4%	0.0	2.47	3.03	3.32
Annealed Sampling Loss	51.3%	0.0	2.56	3.29	3.46
Non-Joint Training for Source	51.3%	0.0	2.53	-	-
Non-Joint Training for Target	-	-	-	3.28	3.46
Full Model (Expectation)	52.2%	0.0	2.44	2.91	3.23
Expectation Model w/ Ensemble	54.1%	0.0	2.35	2.85	3.14
Sampling Model w/ Ensemble	52.8%	0.0	2.41	3.09	3.25

Table 1: Validation results to show ablations of our model components. Our full expectation-based model (third-last row) uses all features (coordinates, relative, stack), CNN attention, and joint training. Each ablation row above that shows the results of changing one component at a time from this full model. Finally, the two last rows represent the final 8-ensemble versions of the full expectation model, as well as the sampling model. (note that lower is better for median and mean distance values)

Model	SOURCE			TARGET	
	Accuracy	Median	Mean	Median	Mean
End-to-End FFN (**?**)	9.0%	3.45	3.52	3.60	3.94
End-to-End RNN (**?**)	10.0%	3.29	3.47	3.60	3.70
Our Expectation Model	56.1%	0.00	2.21	2.78	3.07
Our Sampling Model	56.3%	0.00	2.18	3.12	3.18
Our Expectation Model w/ Ensemble	56.6%	0.00	2.12	2.65	2.91
Our Sampling Model w/ Ensemble	56.8%	0.00	2.11	2.71	2.90

Table 2: Final test results of our final sampling and expectation models (w/o and w/ ensemble), compared to the previous state-of-the-art on this dataset.

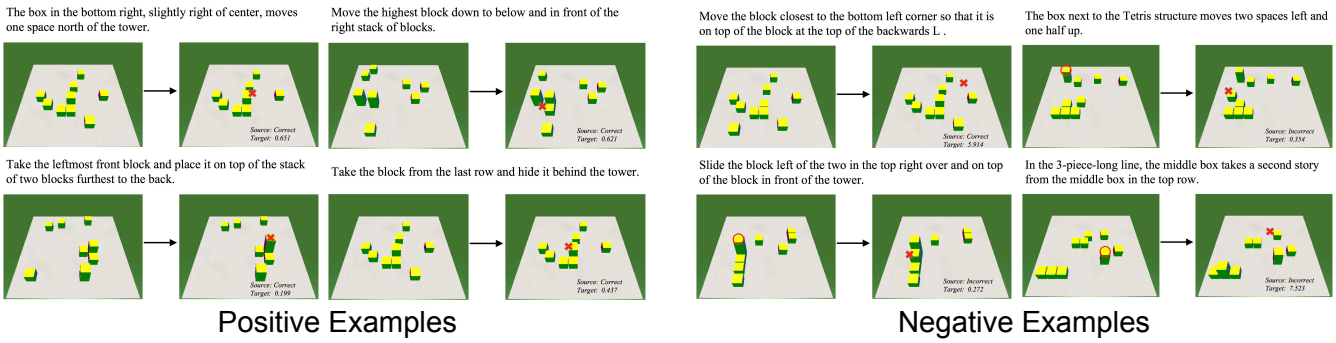

Figure 3: Analysis: positive and negative output examples showing interesting instruction scenarios. The first and second image in each pair depict the ground truth movement of the source block to the target position. We report predicted source accuracy and target distance in bottom-right of each second image. We also use a red cross to represent our predicted target position (ground truth target position can be inferred directly from the image difference between the first and second image). Also, for the cases where our model predicted an incorrect source, we represent that wrongly-predicted source block by a red circle.

major model component choices as discussed in Section 3: the block-world spatial features, the three attention modules, the sampling vs expectation target loss, and the joint vs non-joint training of the embedding representations. In Table 1, the full expectation-based model (third-last row) represents the model which uses all features (coordinates, relative, stack), CNN attention, and joint training. Each ablation row above this third-last row shows the results on changing one component at a time from this full model. Finally, the last two rows of Table 1 add an 8-sized ensemble to the full expectation (as well as the sampling) model; and this setting is used for the final test results in Table 2.

Feature Selection: To show the impact of different representations of the world blocks, we compare the results of using just the coordinate values vs our novel relative and stack-based features (discussed in Section 3.1). As shown

in Table 1, utilizing these new features gives us some decent improvements (2% in source accuracy and 0.18 in target mean distance). **Bilinear Attention Modules**: For both the source block and reference block selection, we model the distributions by three different bilinear attention modules (discussed in Section 3.1): bilinear matching between the last hidden state of the instruction and each block, CNN filters on top of the LSTM-RNN vectors, and dual word-to-block and block-to-instruction attention. The comparison among these three attention modules is shown in Table 1. The models using CNN filters or dual attention outperform the one with the last hidden state, on the source and the target tasks. The CNN filter attention model is slightly better than the dual attention model, and hence we use that in the final full model. Note that the dual attention model is similar (within standard deviation) to the CNN attention model in performance. In Sec. 5.2, we also discuss the complementary nature of the CNN and dual attention models, and report their improved combination results.

Target Training Methods: As shown in Table 1 (and discussed in Section 3.2), the model for target prediction is trained with two types of inference methods and optimization loss functions. Using the expectation loss gives us slightly better performance than using the sampling loss (a 0.23 decrease in the validation target mean prediction, and a 0.11 decrease after an ensemble).[3] This is likely because the two losses use quite different inference procedures. The sampling inference explicitly chooses (samples) one block as the reference block while the expectation inference calculates the reference by the expected (weighted) sum of several blocks, and both inference choices have their advantages vs. disadvantages (e.g., the sampling method actually allows us to output an interpretable single block as the target reference block, as opposed to the expectation approach), hence we report results for both models.

Joint Training: The fourth part of Table 1 compares joint training vs non-joint training of the world-block and language representations across the source and target tasks (as discussed in Sec. 3.3). The non-joint training results for the source and the target tasks are worse than the joint training results, showing the advantage of learning shared spatial world and language representations across source and target tasks, via joint loss function optimization.

Finally, the last row of Table 1 shows the added effects of an 8-sized standard ensemble approach.

5.2 Final Test Results

Next, in Table 2, we present the test-set results for our two inference approaches (expectation and sampling), using the final model choices based on the ablation studies (i.e., all features, CNN attention, joint training), without and with ensemble. Both inference models achieve strong improvements over the previous best work on this dataset from **?** (**?**), who employ three neural models for this task. We compare to their final best model, the RNN-based 'end-to-end' neural model (as well as their second-best feed-forward network FFN model). Our model achieves 47% improvement in source task accuracy, and 22% (0.8 block length) reduction in target distance mean. Moreover, the results are quite stable for both inference models: the **standard deviation** based on 8 runs is around 1% on source accuracy and 0.05 block length on target mean.

Complementarity of Attention Models: We found that our two attention models (CNN and dual) are complementary in nature, achieving a source accuracy of **57.70%** when combining the ensemble models of CNN and dual attention (for the expectation case), i.e., an improvement of 1.1% over the CNN model's 56.6% in Table 2. Further experiments in this direction (as well as the complementarity of the sampling and expectation inference approaches) is future work.

5.3 Analysis

Figure 3 shows several positive and negative examples of the output of our full model. We can correctly understand the semantics in complex source and target descriptions such as 'bottom right, slightly right of center' and 'place it on top of the stack of two blocks furthest to the back'. In the negative examples, we show complex cases that our model cannot handle correctly, mostly due to special scenarios and phrases that it hasn't seen before in the diverse but small dataset. Examples of this include instructions mentioning shape-based block patterns such as 'backwards L', 'Tetris structure', and complex count-based patterns such as '3-piece-long line'.

6 Conclusion

We presented sampling and expectation based models for source and target prediction in configurational robotic instructions (on a challenging blank-labeled blocks dataset). Our models also use spatial-relative features, CNN and dual attention models, and joint-subtask-loss training of world and language representations, achieving substantial improvements over previous work on all metrics.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was partially supported by a Google Faculty Research Award, a Bloomberg Data Science Research Grant, an IBM Faculty Award, and Nvidia GPU awards.

Omnis quis repellendus, consequuntur vitae facilis sit, est quam reprehenderit deserunt, nulla odit quo ducimus aperiam corrupti modi ipsam eum, labore doloribus suscipit quasi corrupti dolores culpa assumenda consequuntur id natus.Iusto blanditiis fugit nostrum minus aliquam id reiciendis animi culpa doloremque, alias ea quas, possimus dolorum molestias quasi itaque consequatur adipisci aliquid eligendi ut quae eveniet, error eaque vel corporis voluptate quos vero quo.Cum nobis a nisi laborum, beatae obcaecati tenetur mollitia placeat ipsum commodi aliquid deleniti, dicta illum odio amet cupiditate natus sunt eius assumenda aspernatur molestias debitis?Dignissimos facilis amet illo illum, asperiores quam similique fuga a nemo fugiat ab debitis alias quis, labore harum temporibus in magni, temporibus aliquam rerum incidunt dolore.Odit mollitia sit fugiat eveniet sunt eius ut error amet nesciunt,

[3]Note that the vanilla, standard sampling approach performed significantly worse than our annealing-based method (described in Sec. 3.2), achieving a target median of 3.57 and a target mean of 3.82 on the dev set.