Energy-based Models are Zero-Shot Planners for Compositional Scene Rearrangement

Nikolaos Gkanatsios*, Ayush Jain*, Zhou Xian, Yunchu Zhang, Chris Atkeson, Katerina Fragkiadaki

Carnegie Mellon University

* Equal contribution

THE ROBOTICS INSTITUTE

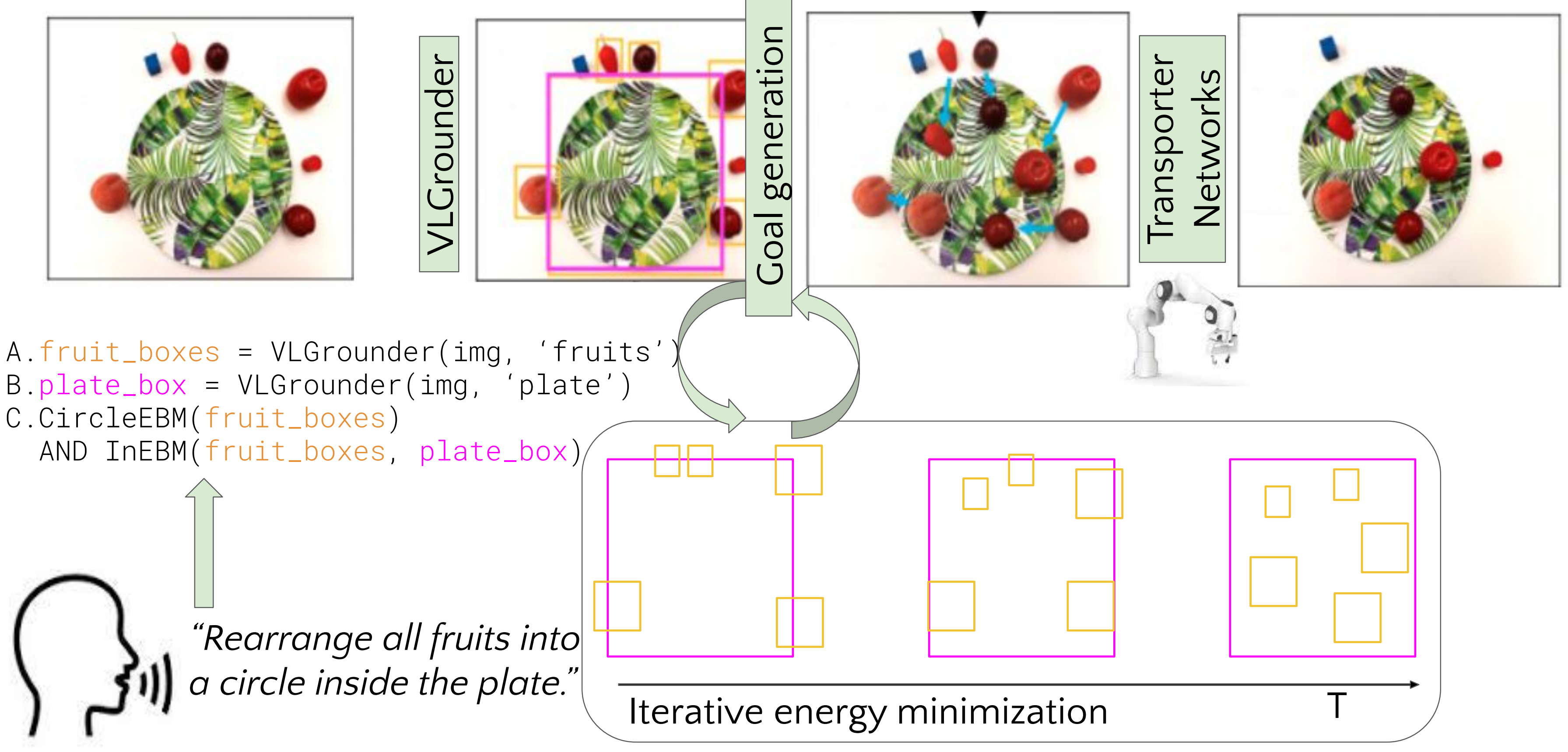
Introduction

If we teach a robot the concepts “***left/right/front of***”, can it generalize to “*place the apple in **front of** the duck, **left of** the avocado and **right of** the green bowl*”?

How can we build robots that:

- decompose complex instructions into familiar concepts,
- are robust to visual variations of the environment.

Scene Rearrangement via Energy Minimization


VLGrounder
Goal generation
Transporter Networks
“Rearrange all fruits into a circle inside the plate.”
Iterative energy minimization
T

```
A.fruit_boxes = VLGrounder(img, ‘fruits’)
B.plate_box = VLGrounder(img, ‘plate’)
C.CircleEBM(fruit_boxes)
  AND InEBM(fruit_boxes, plate_box)
```

SREM is a modular framework for scene rearrangement:

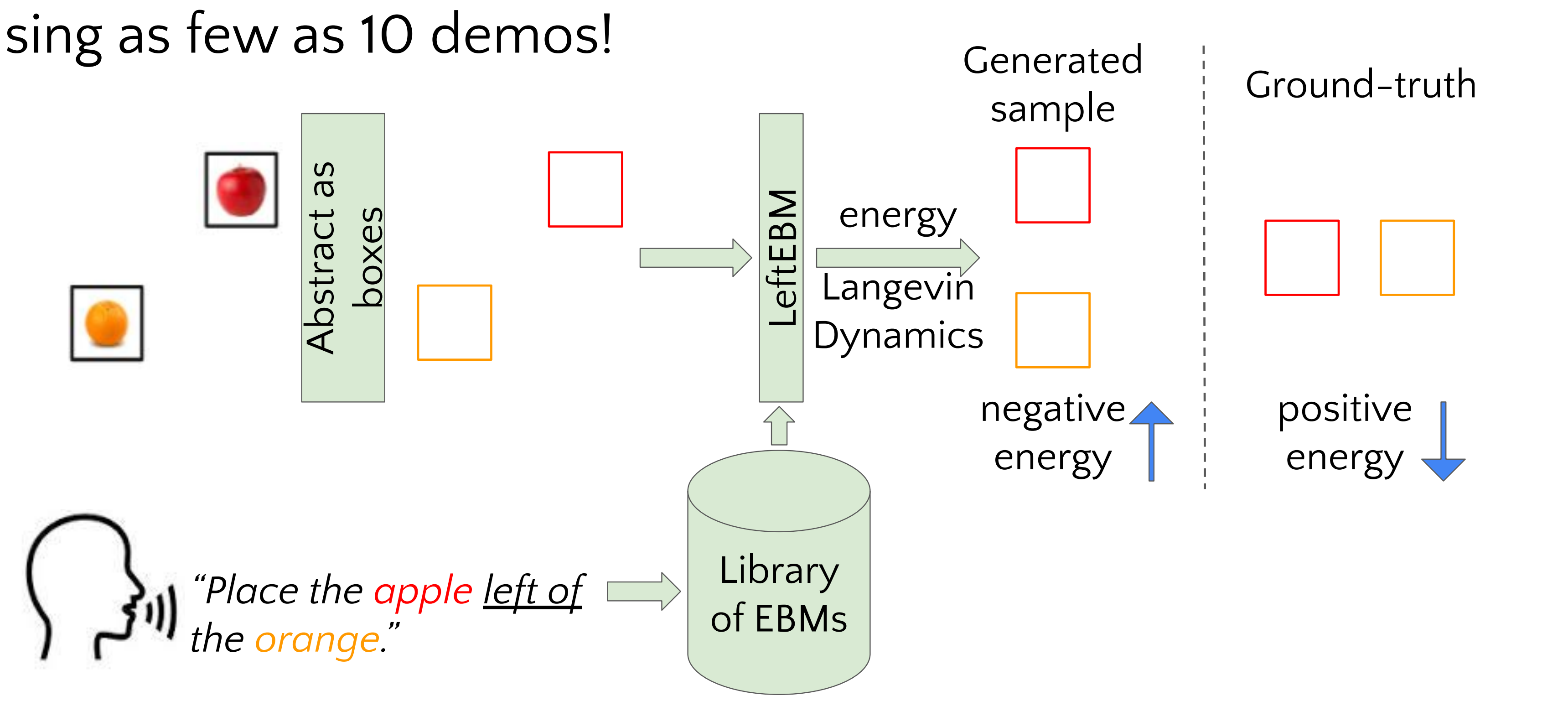
- The instruction is first parsed into a list of calls to a vision-language grounder and energy-based models.
- The grounder locates the objects mentioned in the instruction.
- The energy-based models (EBMs) infer a goal configuration for all mentioned objects jointly.
- A policy moves the objects to the predicted goal locations.

Experimental setup:

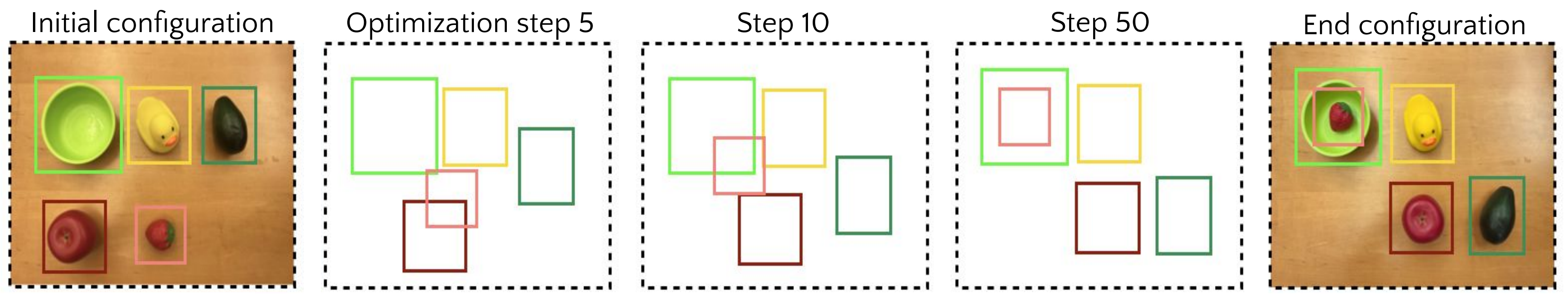
- Training on individual concepts (e.g., “*make a line of fruits*”)
- Test of compositional instructions (e.g., “*put the red bowl to the right of the yellow cube and above the blue cylinder*”).

EBMs are trained on atomic concepts

Using as few as 10 demos!



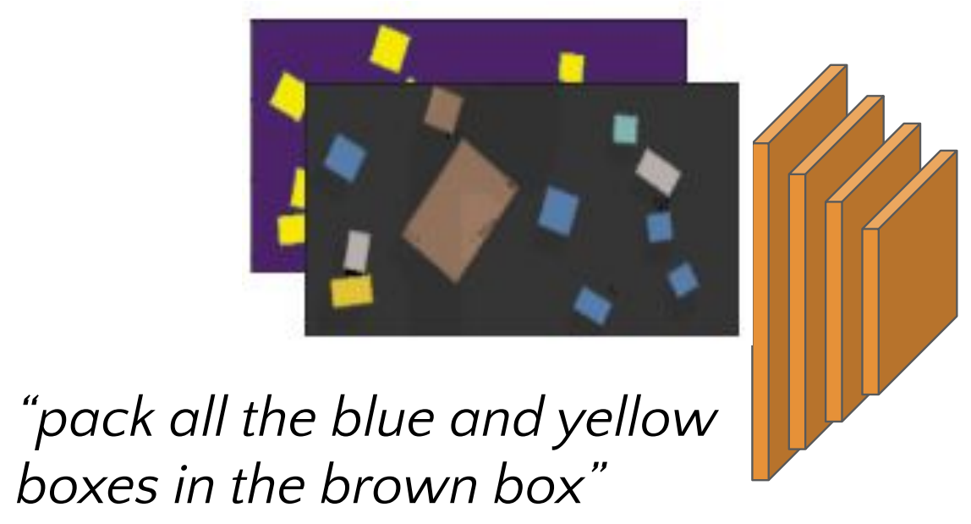
EBMs are composable!



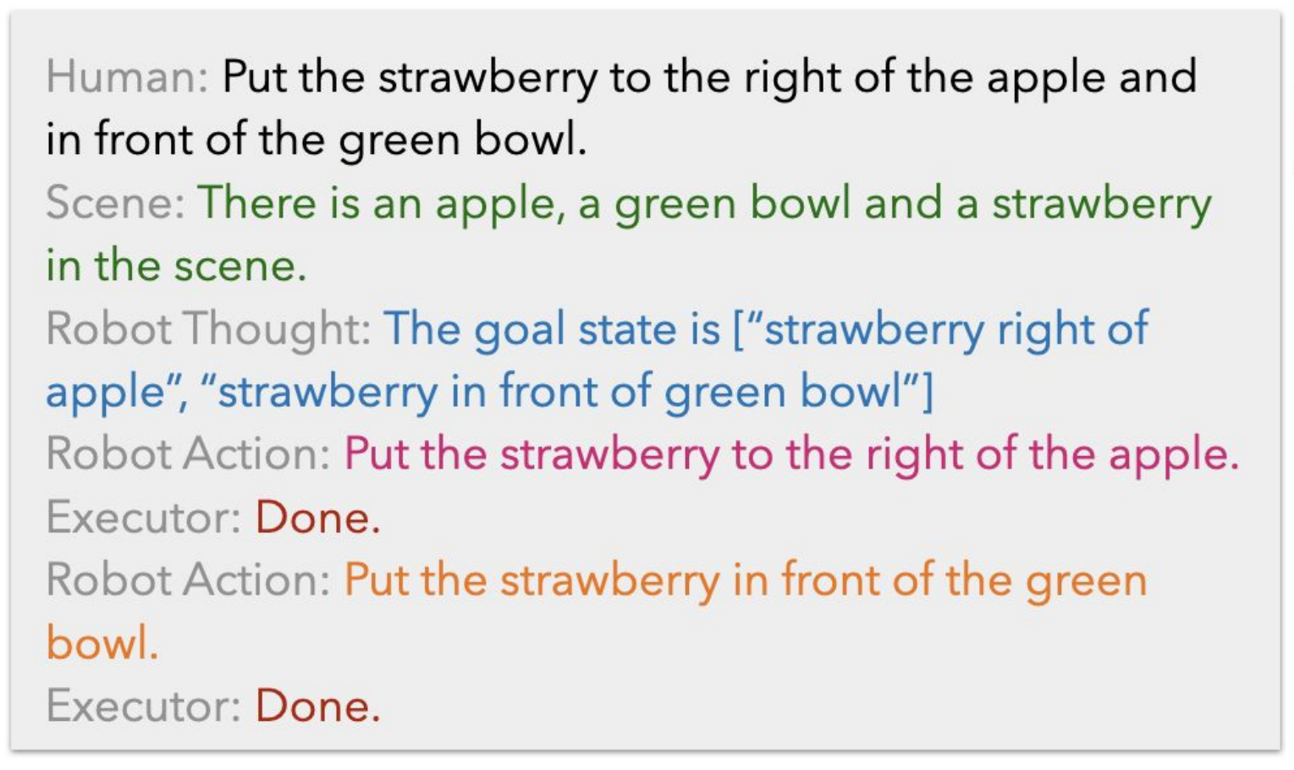
E = FrontEBM([], []) + LeftEBM([], []) + RightEBM([], []) + InsideEBM([], []) + FrontEBM([], [])

“*Place the apple in front of the duck, left of the avocado and right of the green bowl. Put the strawberry inside the green bowl. Move the avocado in front of the duck.*”

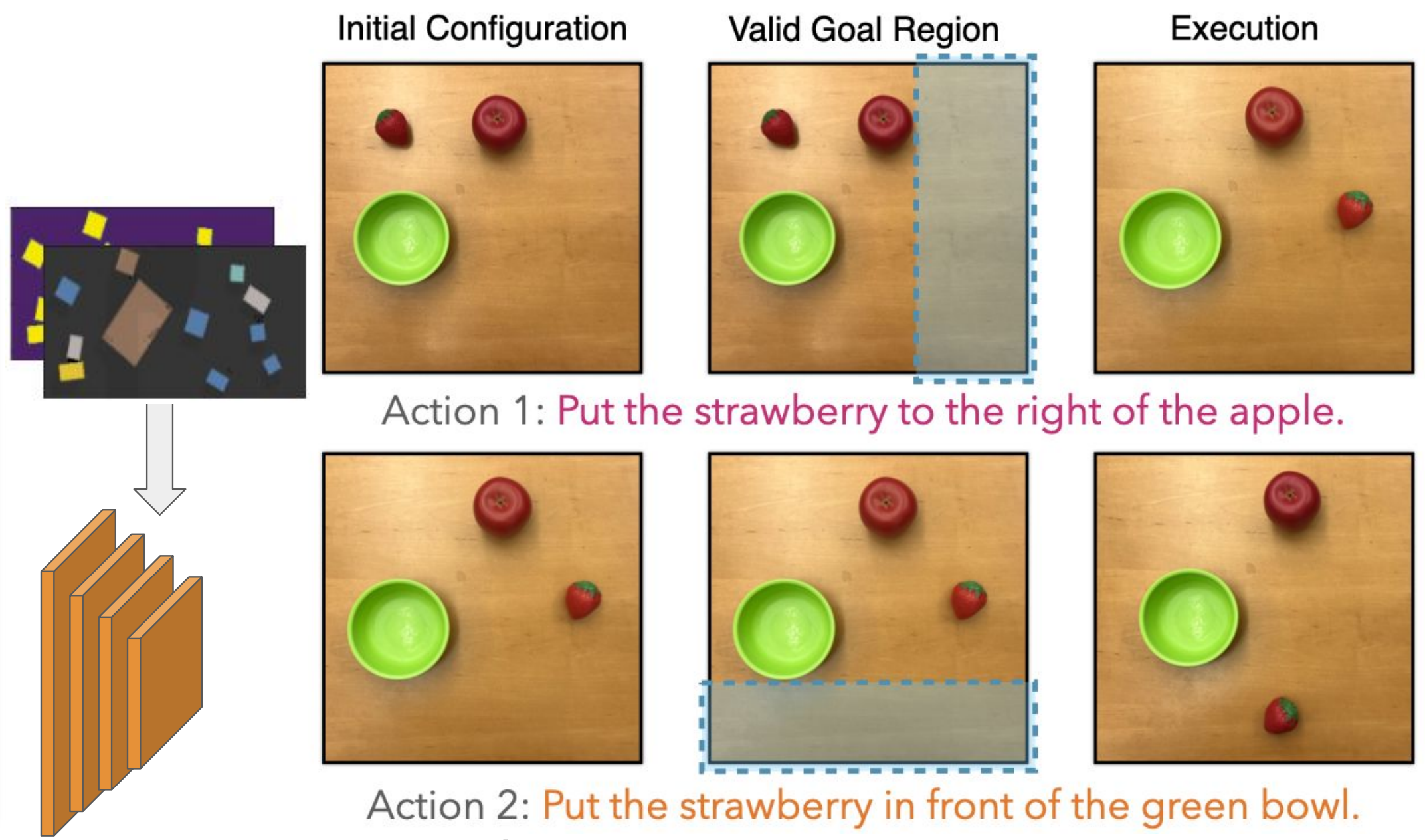
Why is joint optimization important?



End-to-end language-to-action mapping does not generalize to longer instructions.



Human: Put the strawberry to the right of the apple and in front of the green bowl.
Scene: There is an apple, a green bowl and a strawberry in the scene.
Robot Thought: The goal state is [“strawberry right of apple”, “strawberry in front of green bowl”]
Robot Action: Put the strawberry to the right of the apple.
Executor: Done.
Robot Action: Put the strawberry in front of the green bowl.
Executor: Done.



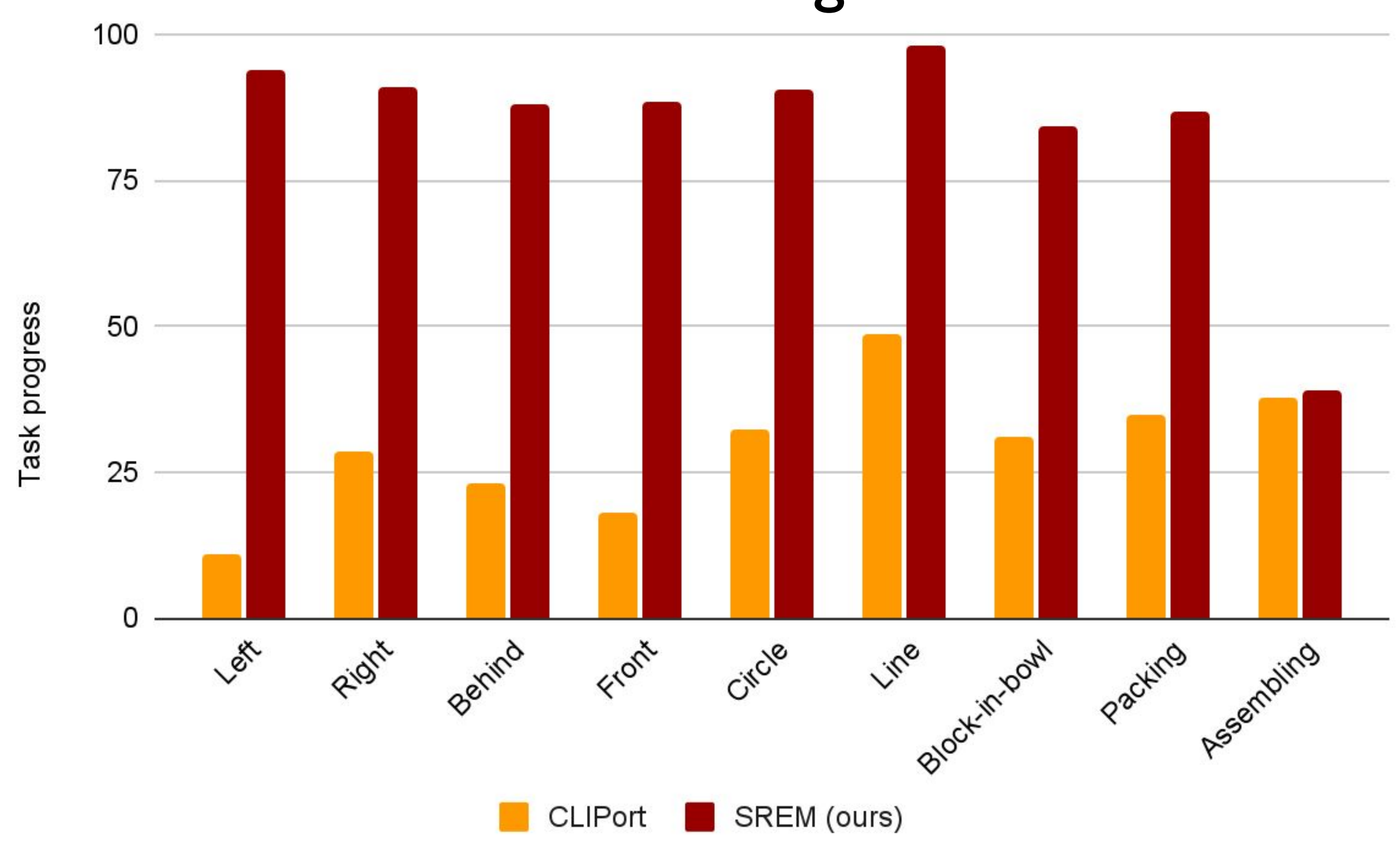
Action 1: Put the strawberry to the right of the apple.

Action 2: Put the strawberry in front of the green bowl.

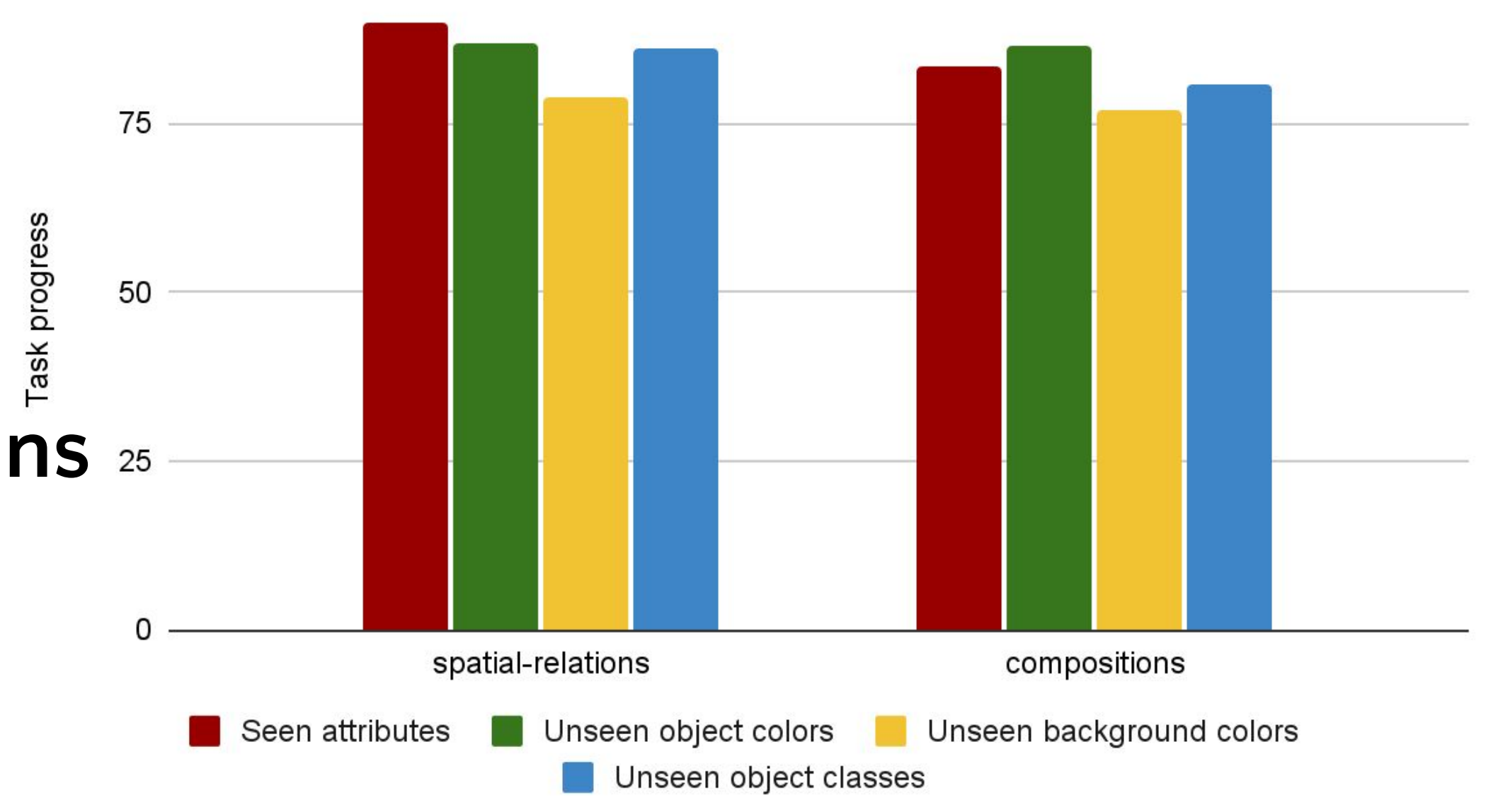
Language subgoals using LLMs: sequential execution does not guarantee correctness!

Results

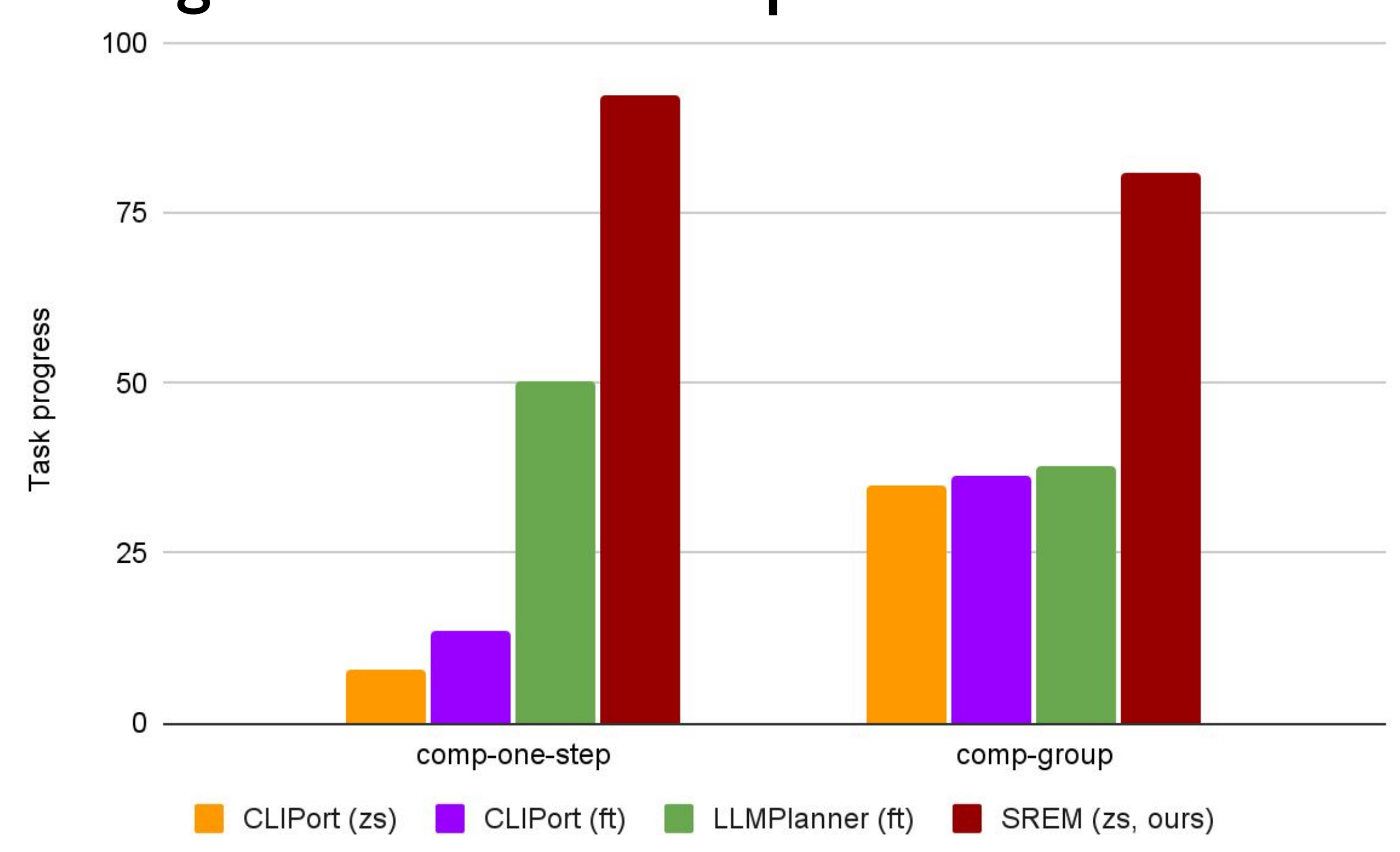
SREM better models the training tasks



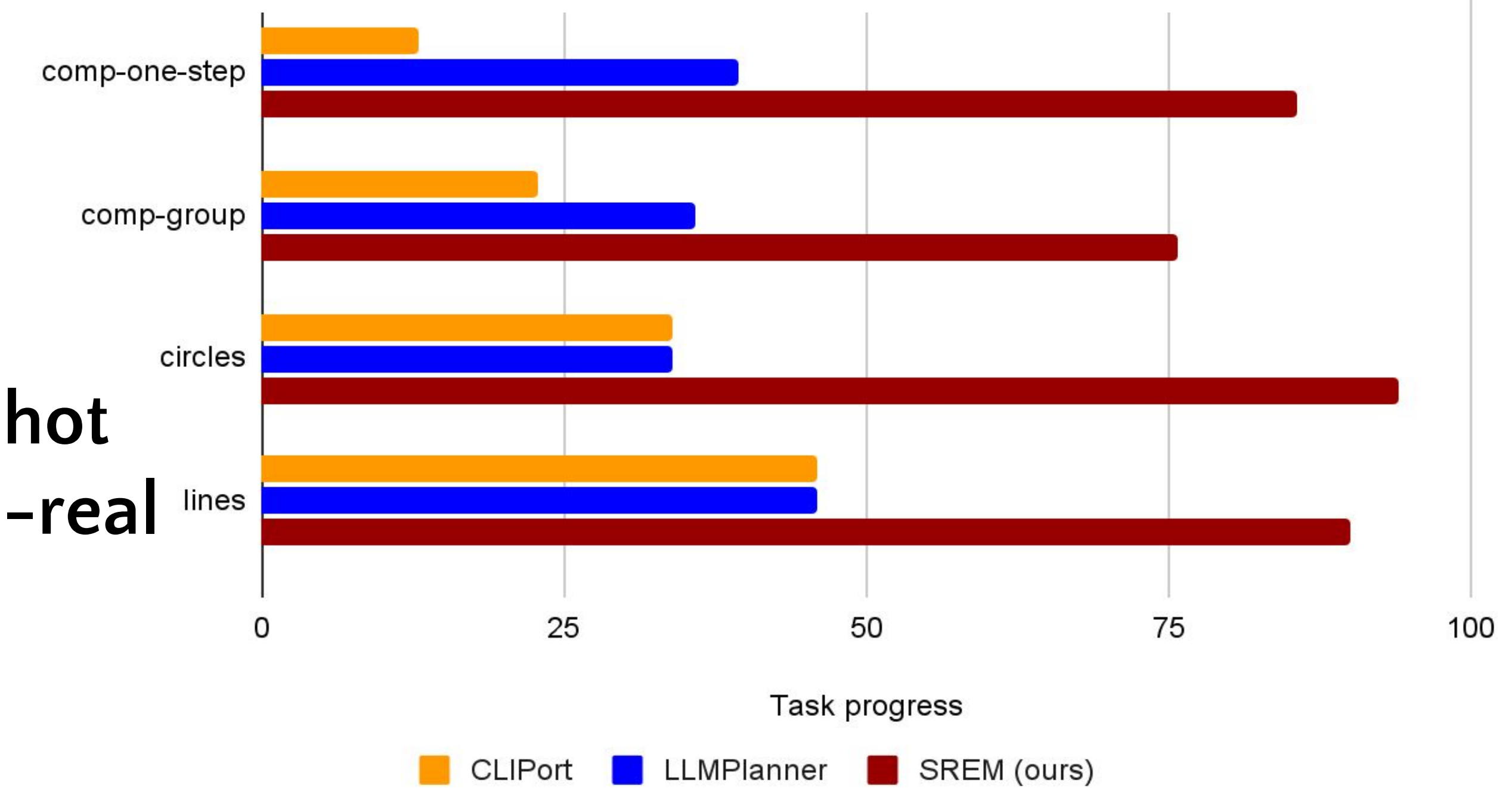
Robust to visual variations



Zero-shot generalizes to compositional instructions



Zero-shot sim-to-real


comp-one-step
comp-group
circles
lines
Task progress
CLIPort
LLMPlanner
SREM (ours)