Learning Interpretable Spatial Operations in a Rich 3D Blocks World

Yonatan Bisk — ISI/UW
Kevin Shih — UIUC
Yejin Choi — UWashington
Daniel Marcu — Amazon.com

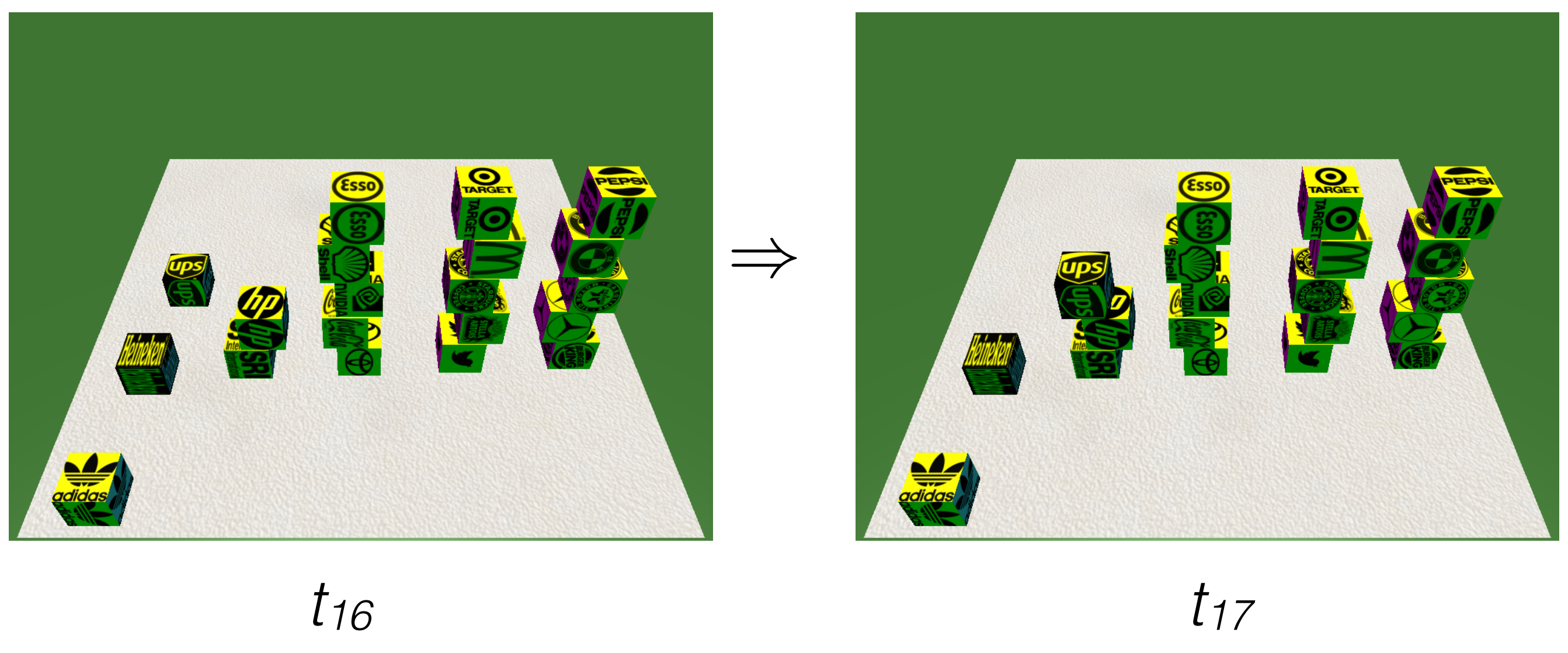
ybisk@cs.washington.edu — https://groundedlanguage.github.io/

Goal: Grounding Spatial Relations

Domain: 3D block configurations and annotated instructions.

Instruction: *“On the (new) fourth tower, mirror Nvidia with UPS.”*

Transition:



Did we correctly place and/or rotate the block?

Component tasks

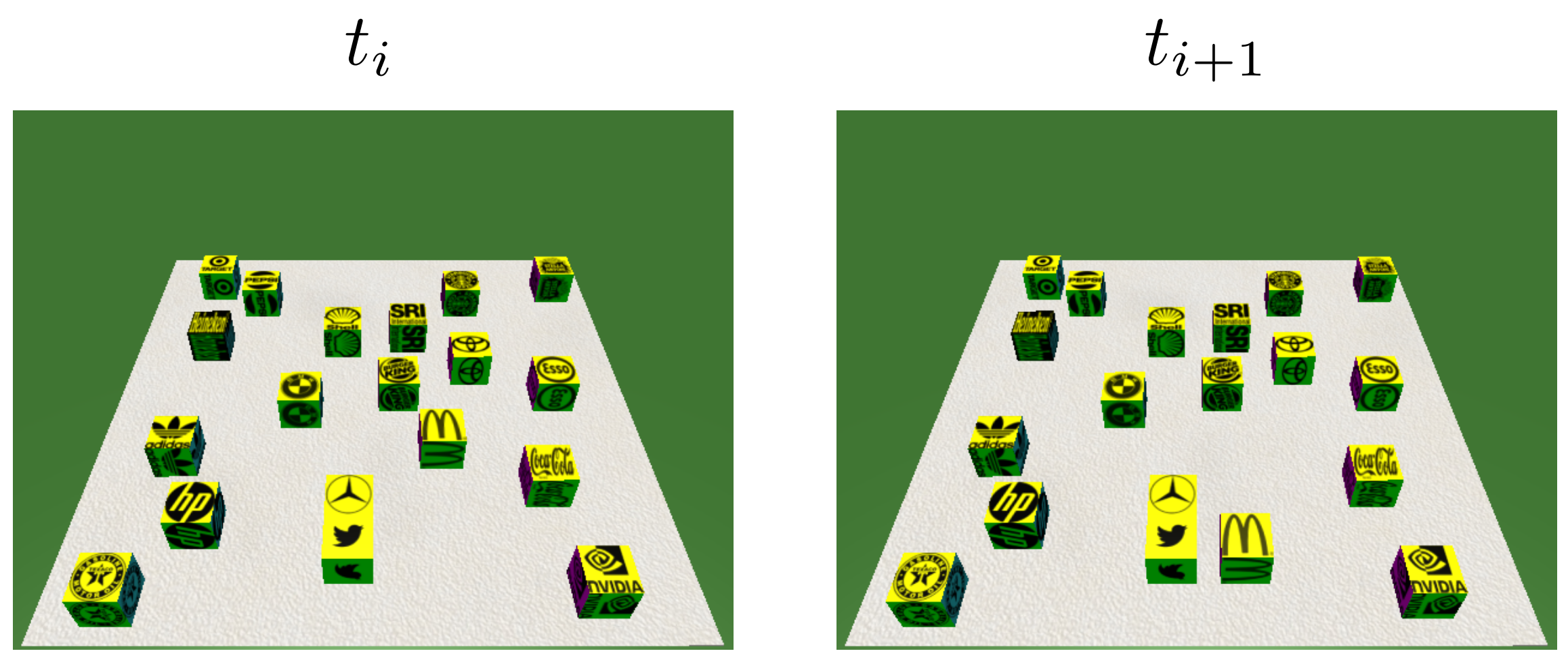
- Grounding referents
- Spatial relations
- Scene understanding
- Abstract Language

Evaluation

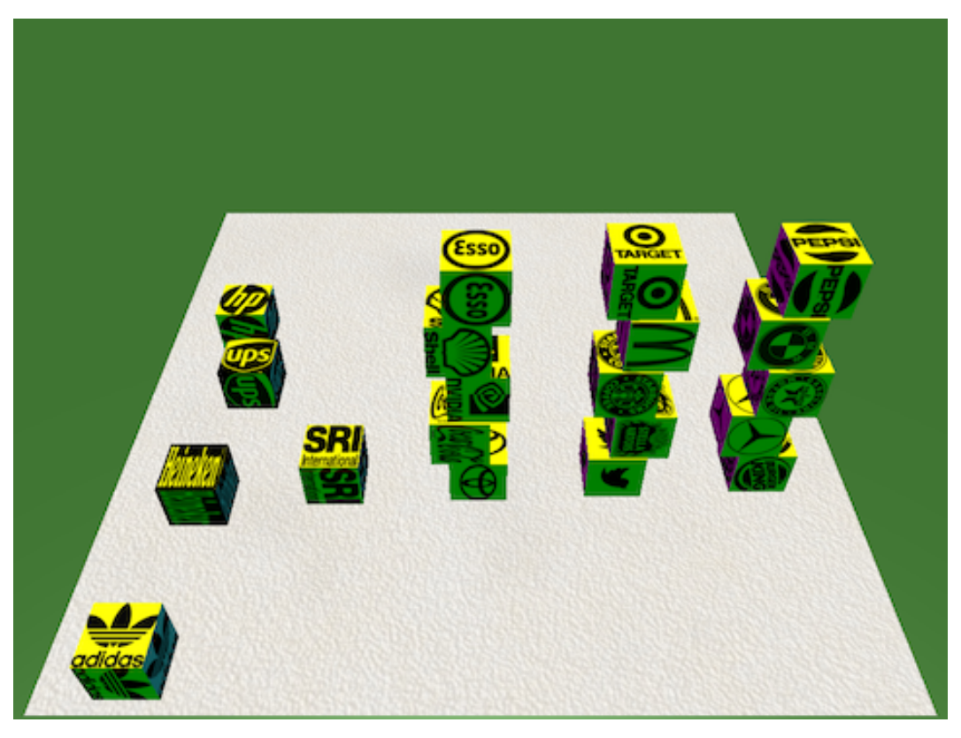
- L_2 in $\mathbb{R}^3$
- radians for angle

Data Collection

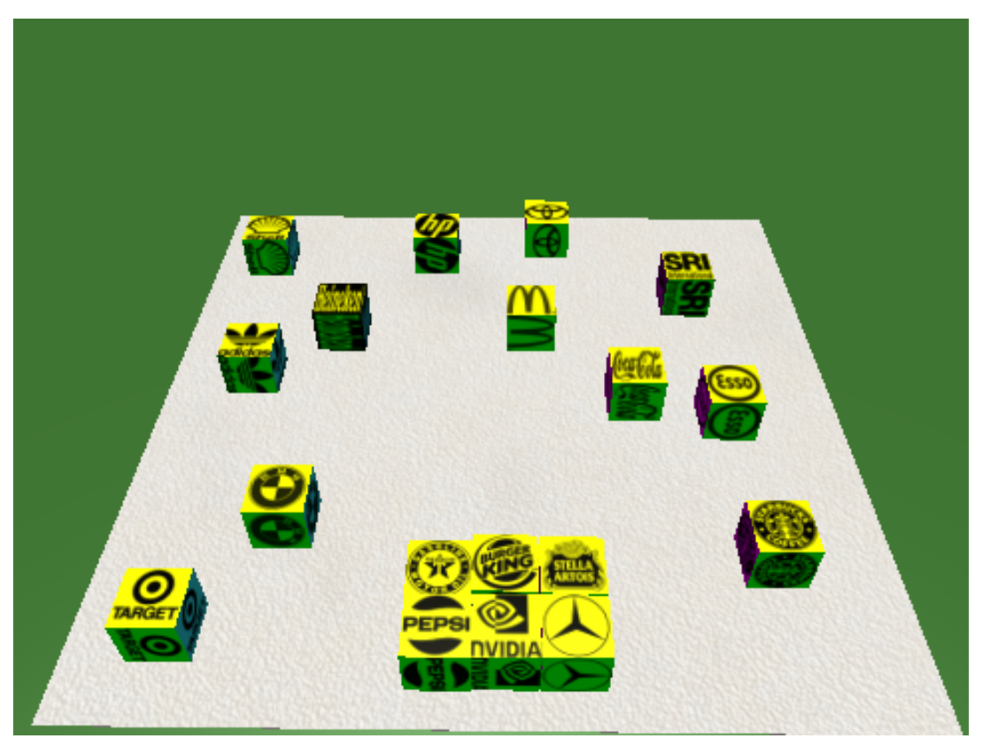
Nine annotations per action collected from Mechanical Turk. The linguistic difficulty of spatial reasoning varies dramatically.



McDonalds
… to the **right** of twitter with a small **space in between**.
… just to the **right** (**not touching**) twitter.



use SRI as the base of a **fourth** tower to the left and **equidistance** with the other tower



in the **emerging 3x3 grid** place texaco in the middle left

Simple relations

Difficult concepts

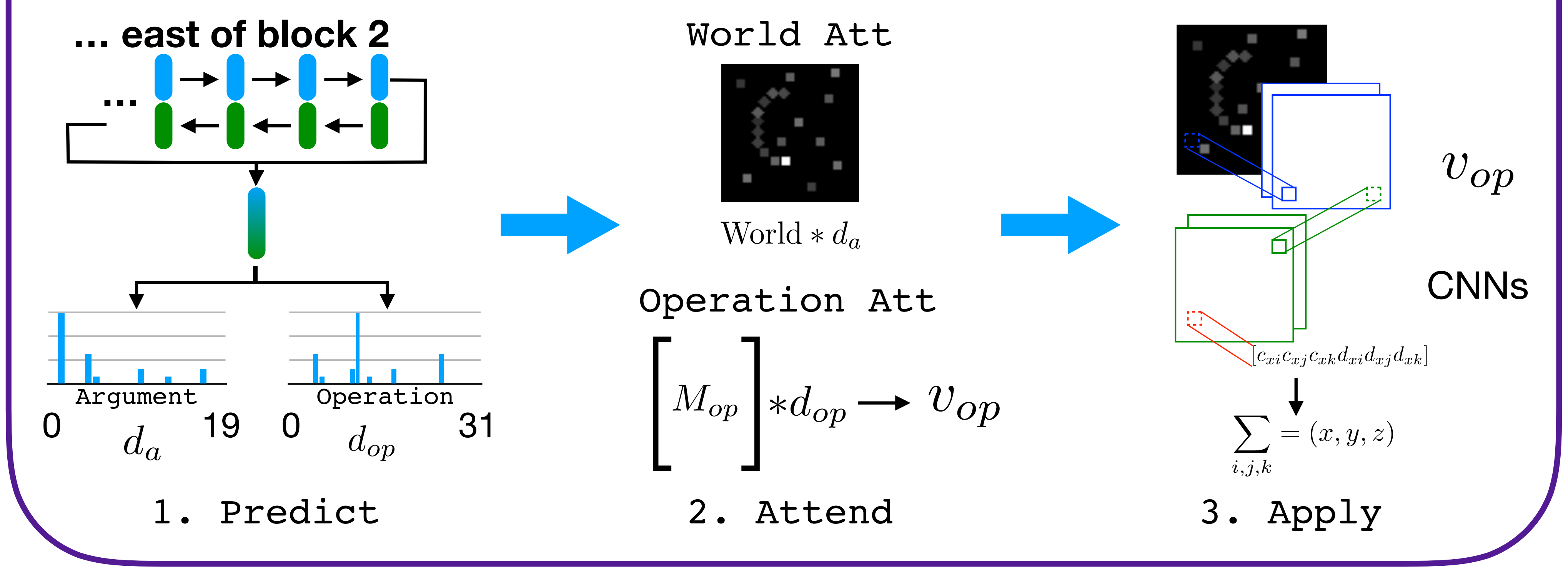
We introduce new concepts and complicate previous ones by having humans perform all actions in R³

Previous: left, up, right, directly, above, until, corner, top, down, below, bottom, slide, space, between, …

This work: degrees, rotate, clockwise, covering, 45, layer, mirror, arch, towers, equally, twist, balance, …

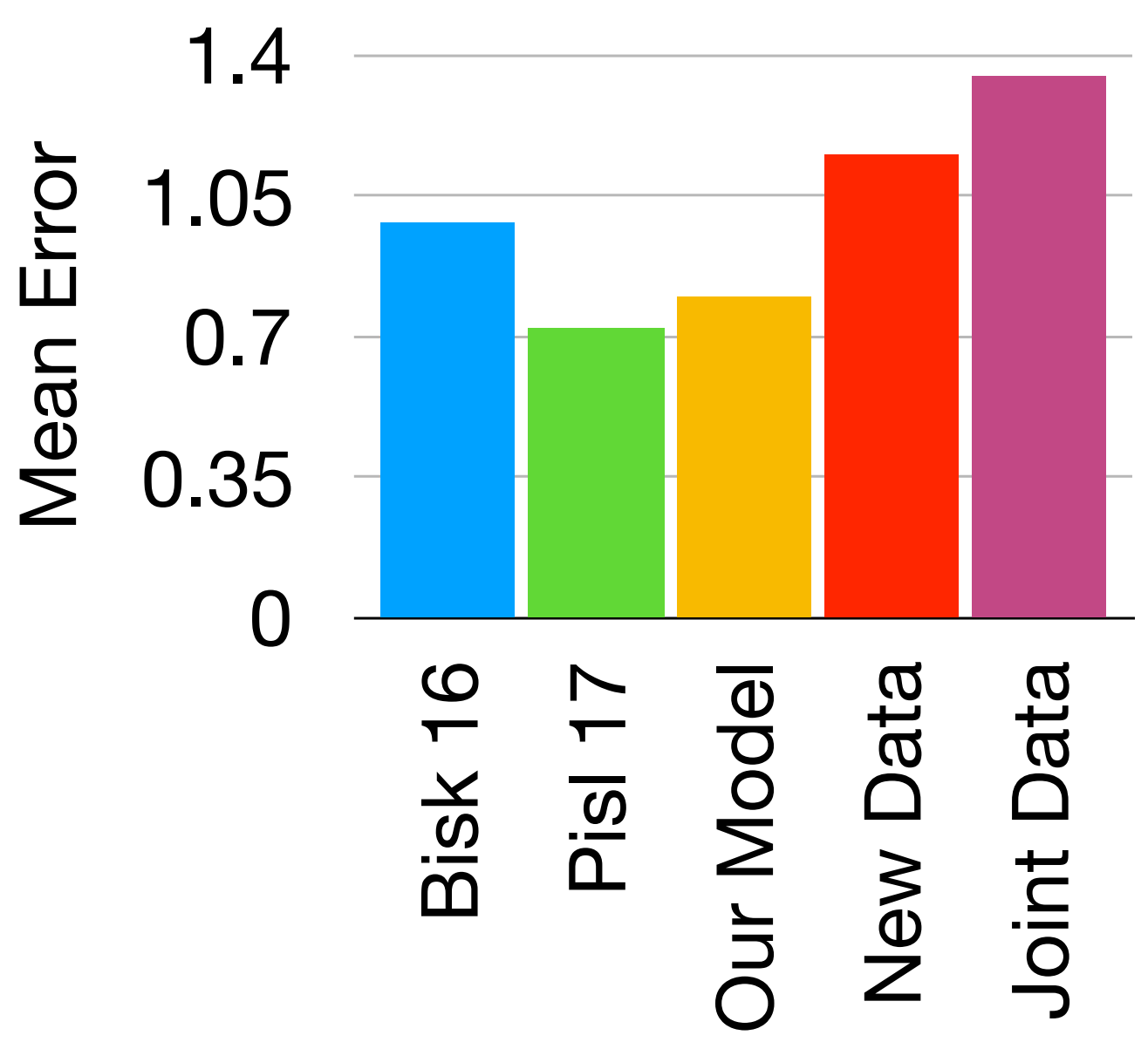
Modeling Operations as Embeddings

The model must both cluster the language into arguments and operations, while jointly learning those operations. Operations are randomly initialized 1x1 convolutions


… east of block 2
…
Argument
0
d_a
19
Operation
0
d_{op}
31
1. Predict
World Att
$\text{World} * d_a$
Operation Att
$\begin{bmatrix} M_{op} \end{bmatrix} * d_{op} \longrightarrow v_{op}$
2. Attend
v_{op}
CNNs
$[c_{xi} c_{xj} c_{xk} d_{xi} d_{xj} d_{xk}]$
$\sum_{i,j,k} = (x, y, z)$
3. Apply

Numbers

We report our average error in block-lengths, on the previous simplified data, our new data and the joint.



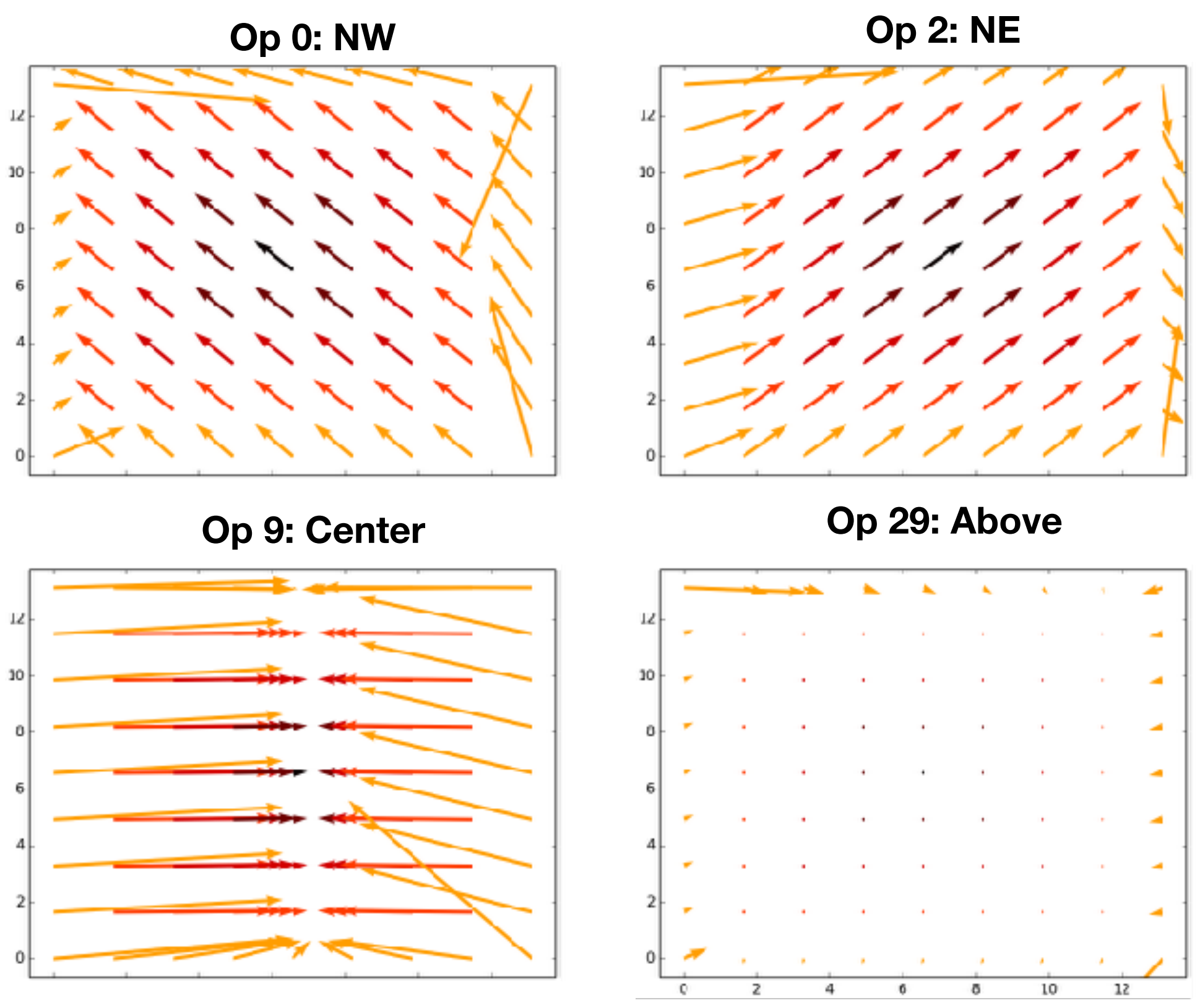
Data statistics

	Configs	Types	Tokens	Ave Len
Previous	100	1,281	258K	15.4
This	100	1,820	233K	18.0
Joint	200	2,299	491K	16.5

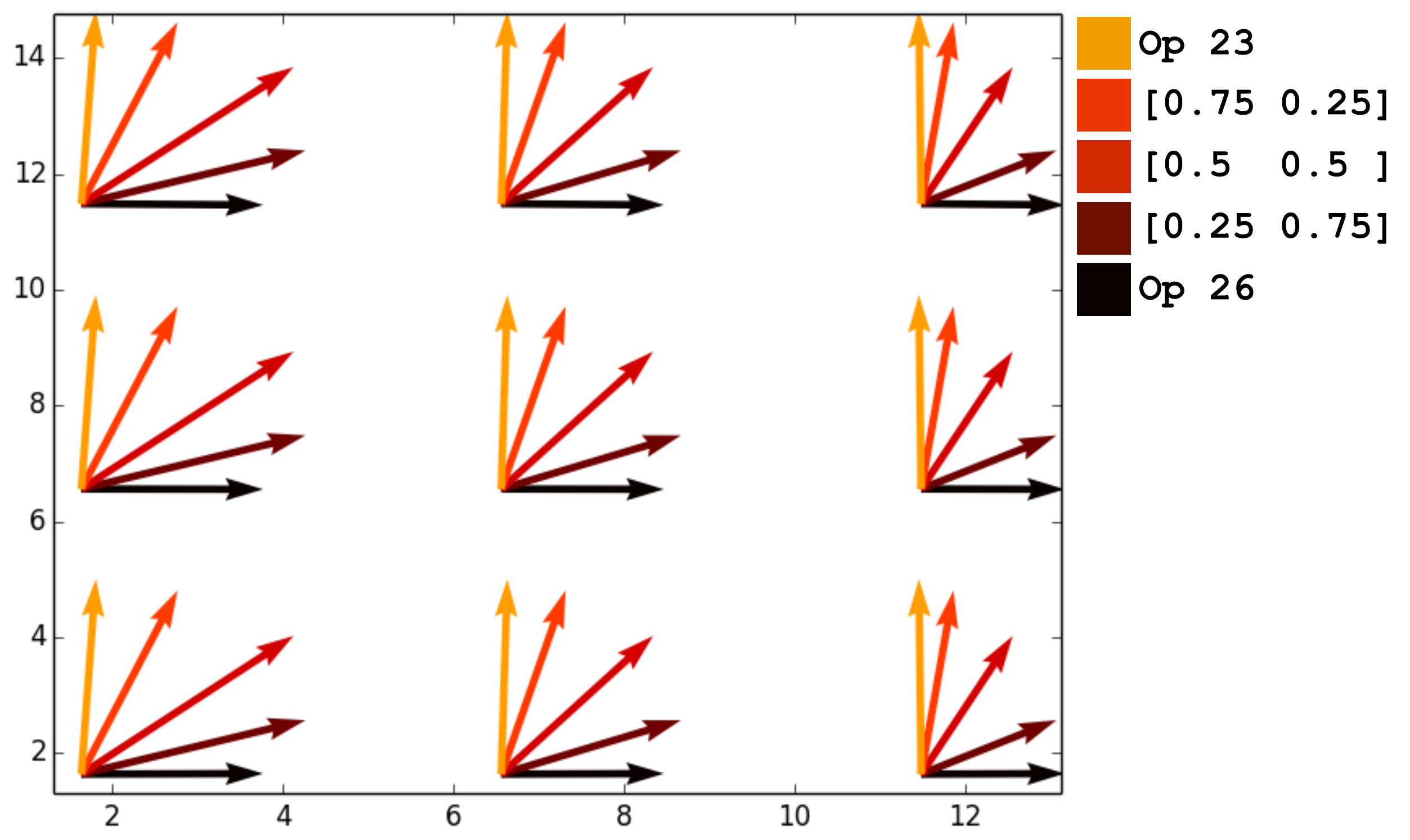
Visualizing Interpretable Operations

After training, each embedding M_{op} can be visualized by multiplying by a one-hot d_{op} and applying the convolution to a single block moved around the image. We visualize four embeddings here (all 32 are presented in the paper).

Replacing d_{op} with a distribution allows us to interpolate between operations.



Interpolating between Op23 (north) and Op26 (east). Note, that the angles and lengths of the vectors shift as a function of their location in the world, they are not absolute offsets. One can only go so far east on the eastern edge.


Op 23
[0.75 0.25]
[0.5 0.5]
[0.25 0.75]
Op 26