

	Source Acc.	Target			
		Gold	Source	End-to-End	
		Mean	Med	Mean	Med
Bisk 16	98	–	–	0.98	0.0
Pišl 17	98.5	–	–	0.72	–
Ours	97.5	0.7	0.14	0.80	0.14
v2	91.3	1.2	0.85	1.15	0.88
v1 + v2	95.9	1.0	0.50	1.10	0.51
v1 + v2 → v1	98.1	0.8	0.15	0.84	0.15
v1 + v2 → v2	93.1	1.2	0.88	1.35	0.91

Table 3: A comparison of our interpretable model with previous results (top) in addition to our performance on our new corpus (v2). Finally, we show how training jointly on both corpora has only a very moderate effect on performance, indicating the complementarity of the data. Target values are error measurements in block-lengths (lower is better).




Error	Goal	Instruction
4.8		use sri as the base of a fourth tower to the left and equidistant with the other tower
5.2		spin sri slightly to the right and then set it in the middle of the 4 stacks
6.4		in the emerging 3x3 grid place texaco in the middle left

Table 4: Several of our worst performing results. Errors are in block lengths, the images are the goal configuration, and the instructions have been lowercased and tokenized.

(three degrees). In validation, 46% of predictions require a rotation. 1,374 of 1491 predictions are within 2 degrees of the correct orientation. The remainder have dramatically larger errors (36 at 30°, 81 at 45°). This means that the model is learning to interpret the scene and utterance correctly in the vast majority of cases.

### Error Analysis

Several of our model’s worst performing examples are included in Table 4. The model’s error is presented alongside the goal configuration and misunderstood instruction.

The first example specifies the goal location using an abstract concept (tower) and the offset (equidistant) implies recognition of a larger pattern. The second example specifies the goal location in terms of “the 4 stacks”, again without naming any of them and in 3D. Finally, the third demonstrates a particularly nice phenomenon in human language where a plan is specified, the speaker provides categorizing information to enable its recognition, and then can use this newly defined concept as a referent. No models to our

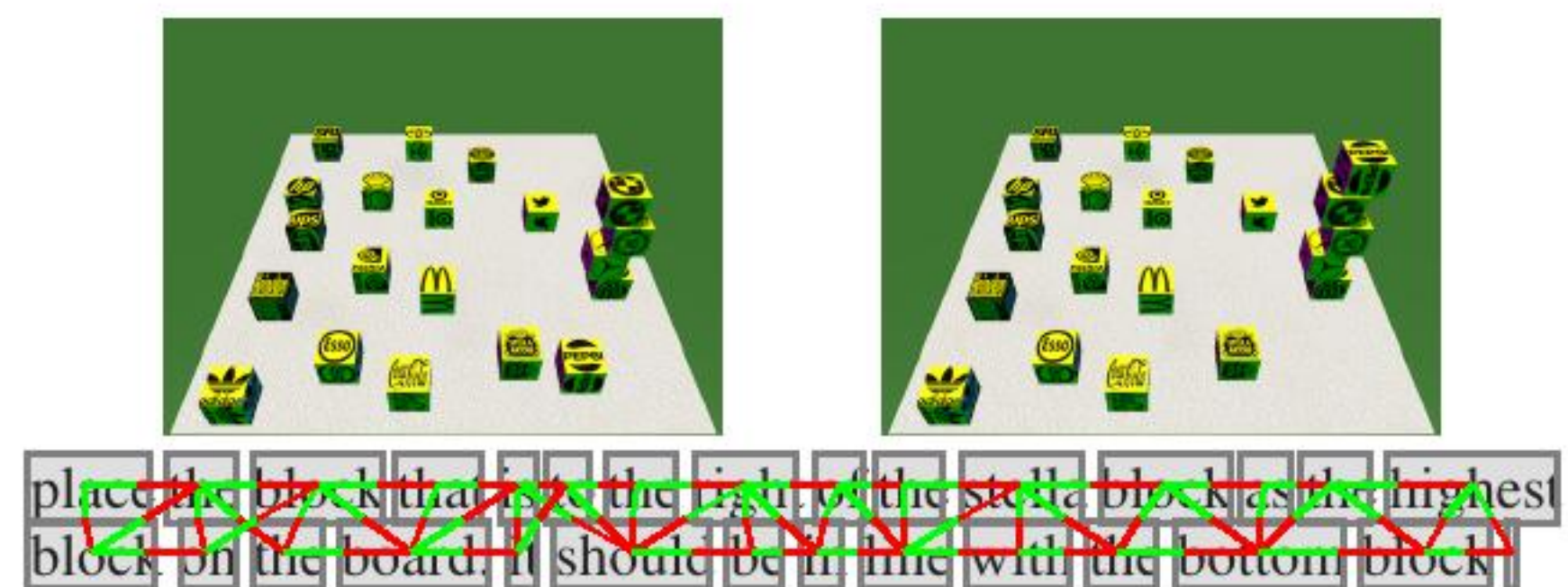


Table 5: Example utterance which requires both understanding that highest is a 3D concept, and inferring that the 2D concept of a line has been rotated to be in the z-dimension.

knowledge have the ability to dynamically form new concepts in this manner.

**Rotations** Despite a strong performance by the model on rotations, there are a number of cases that were completely overlooked. Upon inspection, these appear to be predominantly cases where the rotation is not explicitly mentioned, but instead assumed or implied:

- place toyota on top of sri in the **same direction** .
- take toyota and place it on top of sri .
- ... making part of the inside of the **curve of the circle** .

The first two should be the focus of immediate future work as they only require trusting that a new block should trust the orientation of an existing one below it unless there is a compelling reason (e.g. balance) to rotate it. The third case, returns to our larger discussion on understanding geometric shapes and is probably out of scope for most approaches.

### Conclusions

This work presents a new model which moves beyond simple spatial offset predictions (+x, +y, +z) to learn functions which can be applied to the scene. We achieve this without losing interpretability. In addition, we introduce a new corpus of 10,000 actions and 250,000 tokens which contains a plethora of new concepts (subtle movements, balance, rotation) to advance research in action understanding.

### Acknowledgments

We thank the anonymous reviewers for their many insightful comments. This work was supported in part by the NSF grant (IIS-1703166), DARPA CwC program through ARO (W911NF-15-1-0543), and gifts by Google and Facebook.

### References

- Andreas, J., and Klein, D. 2015. Alignment-based compositional semantics for instruction following. In *Proc of EMNLP*, 1165–1174.
- Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Learning to compose neural networks for question answering. In *Proc of NAACL*, 1545–1554.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*.