A Preprocessing Rules

In order to distinguish the target block and landmark references in an instruction, we first identified verbs in the training split that were used to denote the main action, such as *move*, *place*, *take*, *or put*. We also identified other lexical items that indicate target blocks and pick the target block via the following set of rules that are applied in order until a target is found:

- 1. if there is a action verb (such as *move*), is its argument a block reference?
- 2. if there is an action verb, the first linearly succeeding block reference (not all of them were identified as VERB by spacy, we override the parse)
- 3. does the instruction start with a block reference? *10 to the left of 4*
- 4. is the parse's root a verb (other than the annotated ones) that has a block reference as argument?
- 5. is there a with-PP? *Continue the horizontal line with hp*

All other block references are annotated to be *landmark blocks*.

We extract string spans denoting the spatial relation by extracting phrases that have as head one of the following dependency or POS tags:

- dependency tags: advmod, prep, mark, attr, acomp, advcl
- POS tags: ADP

Prepositional modifiers starting with "of" and single-token spans of the word "so" are extended to include the head, and the attached phrase, respectively.

B Template Generator

The template generator builds synthetic instructions based on 3 slots: the target block, a landmark block, and the relation between them. We extract slot information for a given image pair using the scene parser described above. For a landmark, we choose the block closest to the target block's landing position. For relation, we compute the compass direction. The template generator randomly chooses from a small number of instruction predicates and translates the compass relation into a fixed set of relation descriptors. Block references follow the pattern "block \$name". Table 8 shows the translation from slots to natural language.

Slot	Value	Mapping
Action		add, move, pick up, place,
predicate		position, put, slide, take
Block	\$digit	block \$digit
Block	\$logo	block \$logo
Relation	N	above
Relation	S	below
Relation	E	to the right of
Relation	W	to the left of
Relation	NE	above and to the right of
Relation	SE	below and to the right of
Relation	NW	above and to the left of
Relation	SW	below and to the left of

Table 8: How the template generator maps slots to natural language

Natural Instruction	Synthetic Instruction				
position ups so its top	pick up block ups and				
edge touches twitter's bot-	move it below block twit-				
tom edge.	ter				
put the target block in the	pick up block target and				
first open space above the	move it above block tex-				
texaco block.	aco				
slide the stella artois box	pick up block stella artois				
up and to the right until it	and move it above block				
is directly above and lined	target				
up with the target box.					

Table 9:Samples of instructions from logo data ofBLOCKS dataset.

C Results on digit data

Following, results of the experiments with all model types are reported on the digit data in table 10. As mentioned in section 5 the digit data is more difficult to learn, most likely because the digits are harder to identify by the model.

Model	BLEU	METEOR	CIDEr	ROUGE-L	GT_T	GT_{LM}	Ref_T	Ref_{LM}
NN-retrieve	0.1169	0.1906	0.1186	0.3815	0.0756	0.1512	0.0788	0.157
CNN+LSTM	0.2391	0.2135	0.2373	0.5724	0.0698	0.2093	0.0872	0.2442
$CNN+LSTM+I_b$	0.2763	0.2296	0.2936	0.5882	0.1047	0.2267	0.1047	0.2616
CNN+LSTM+Att	0.2164	0.2282	0.1642	0.4651	0.0814	0.1802	0.0988	0.1628
$CNN+LSTM+Att+I_b$	0.3816	0.2752	0.2994	0.5596	0.1686	0.2616	0.1918	0.2849
Template	0.4501	0.3778	1.1809	0.6304	0.9419	1.0	0.9419	0.9884

Table 10: Model performance on the block data with digits with natural instructions. I_b specifies a modified image input using the concatenation of the Add/Subtract image modifications. For the Template model, we compare each template instruction with the human made references. In each column, the highest score (except those from template model) are marked in bold.

D Model study on digit data

To extend the model study, in table 11 are the delta values of the scores from different task variants achieved by the LSTM+CNN model.

Modification	Δ BLEU	ΔGT_T	ΔGT_{LM}
none	0.2391	0.0698	0.2093
add	0.0276	-0.0175	0.0291
sub	0.0598	0.0174	0.0174
both	0.0372	0.0349	0.0174
state	0.25	0.8662	0.5988
state + synthetic	0.6493	0.8721	0.6802

Table 11: Delta of model performance with modifiedimage inputs on digit data and natural instructions.