SORNet: Spatial Object-Centric Representations for Sequential Manipulation

Wentao Yuan[1]   Chris Paxton[2]   Karthik Desingh[1]   Dieter Fox[1,2]
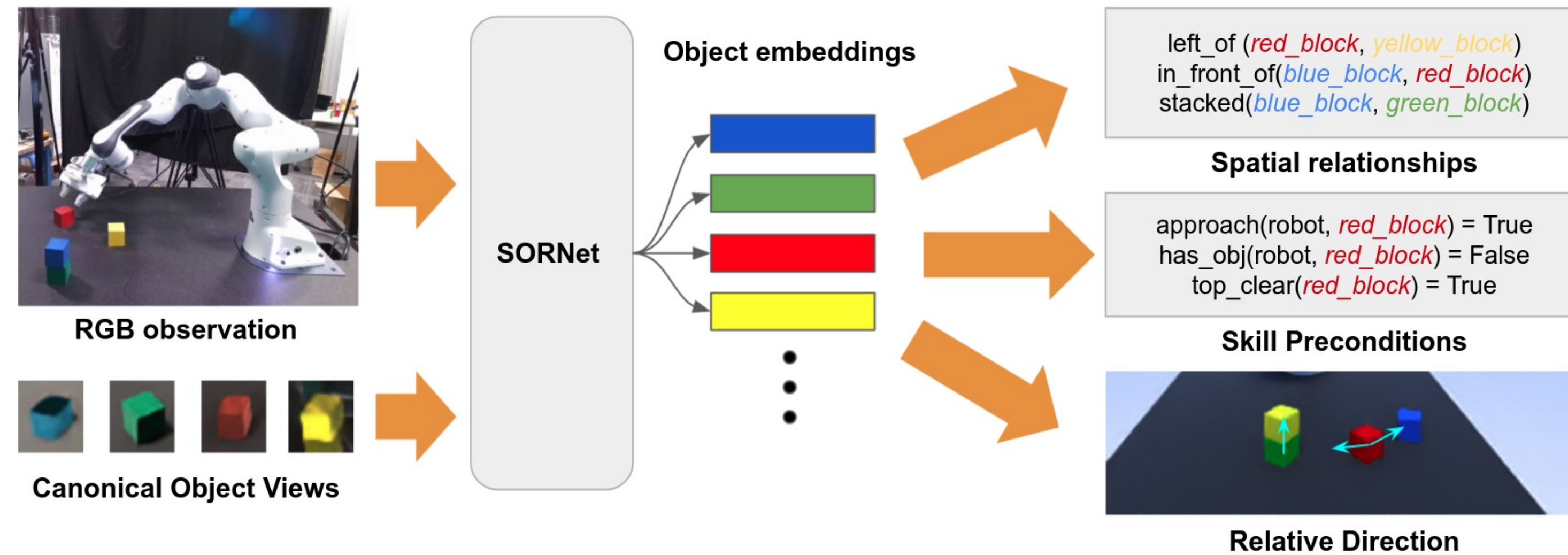
[1]University of Washington   [2]NVIDIA

Overview

We propose **SORNet**: **S**patial **O**bject-centric **R**epresentation **Net**work to learn object-centric embeddings that encode spatial relationships

RGB observation

Canonical Object Views

Object embeddings

left_of(*red_block*, *yellow_block*)
in_front_of(*blue_block*, *red_block*)
stacked(*blue_block*, *green_block*)

Spatial relationships

approach(robot, *red_block*) = True
has_obj(robot, *red_block*) = False
top_clear(*red_block*) = True

Skill Preconditions

Relative Direction

- SORNet is trained only on classification of **logical** predicates, but captures **continuous** spatial relationships;
- SORNet generalizes **zero-shot** to scenes with unseen objects and different number of objects.

Training Objective (logical predicates)
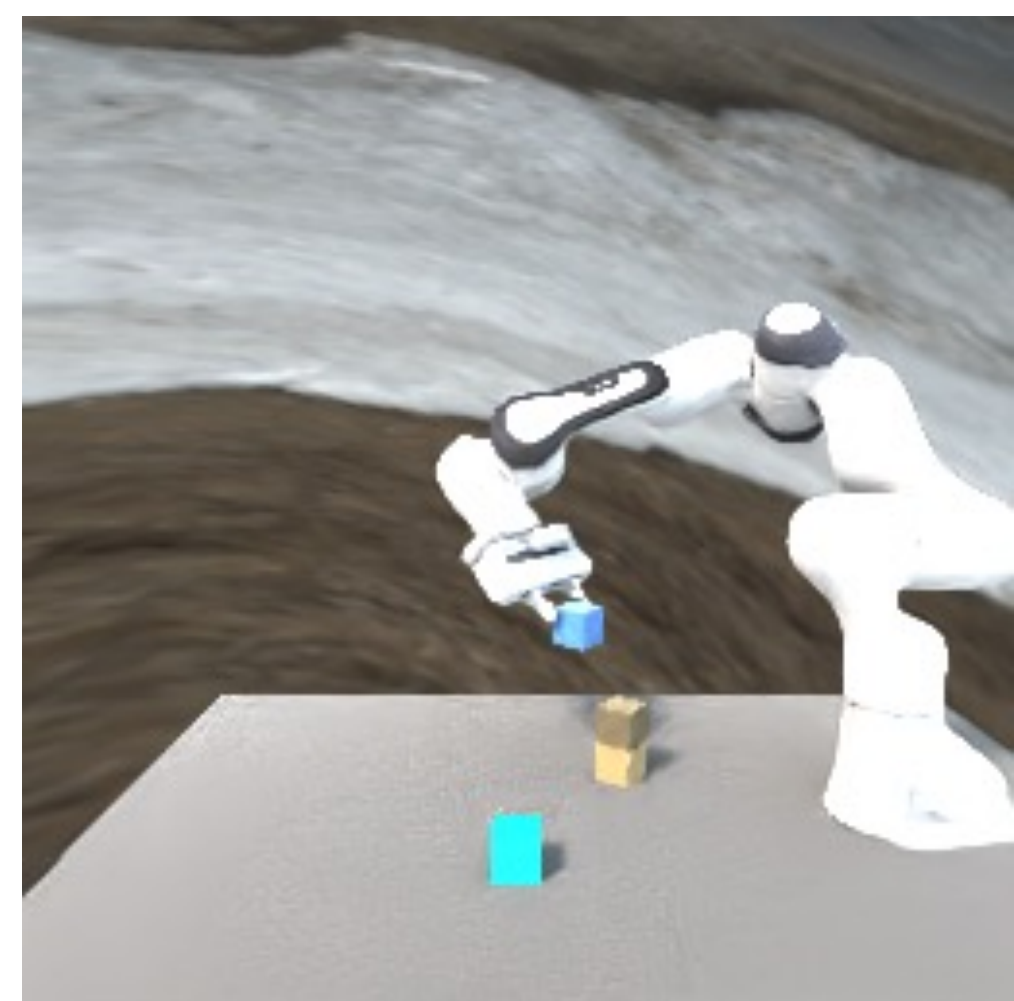
has_obj(robot, azure_block)
top_is_clear(beige_block)
on_surface(cyan_block, right)
stacked(brown_block, beige_block)
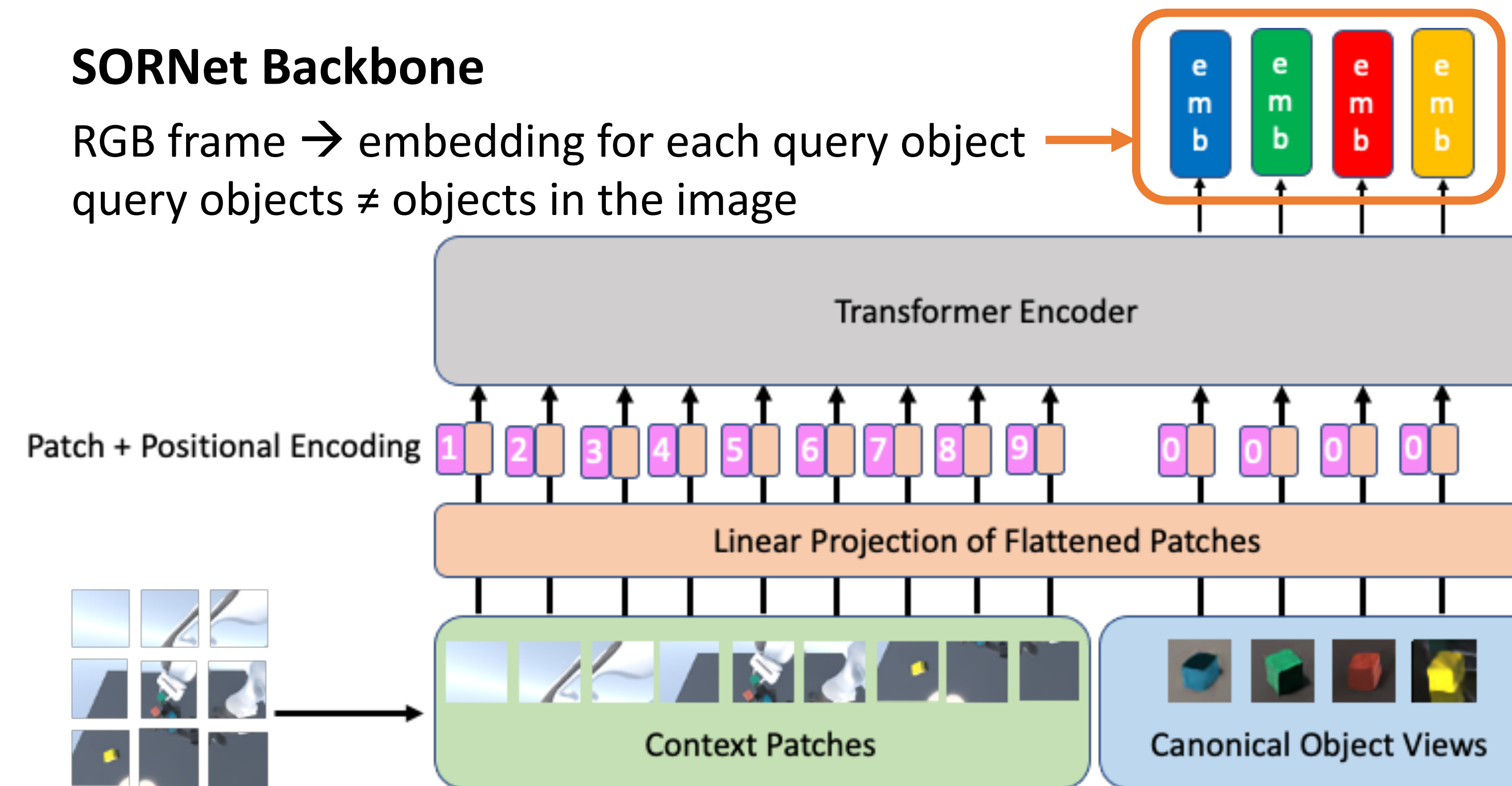
Inference Objective (continuous direction)
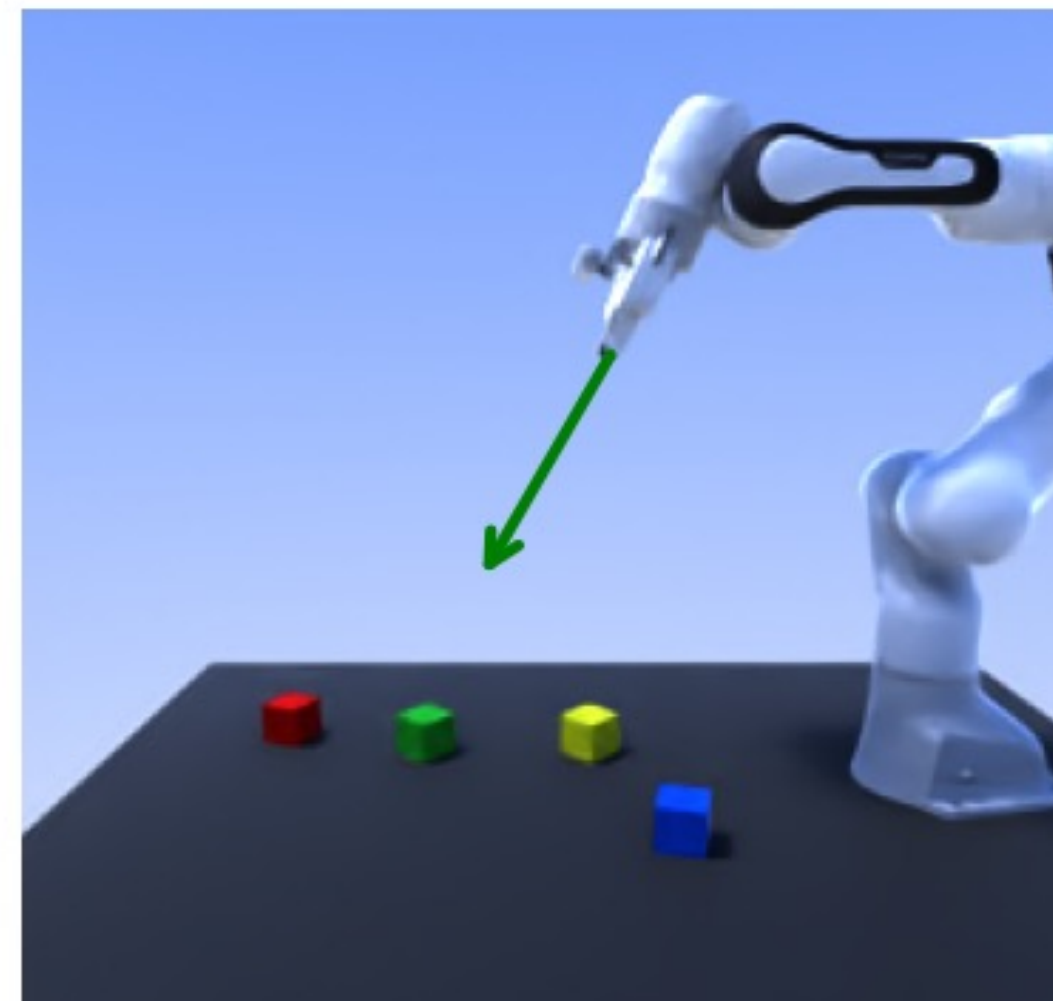
Arrow = predicted unit direction scaled by predicted distance from end effector to green block (unseen during training)
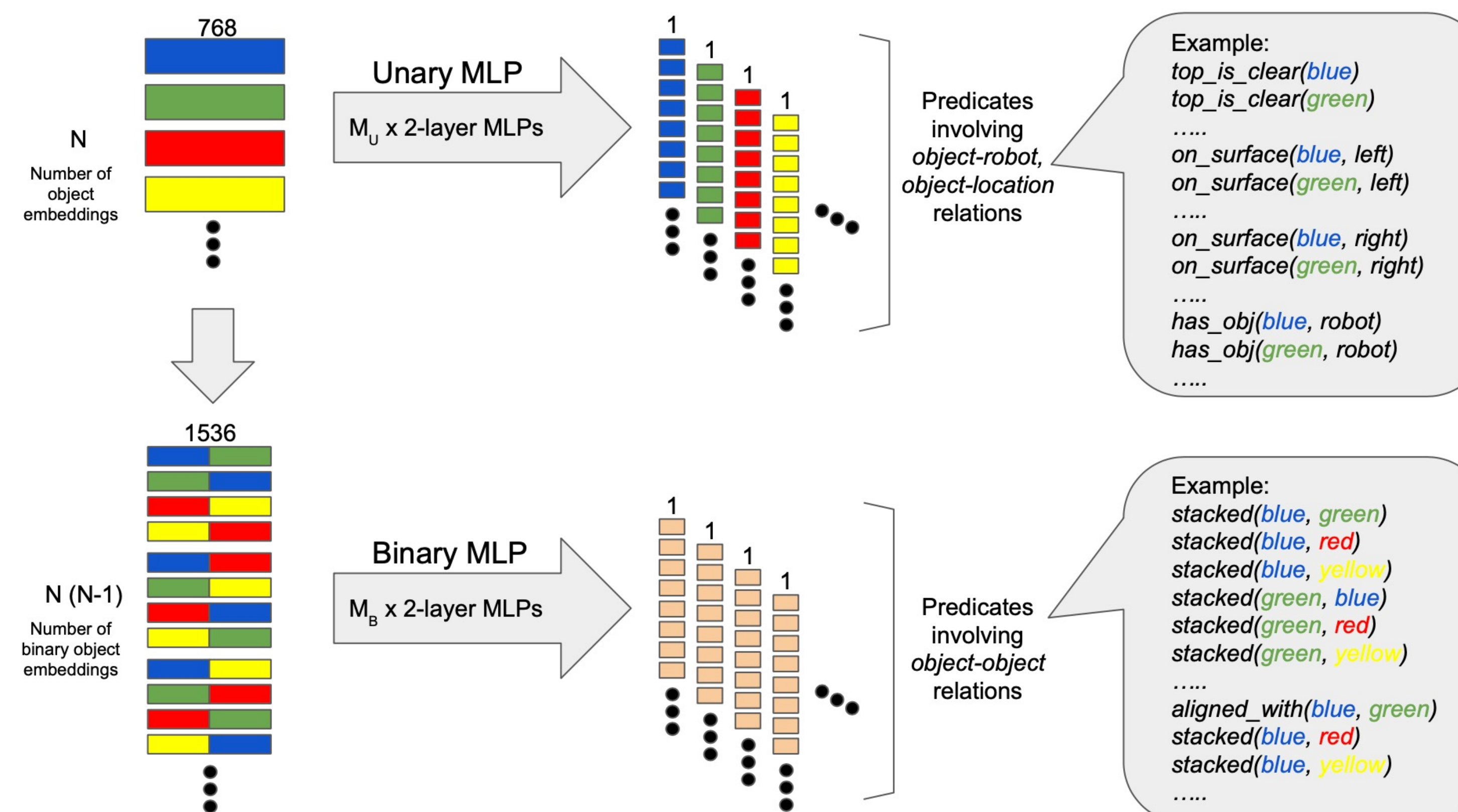
Approach

SORNet Backbone

RGB frame → embedding for each query object
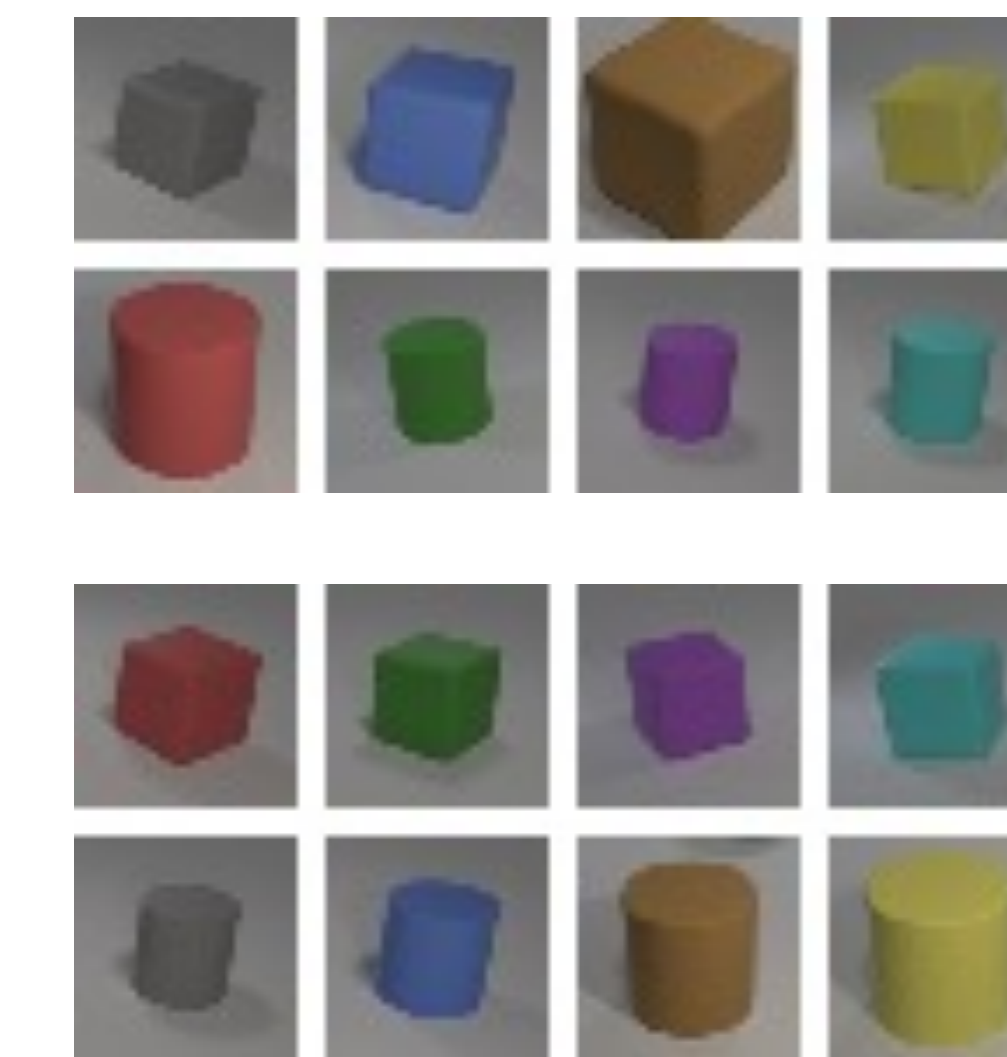query objects ≠ objects in the image

Transformer Encoder

Patch + Positional Encoding   1 2 3 4 5 6 7 8 9   0 0 0 0

Linear Projection of Flattened Patches

Context Patches   Canonical Object Views

Predicate Classifier

Number of outputs changes adaptively with number of input embeddings

768

N
Number of object embeddings

Unary MLP
M_U x 2-layer MLPs

Predicates involving *object-robot*, *object-location* relations

Example:
top_is_clear(blue)
top_is_clear(green)
.....
on_surface(blue, left)
on_surface(green, left)
.....
on_surface(blue, right)
on_surface(green, right)
.....
has_obj(blue, robot)
has_obj(green, robot)
.....

1536

N (N-1)
Number of binary object embeddings

Binary MLP
M_B x 2-layer MLPs

Predicates involving *object-object* relations

Example:
stacked(blue, green)
stacked(blue, red)
stacked(blue, yellow)
stacked(green, blue)
stacked(green, red)
stacked(green, yellow)
.....
aligned_with(blue, green)
stacked(blue, red)
stacked(blue, yellow)
.....

Results

Training (Condition A)
- Cubes are **gray**, **blue**, **brown**, or **yellow**
- Cylinders are **red**, **green**, **purple**, or **cyan**
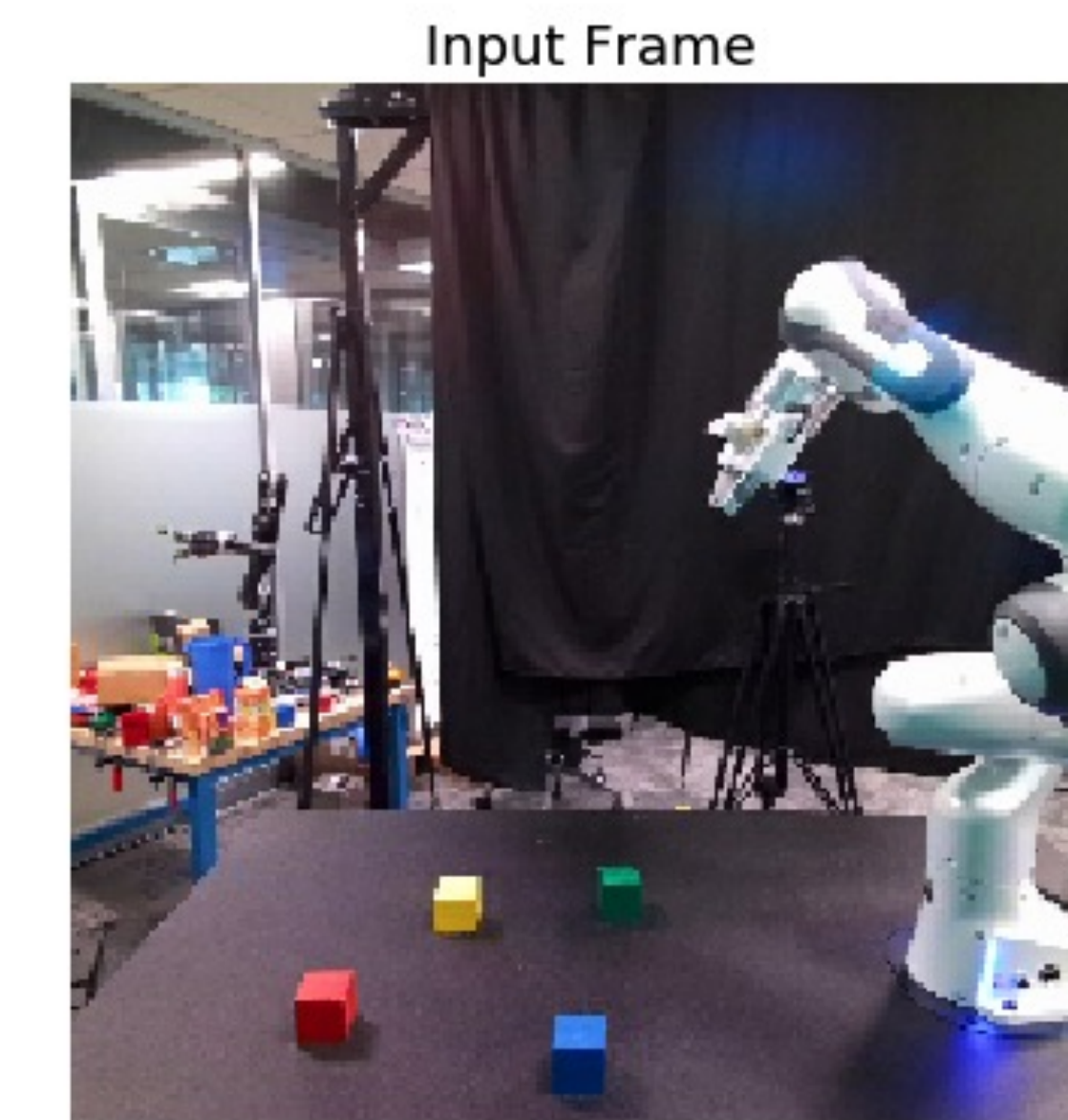- Spheres can have any color

Testing (Condition B)
- Cubes are **red**, **green**, **purple**, or **cyan**
- Cylinders are **gray**, **blue**, **brown**, or **yellow**
- Spheres can have any color

Zero-shot Spatial Relationship Prediction Accuracy

	MDETR [29]	SORNet (ours)
ValA Accuracy	84.950	90.909
ValB Accuracy	59.627	89.403

Real-world Open-loop Planning

Input Frame

Predicates Predicted
on_surface(red_block, left)
on_surface(blue_block, left)
on_surface(green_block, right)
on_surface(red_block, far)
on_surface(yellow_block, center)
top_is_clear(red_block)
top_is_clear(green_block)
top_is_clear(blue_block)
top_is_clear(yellow_block)

Goal Conditions
has_anything(robot) = False
on_surface(red_block, right) = True

Plan Generated
approach_obj(red_block)
grasp_obj(red_block)
lift_obj_from_tabletop(red_block)
place_on_right(red_block)