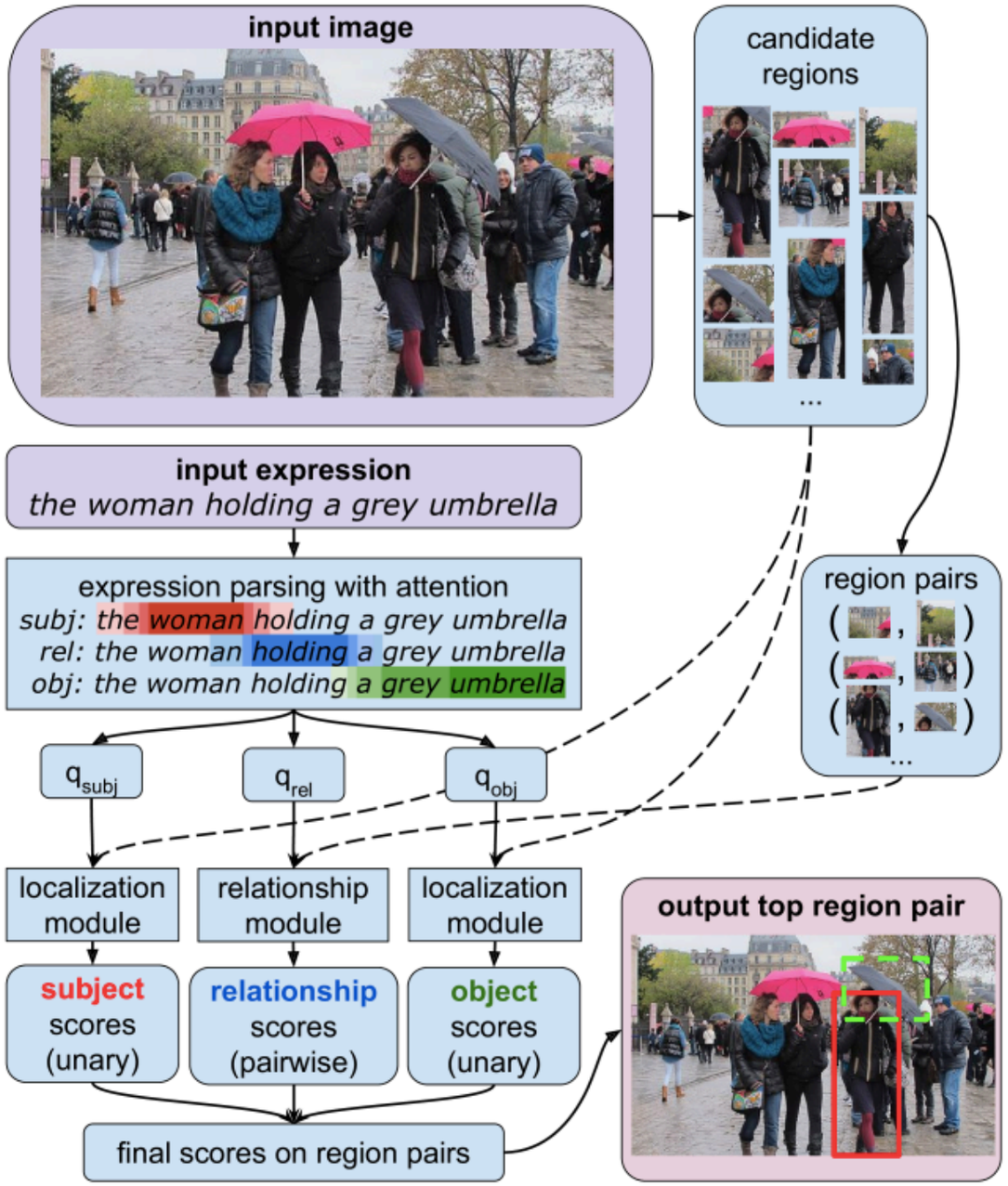
Embodied Language Grounding

Katerina Fragkiadaki

Carnegie Mellon University

What does it mean to comprehend natural language?

What does it mean to comprehend natural language?



Localize the referrents it mentions

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

What does it mean to comprehend natural language?

Generate the image it describes

Probabilistic Neural Programmed Networks for Scene Generation, Deng et al.

What does it mean to comprehend natural language?

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

Caption an image

Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Li,Fei fei

What does it mean to comprehend natural language?

- After wading barefoot in the lake, Erik used his shirt to dry his feet.
- After wading barefoot in the lake, Erik used his glasses to dry his feet.

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

Embodied Cognition

- Words and phrases are indexes to objects in the world or to prototypical symbols of those objects
- We derive affordance from those objects
- The derived affordances constrain the way ideas can be coherently combined

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

What does it mean to comprehend natural language?

- The bowl inside the cube
- The cube inside the bowl

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

Simulation Semantics

- Words and phrases are indexes to objects in the world or to prototypical symbols of those objects
- We derive affordance from those objects
- The derived affordances constrain the way ideas can be coherently combined

To act upon it, and infer its affordability

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

Simulation Semantics

Many philosophers agree and support such simulation semantics, no computational model for language grounding

Visually grounded interaction and language, NIPS workshop 2018

Embodiment, simulation and meaning, Bergen, How reading comprehension is embodied and why that matters , Glenberg, Grounding language in action, Glenberg and Kaschak, Grounding Meaning in Affordances, Glenberg

Language grounding to visual cues

Remains disconnected from affordances

2D boxes or 2D CNN features do not have any affordability attached
Are themselves ungrounded

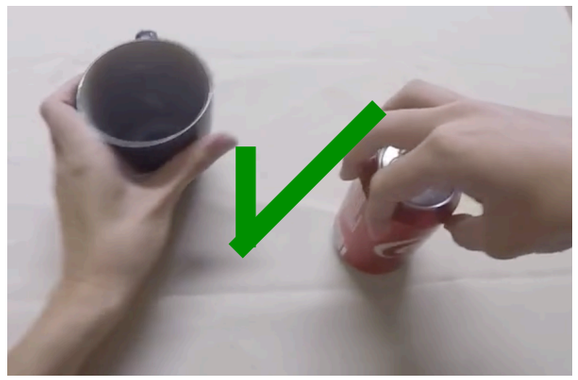
- The bowl inside the cube
- The cube inside the bowl

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

Reward learning using natural language

Given a NL utterance, learn a visual detector

“Can is to the right of the mug”

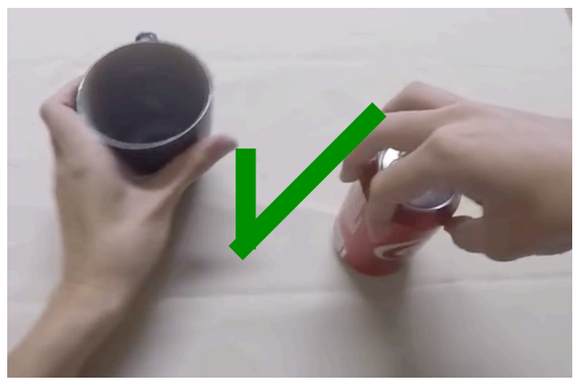


Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

Reward learning using natural language

Given a NL utterance, learn a visual detector

“C*an is to the right of the mug”*



Use the learned visual detector to guide policy learning for achieving the NL described goal

It did not really worked, the reward detector could not effectively generalize across camera placements

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

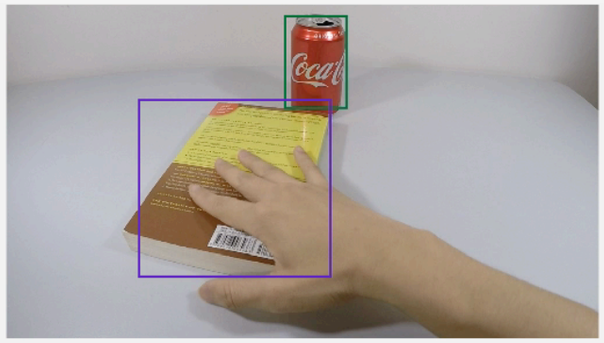
3. Language Grounding from Narrated Demonstrations

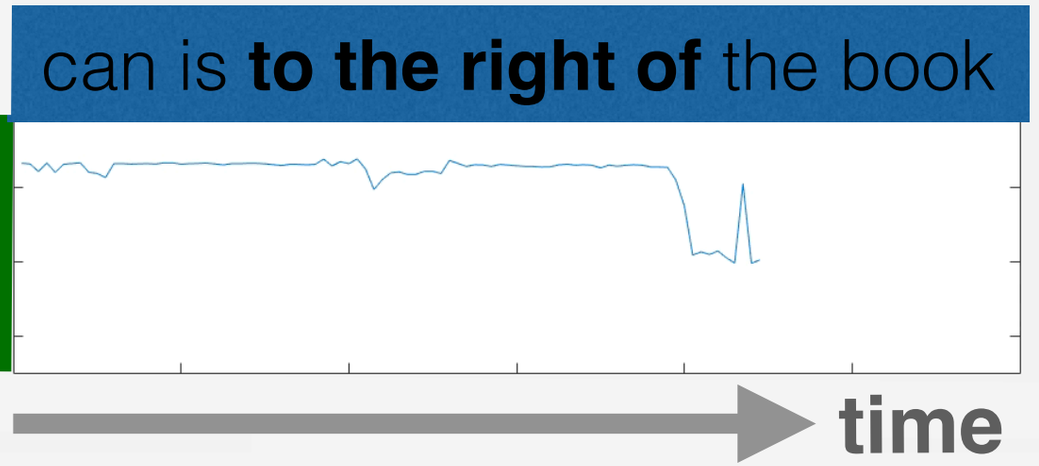
CVPR'18

"Can is to the right of the mug"

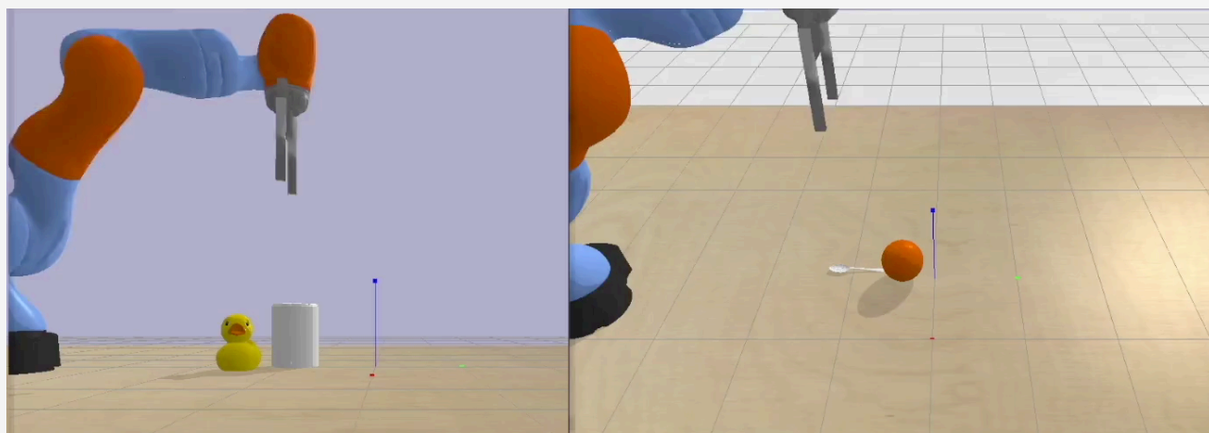
reward detector

Learned reward detector



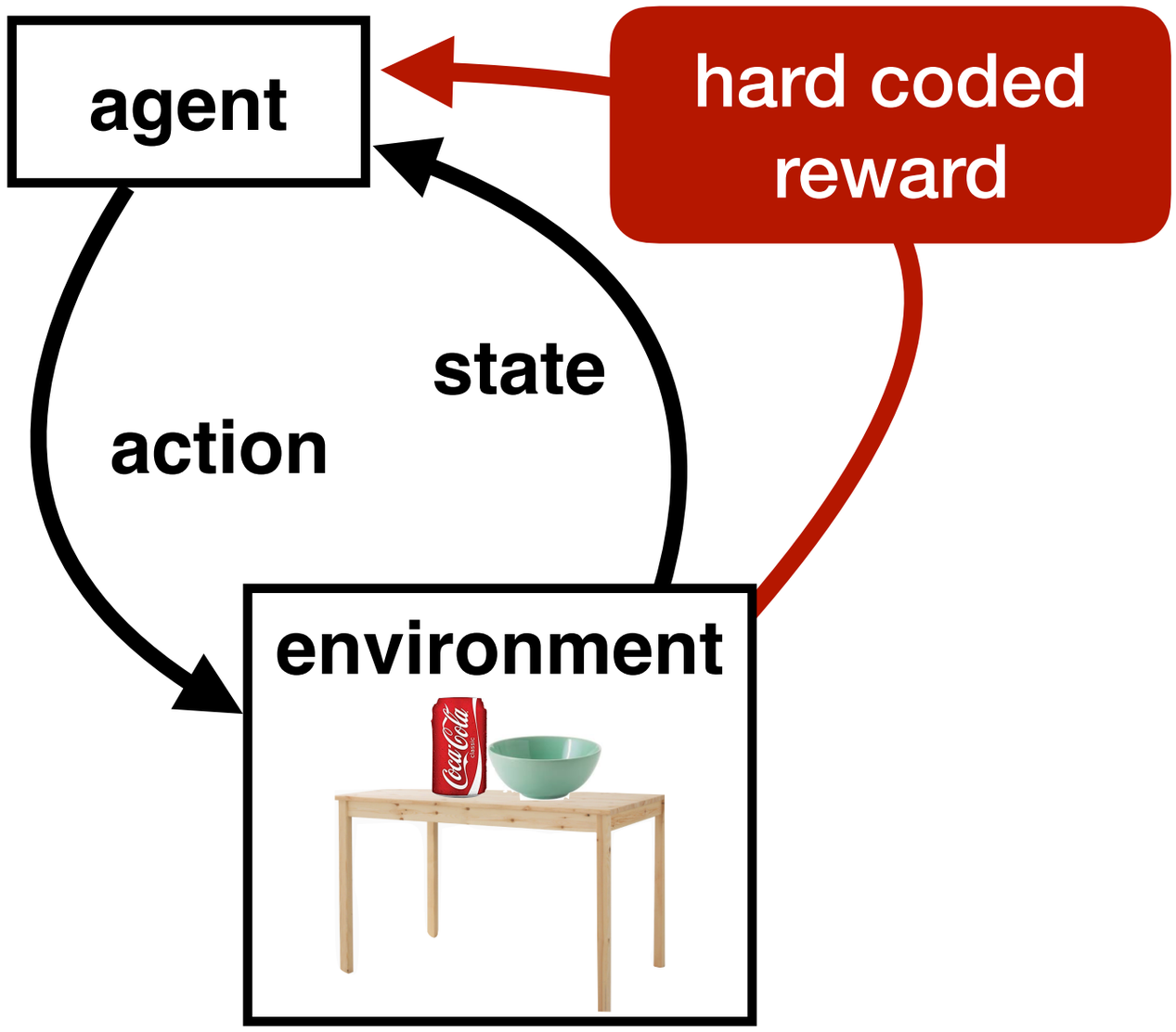


Learned policy

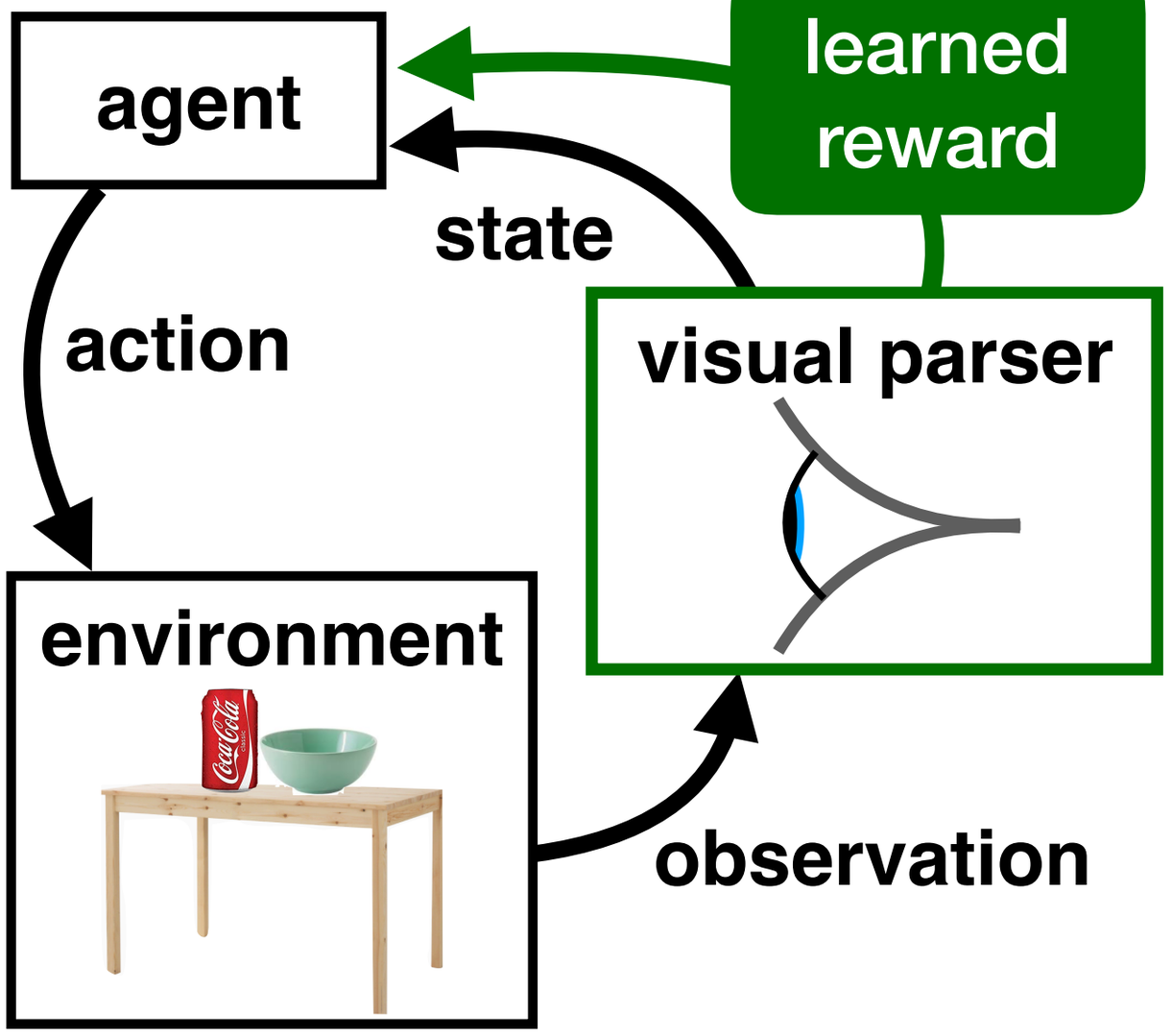


3. Language Grounding from Narrated Demonstrations

CVPR'18



Prior work used manually coded rewards



We ground rewards to the sensory input of the agent

Affordandable visual representations

We seek visual cues that obey basic common sense constraints and basic affordability reeasoning:

- Objects have 3D extent
- Objects do not interpenetrate in 3D
- Objects come in regular sizes
- Objects persist over time

Grounding language on such representations would be able to support affordability inference

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

Affordandable visual representations

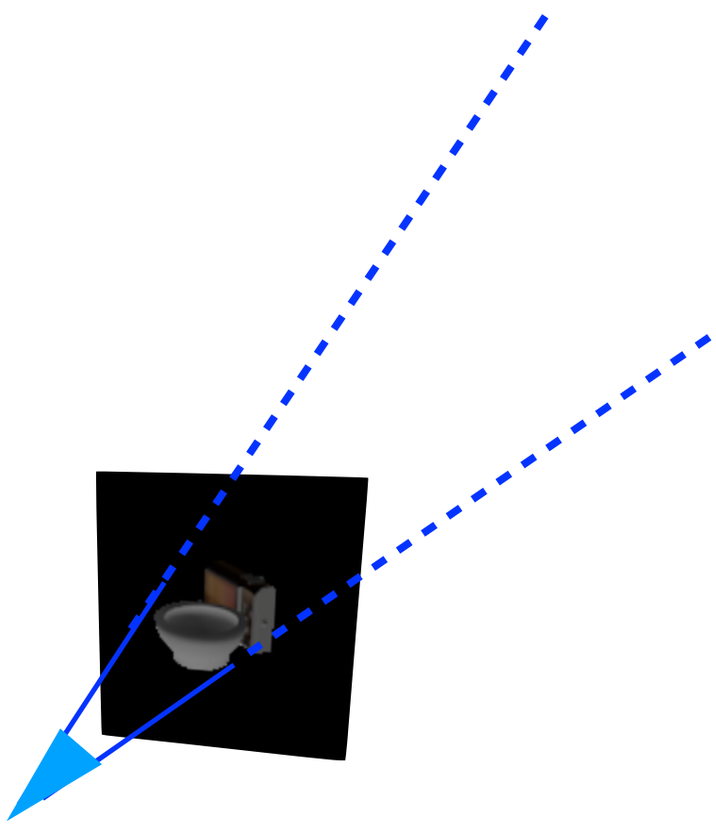
We seek visual cues that obey basic common sense constraints and basic affordability reeasoning:

- Objects have 3D extent
- Objects do not interpenetrate in 3D
- Objects come in regular sizes
- Objects persist over time

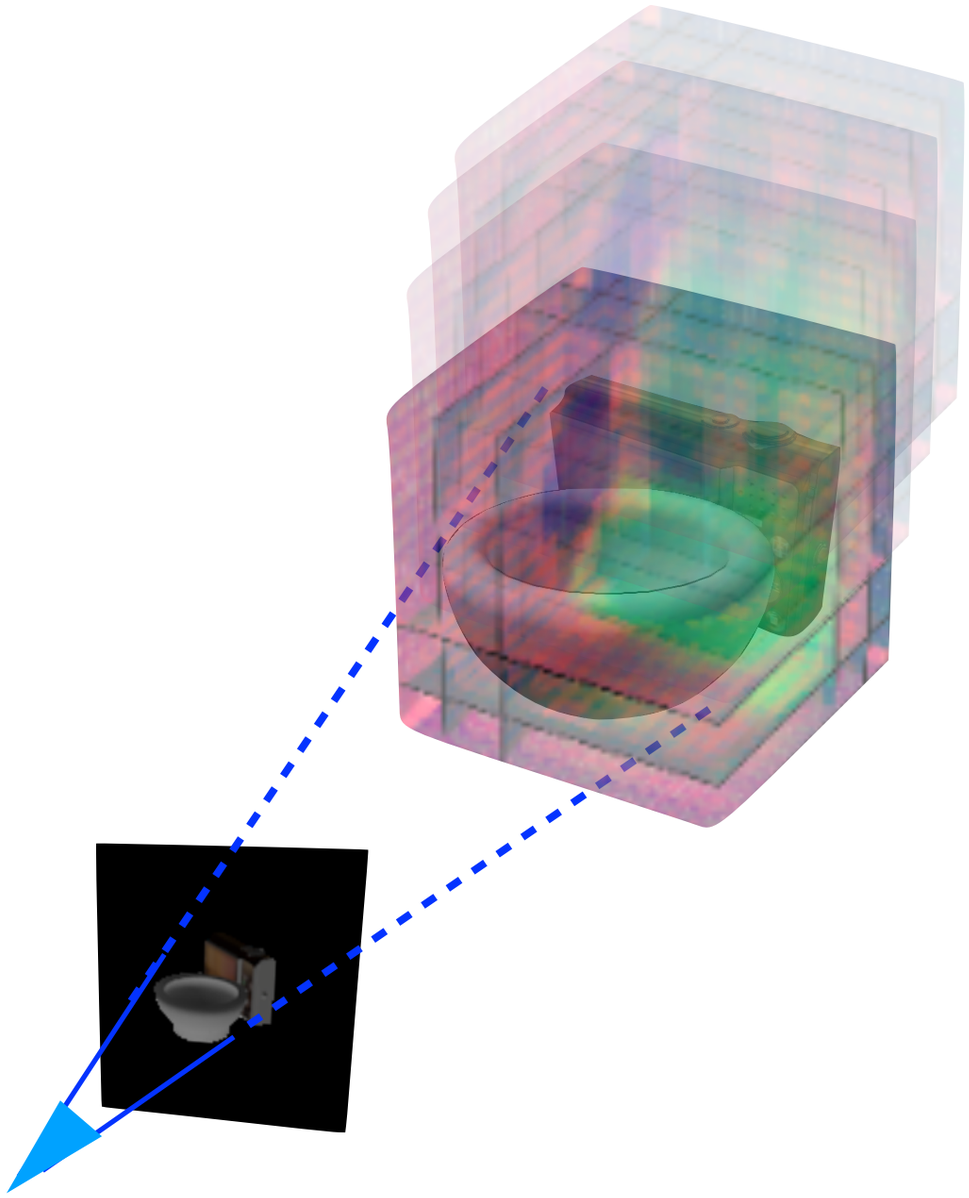
Grounding language on such representations would be able to support affordability inference

Modeling Relationships in Referential Expressions with Compositional Modular Networks, Ronghang et al.

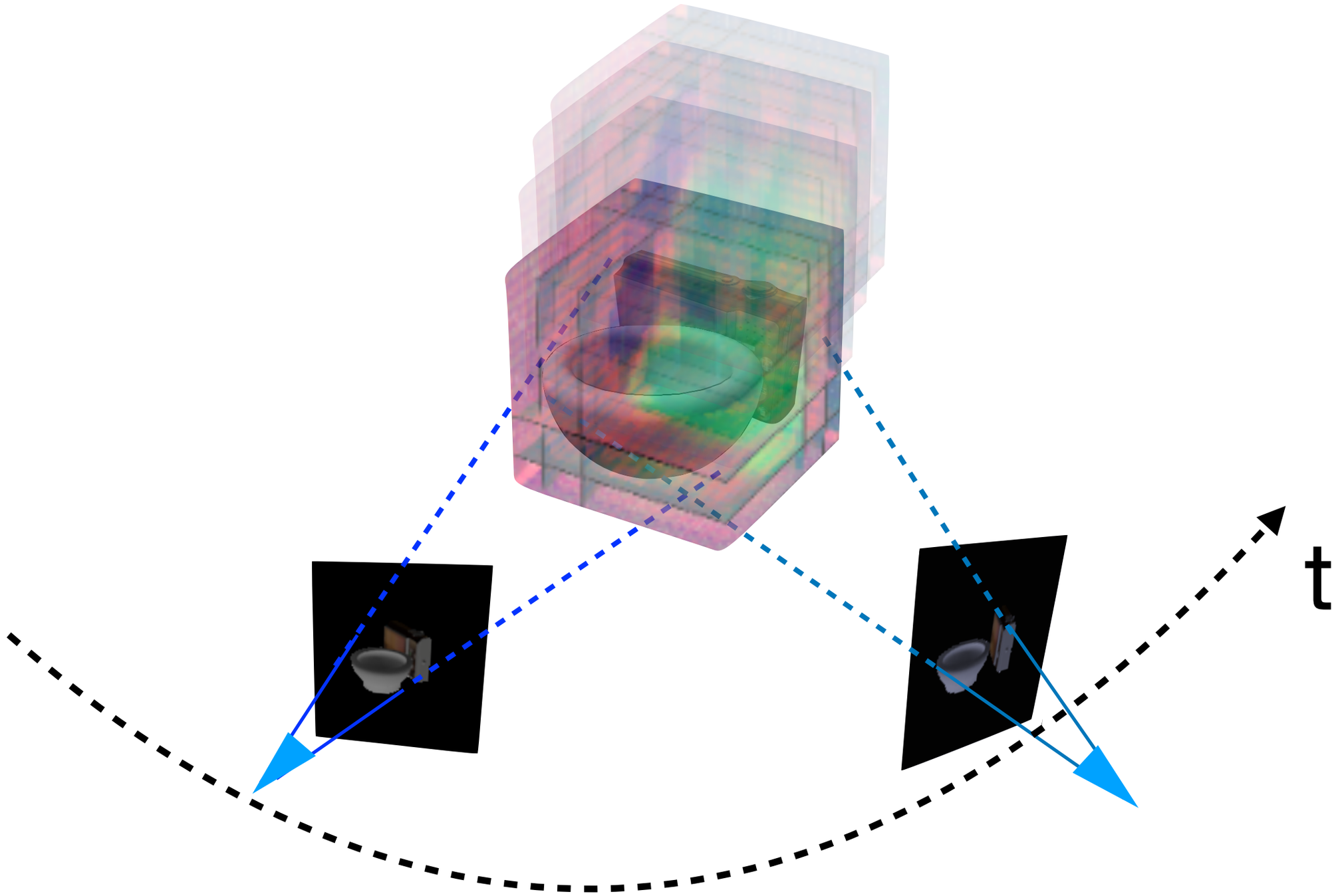
Geometry-Aware Recurrent Networks



Geometry-Aware Recurrent Networks



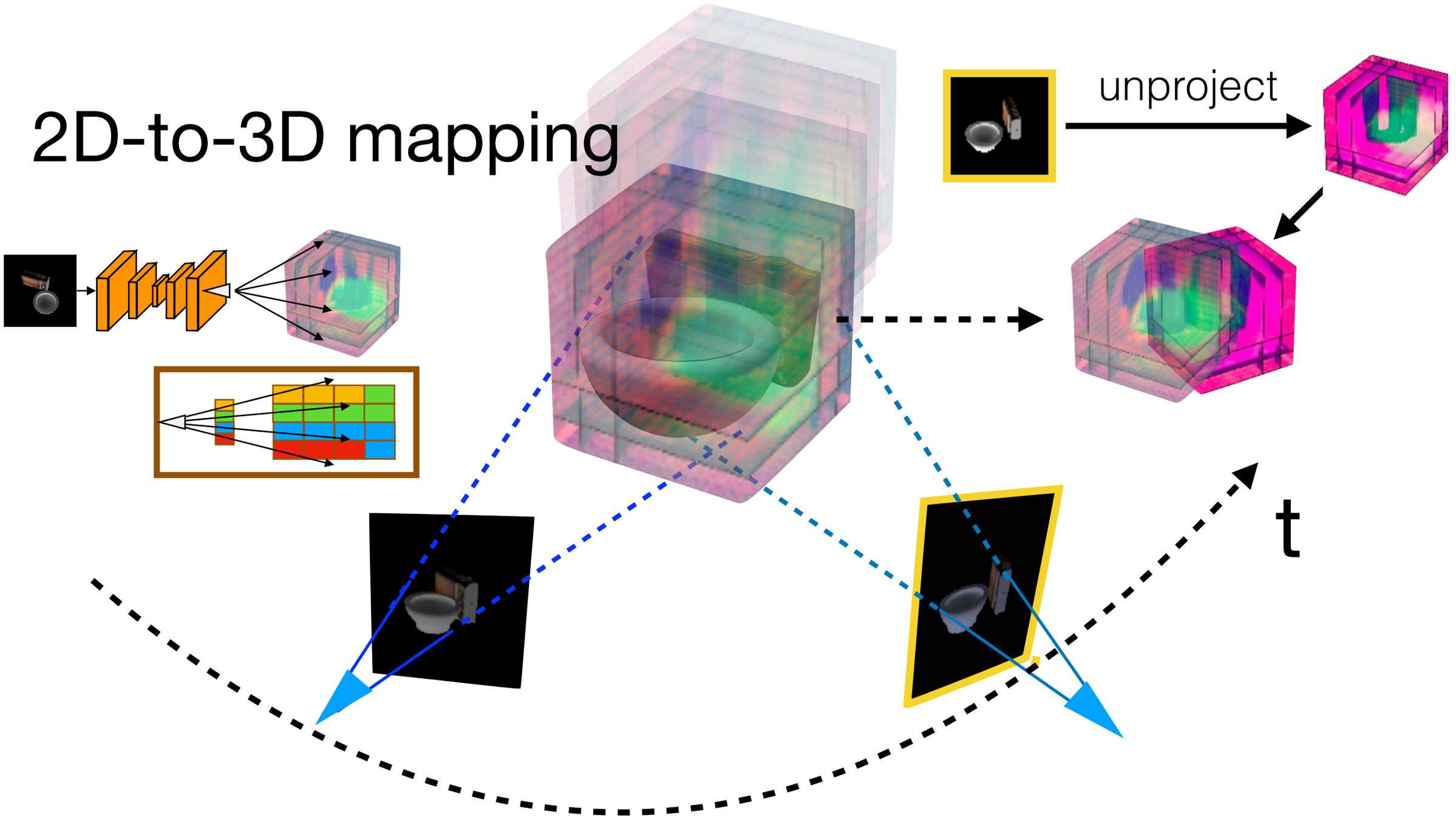
Geometry-Aware Recurrent Networks



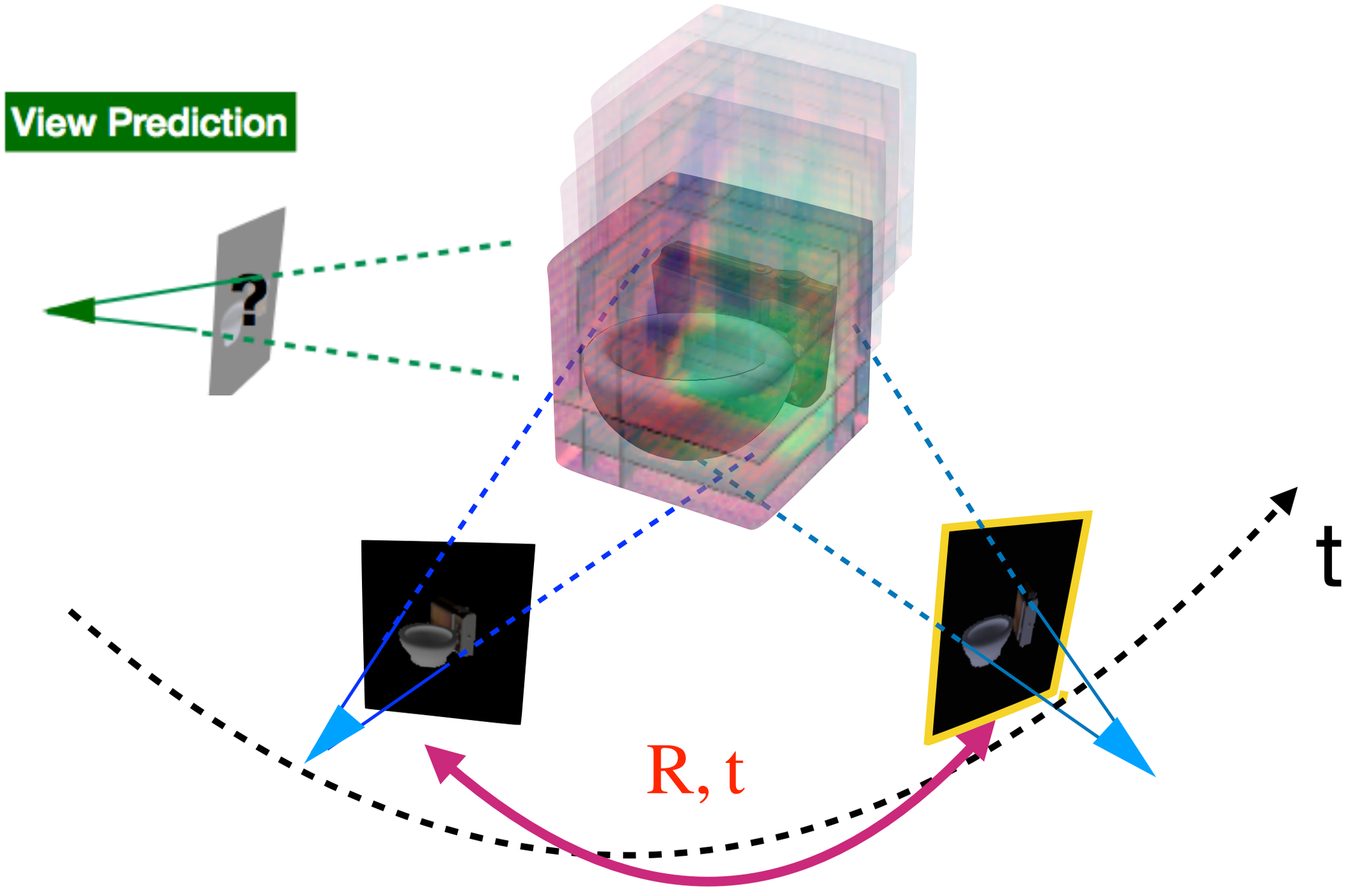
Geometry-Aware Recurrent Networks

egomotion-stabilized update

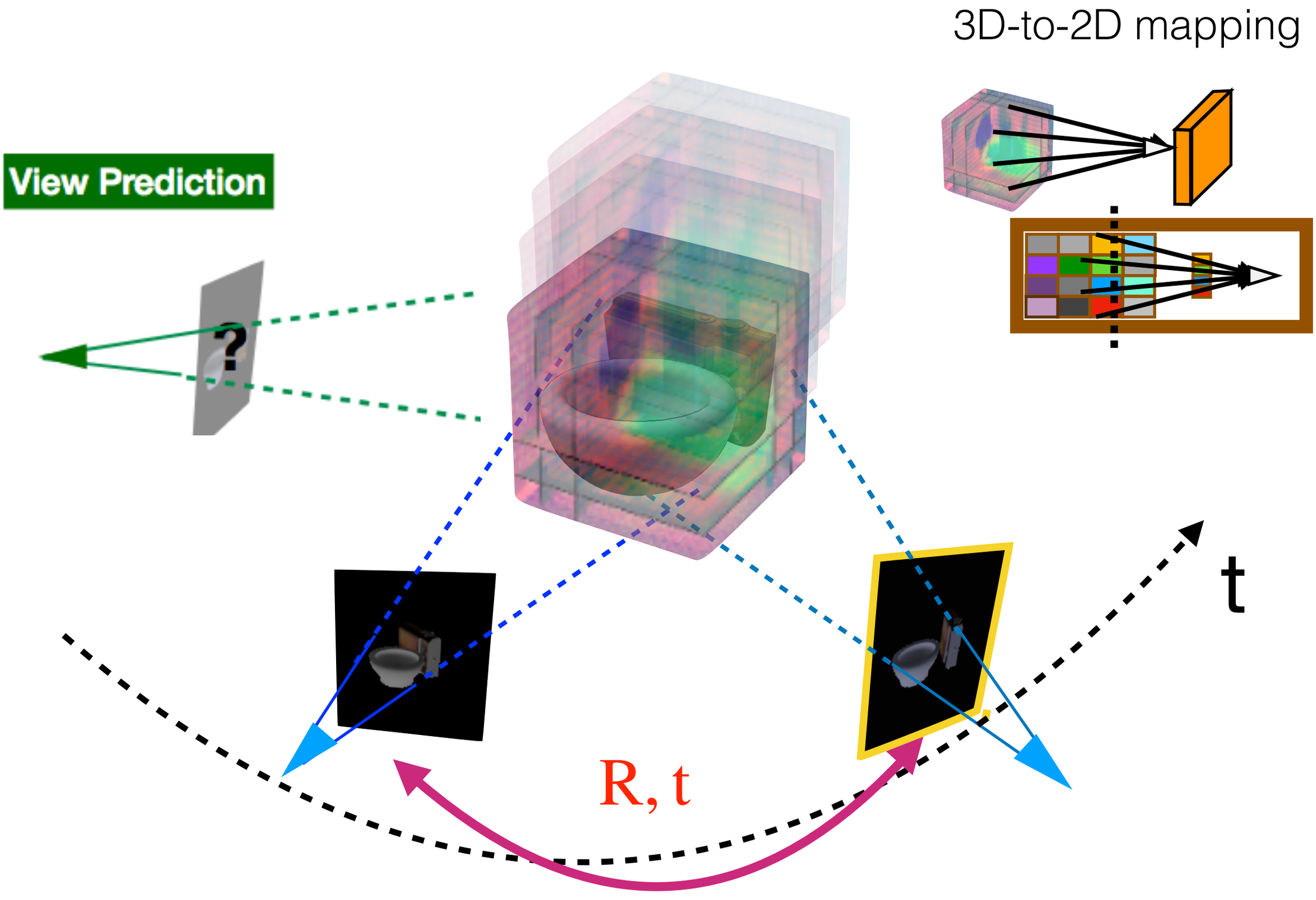
2D-to-3D mapping



View prediction

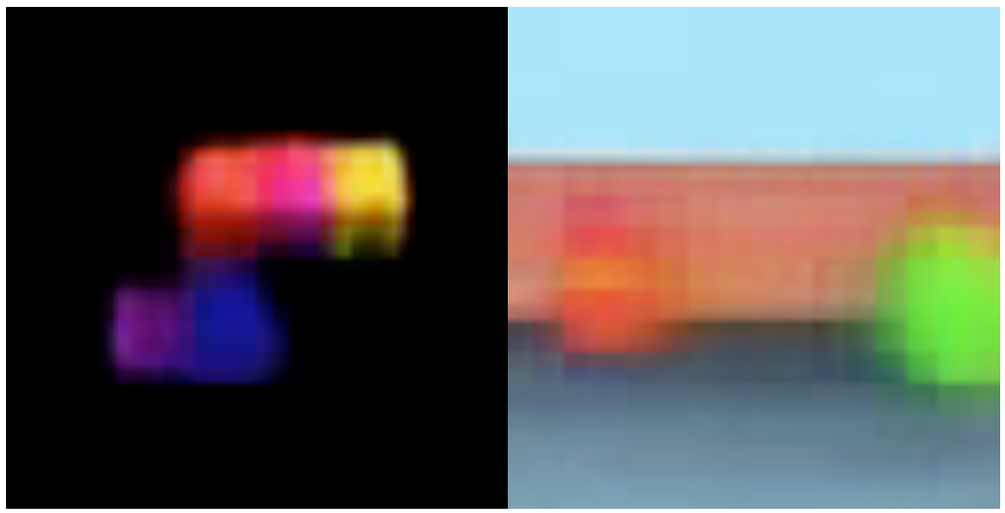


View prediction



View prediction

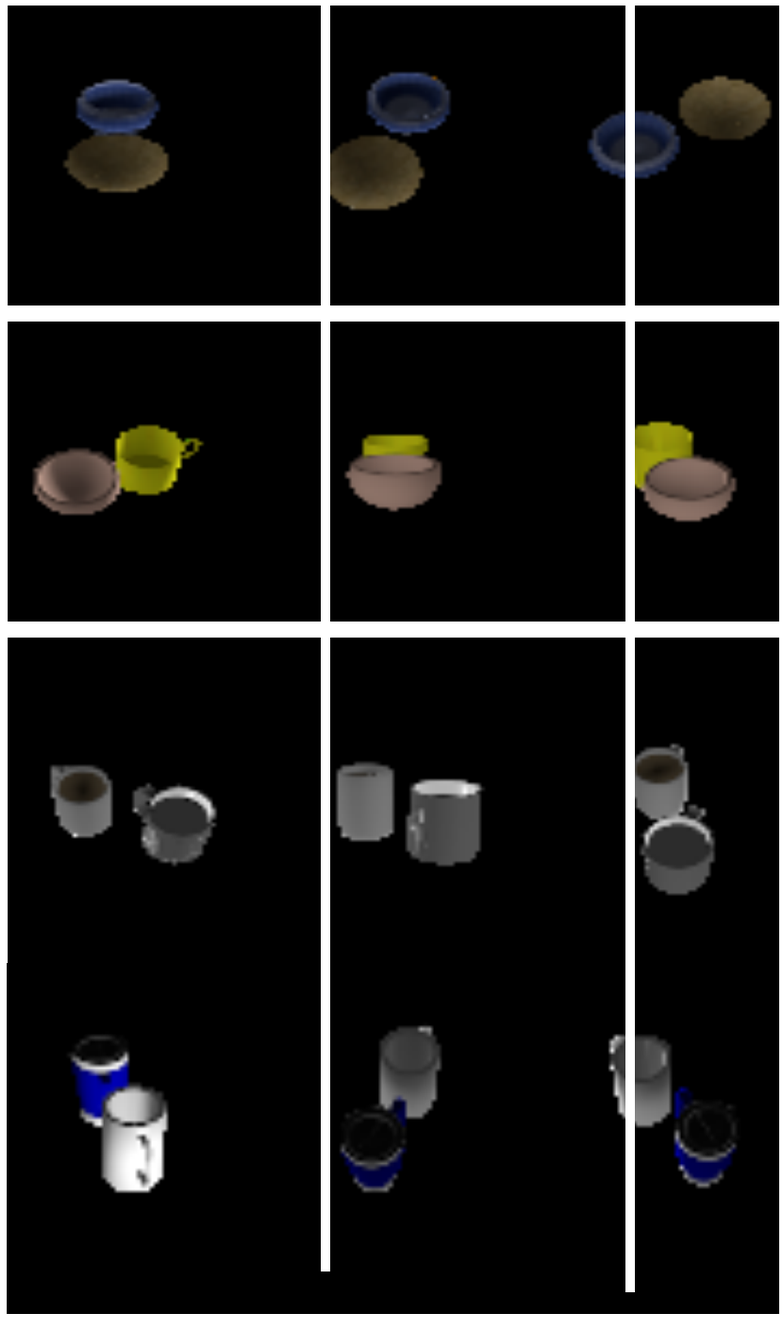
geometry-aware RNN



2D RNN [1]

View prediction

3 input views



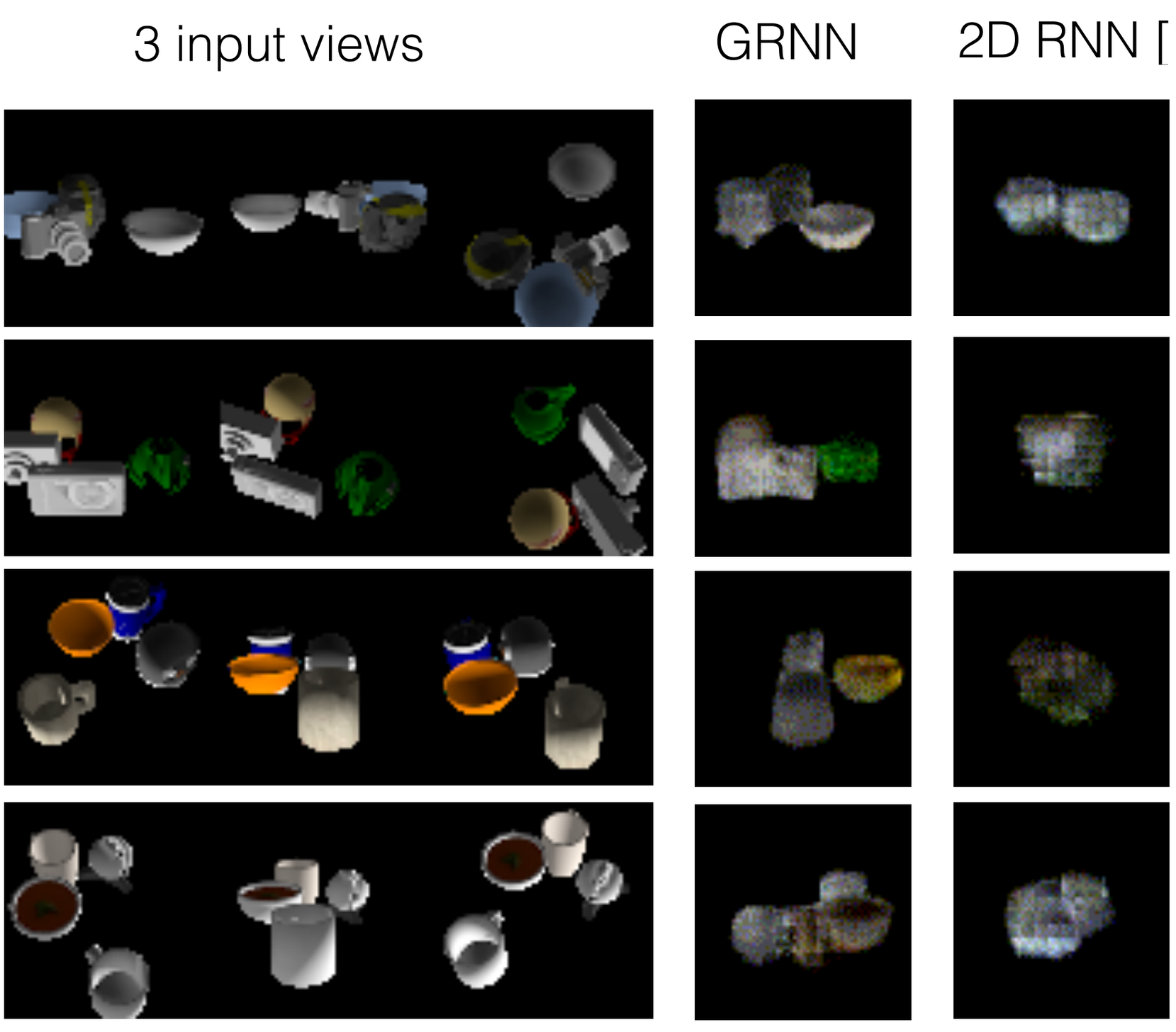
GRNN

2D RNN [1]

[1] Neural scene representation and rendering DeepMind, Science, 2018

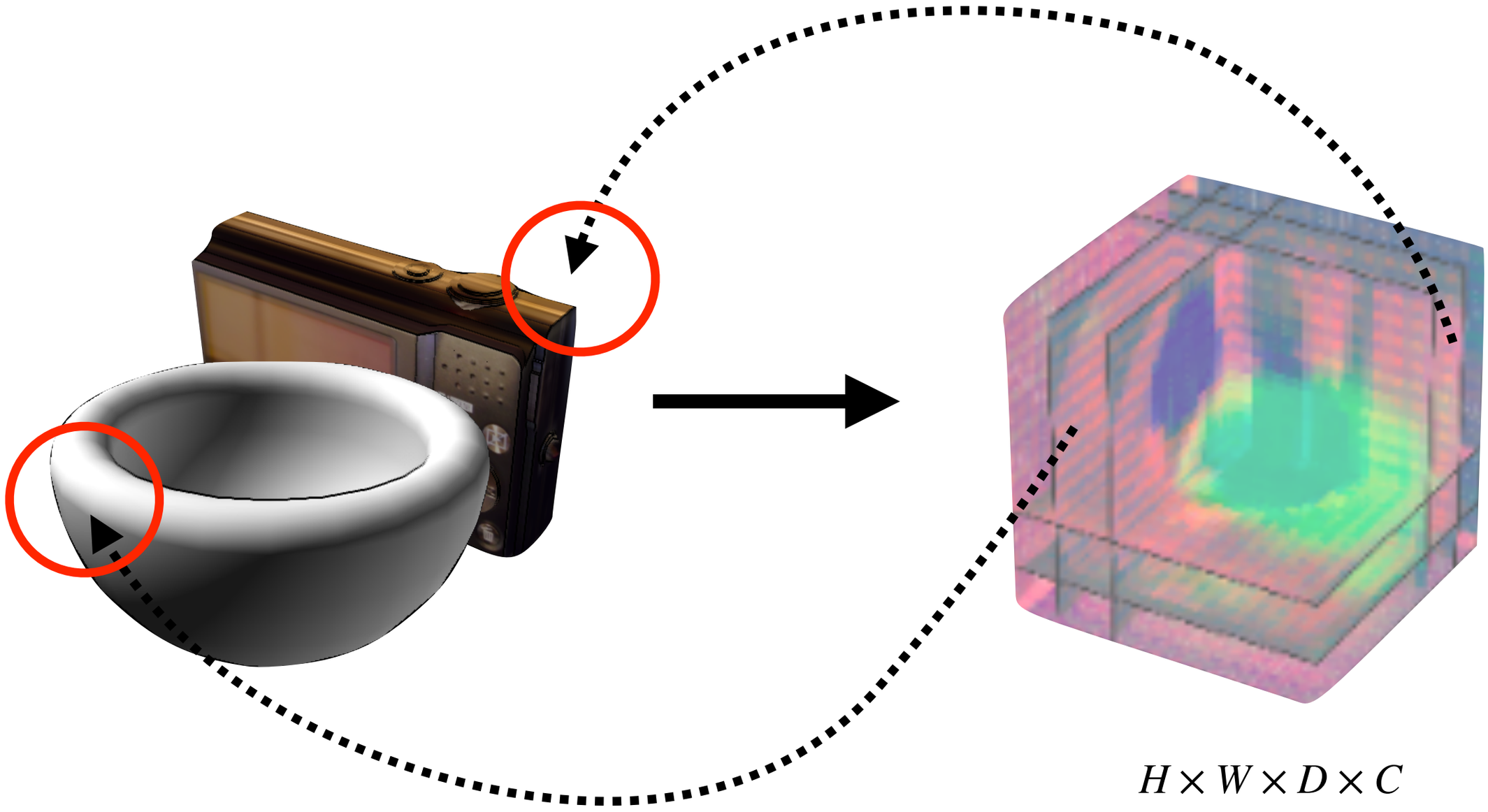
View prediction

Truly novel scenes

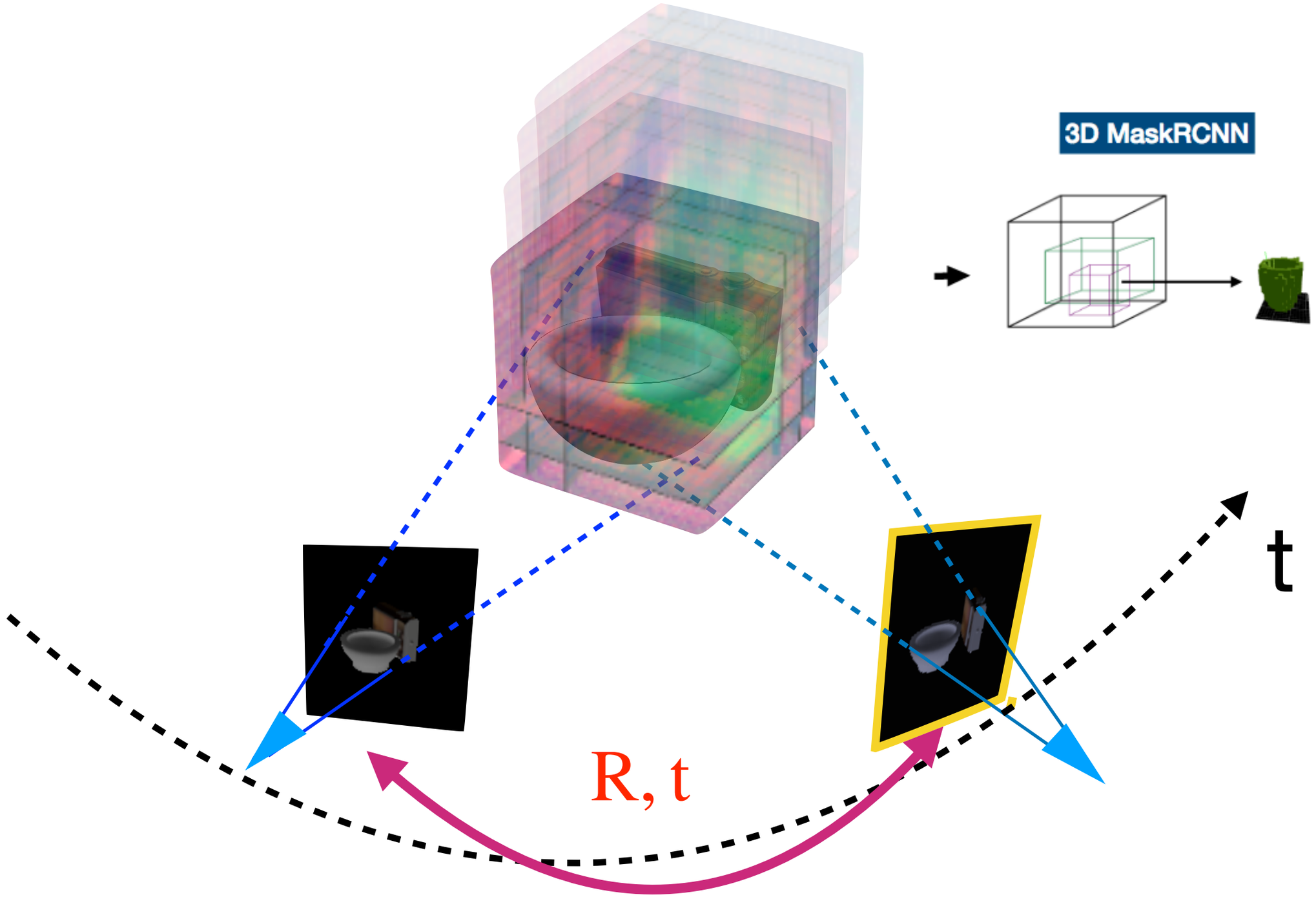


[1] Neural scene representation and rendering DeepMind, Science, 2018

Geometry-Aware Recurrent Networks (GRNNs)

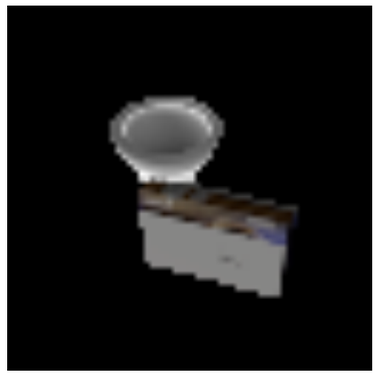


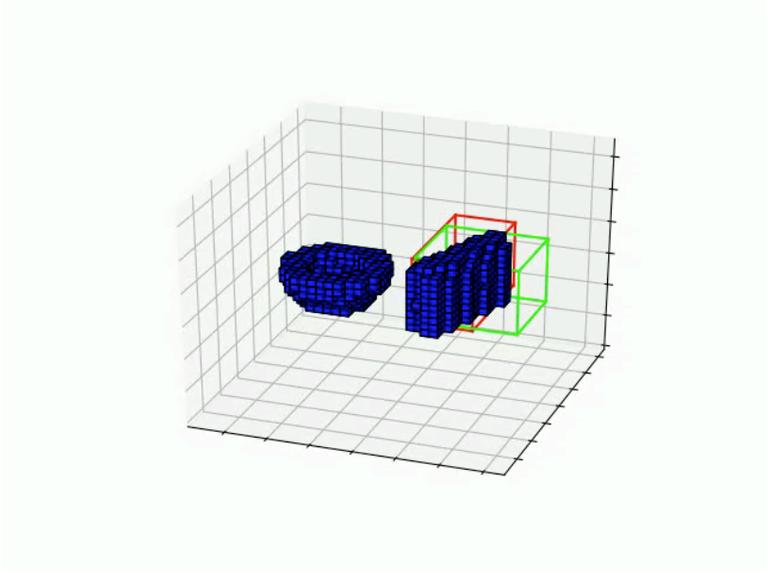
3D object detection



3D object detection

of input views

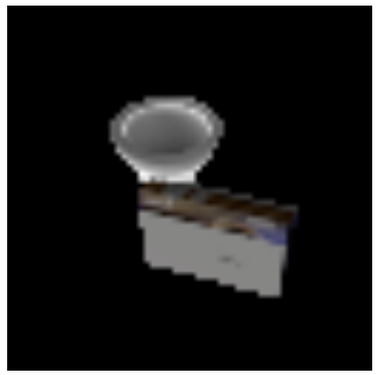


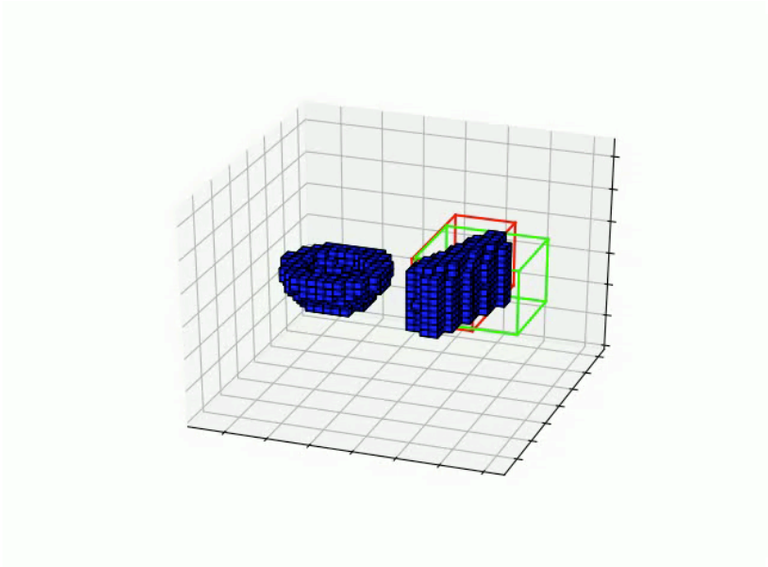


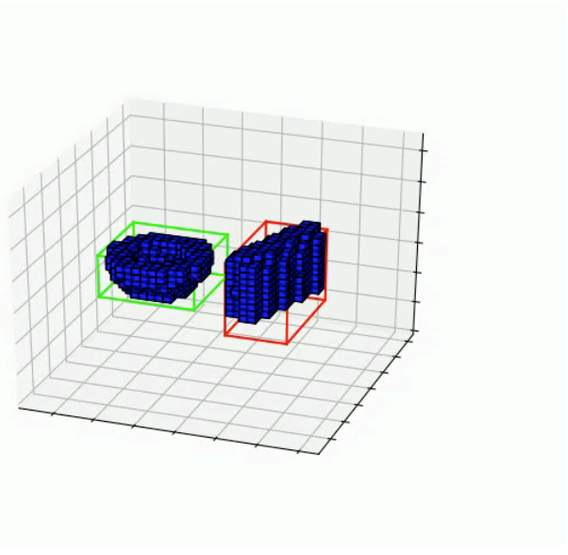
3D object detection

Results - 3D object detection

of input views

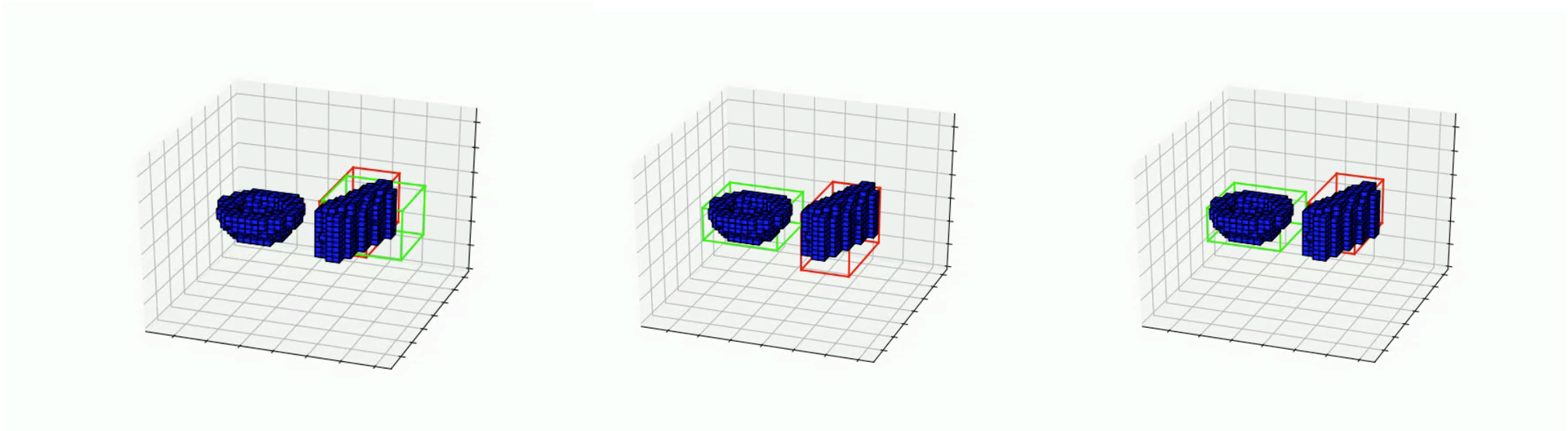






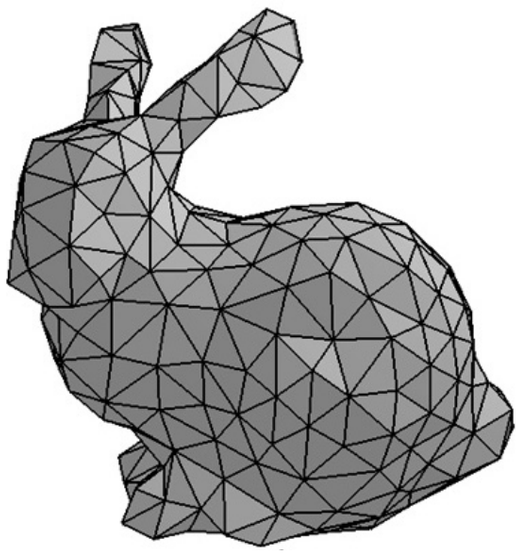
3D object detection

of input views

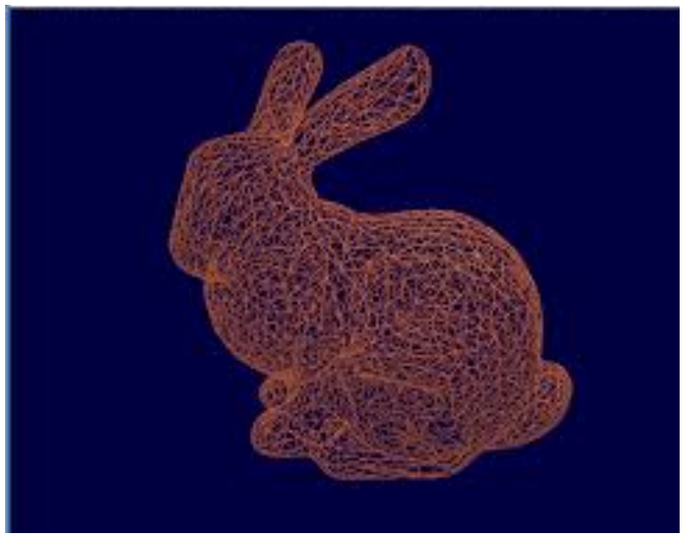


3D representations

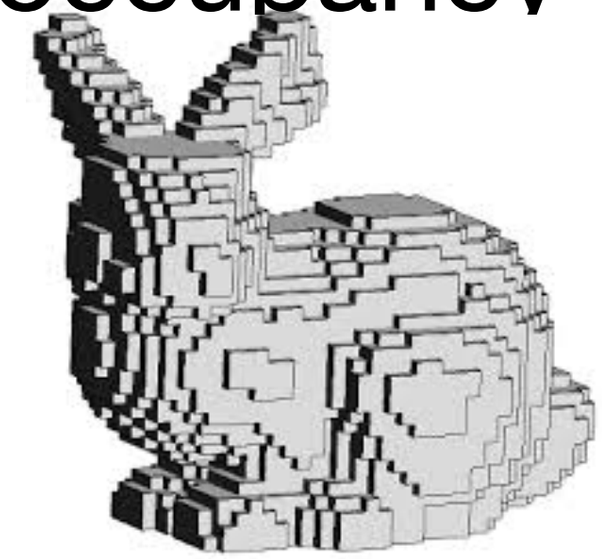
3d mesh



3d point

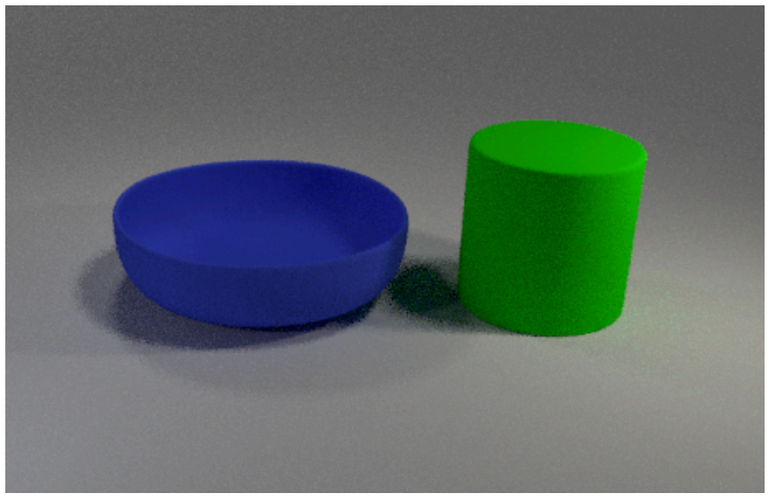


3d voxel occupancy



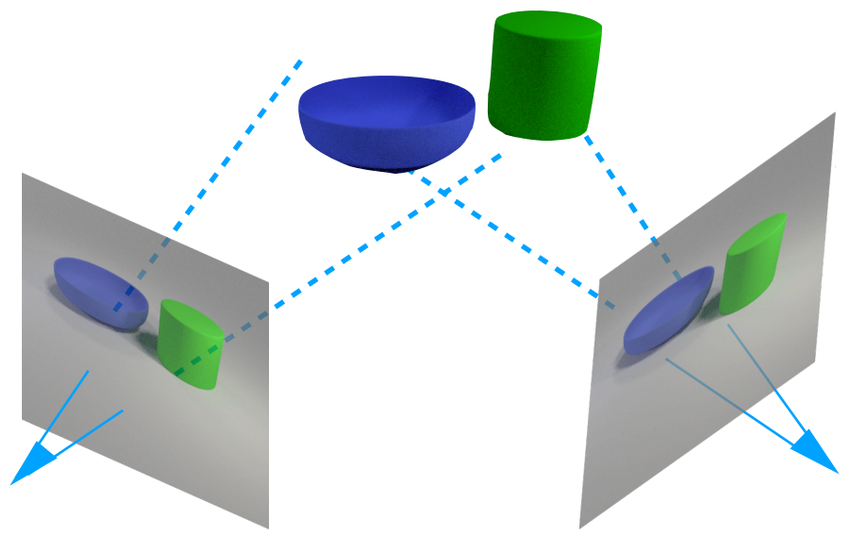
- ask too much: high level of 3D details may be impossible to obtain
- ask too little: information about semantics of the objects is not prese

“The green rubber cylinder is on the right of the blue bowl”



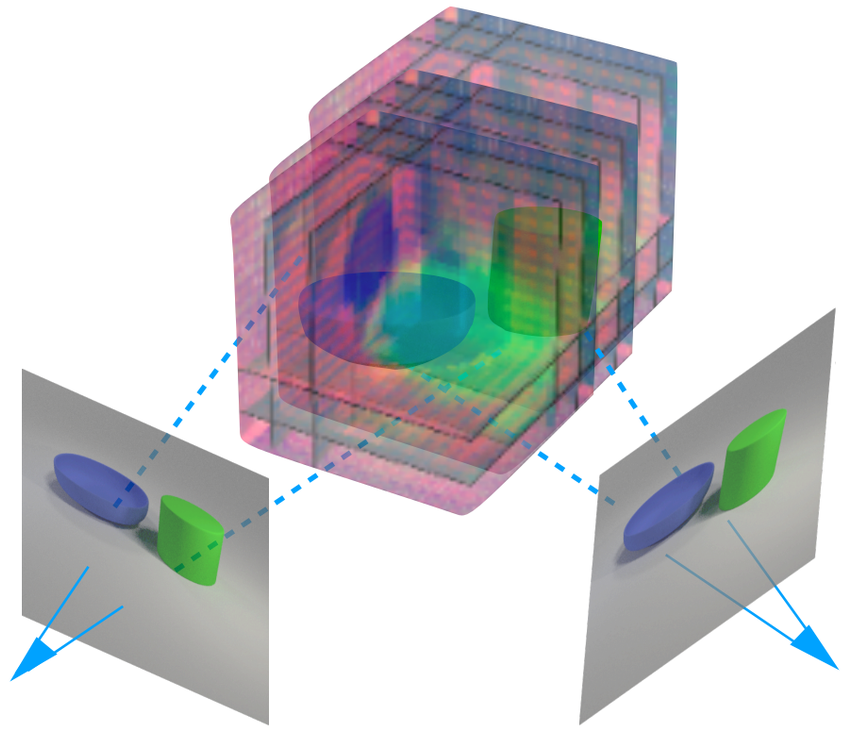
1. We consider an embodied agent that can see a scene from multiple viewpoints

“The green rubber cylinder is on the right of the blue bowl”



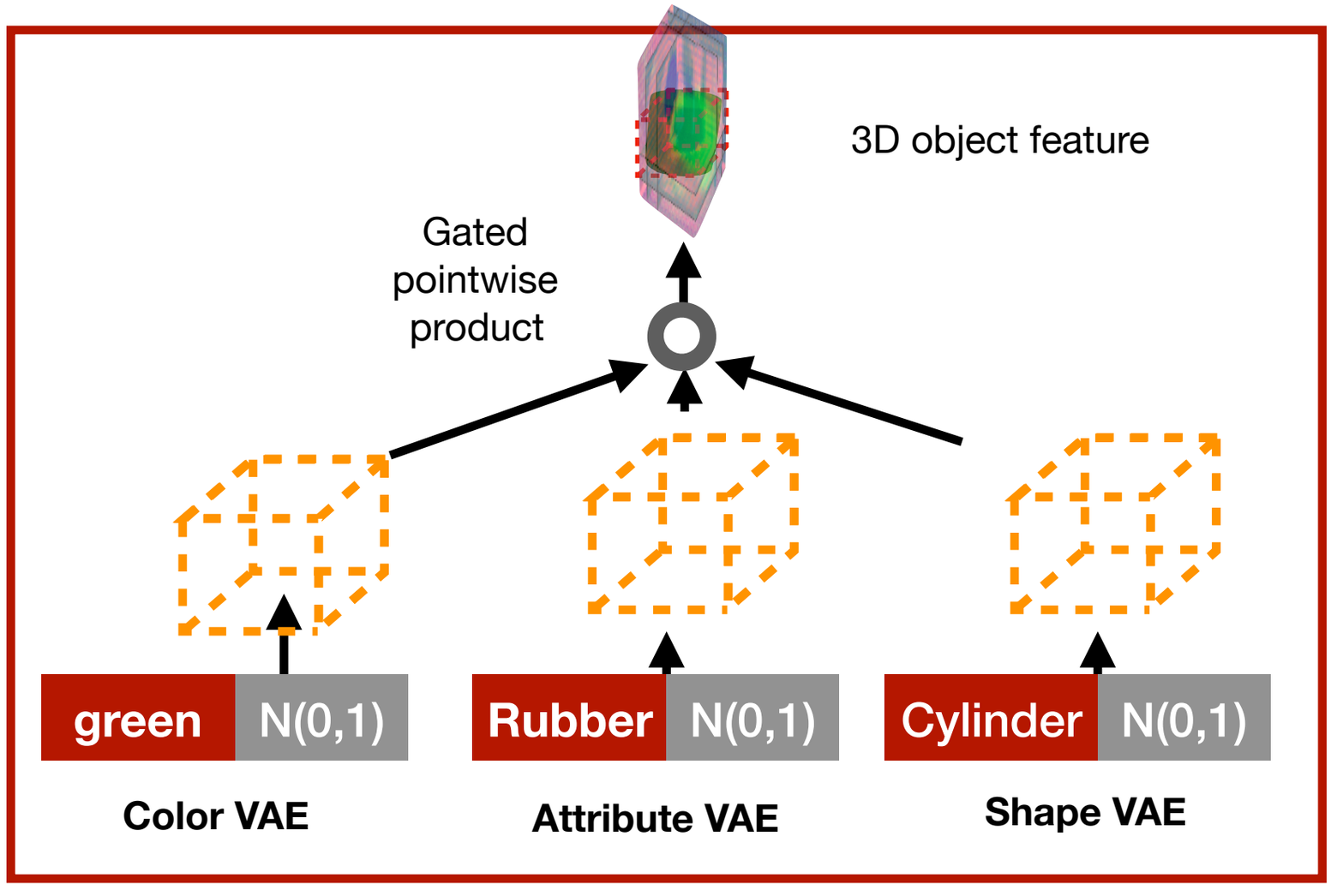
1. We consider an embodied agent that can see a scene from multiple viewpoints

“The green rubber cylinder is on the right of the blue bowl”



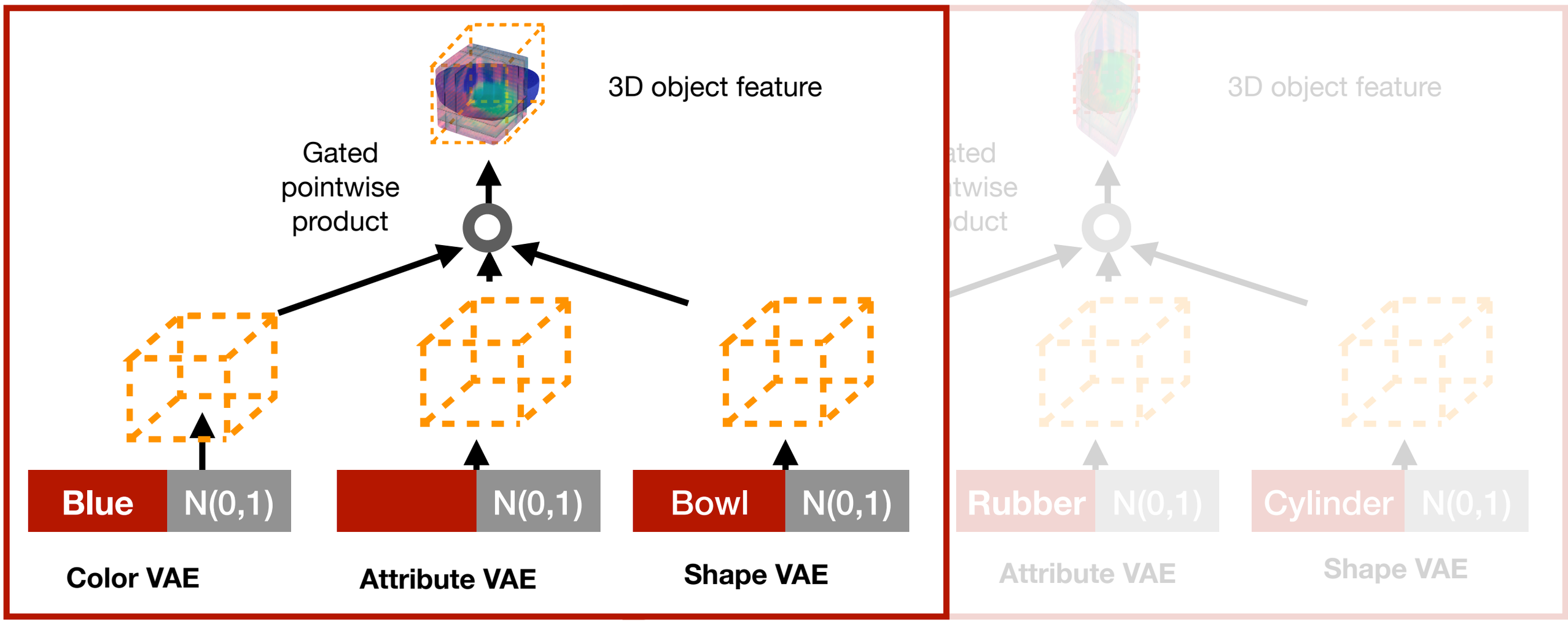
1. Our agent learns to map an RGB image to a set of 3D feature maps by training GRNNs to predict views

*"The **green rubber cylinder** is on the right of the blue bowl"*



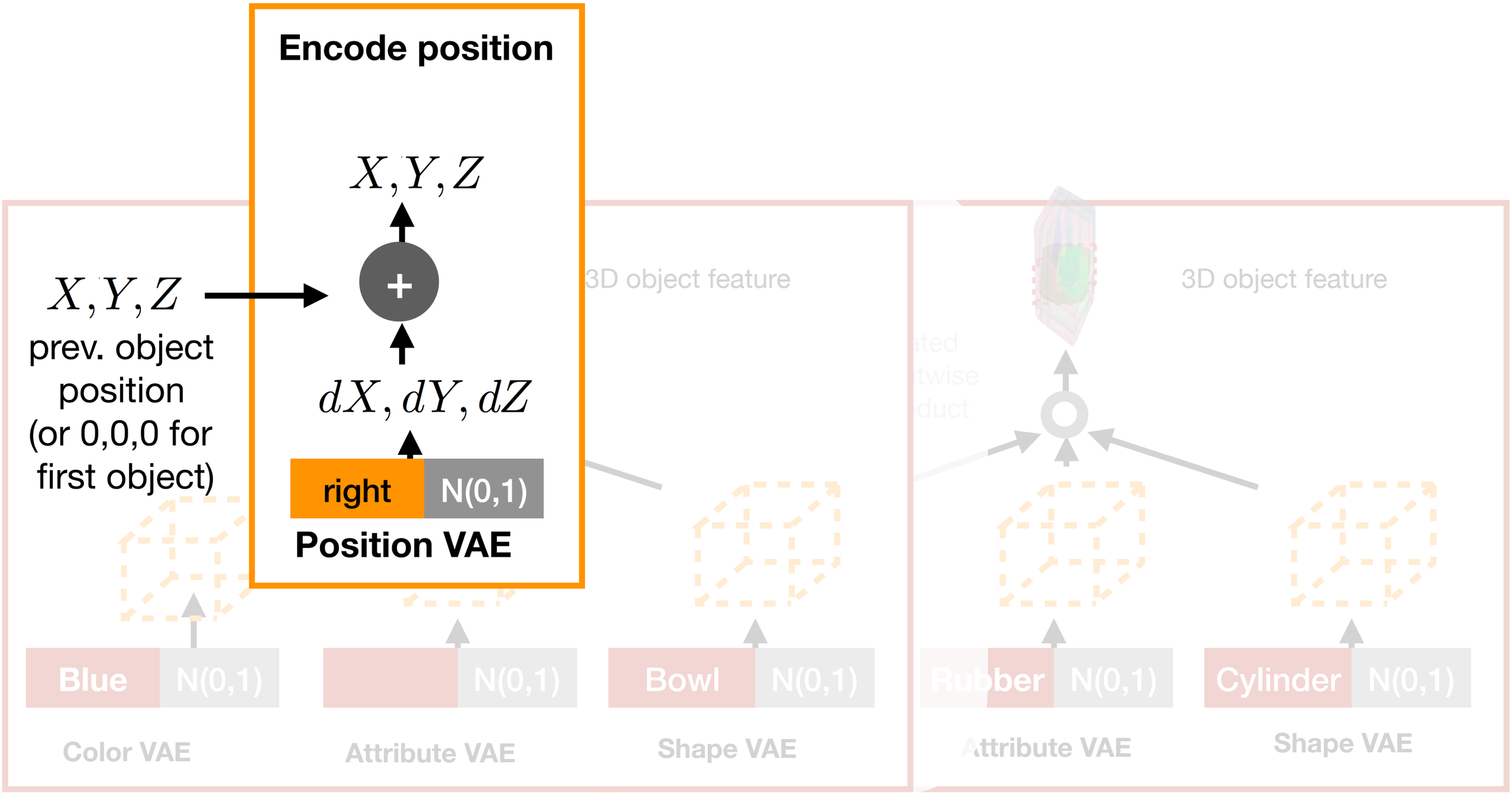
1. Our agent maps noun phrases to object-centric 3D feature maps

*"The green rubber cylinder is on the right of the **blue bowl**"*



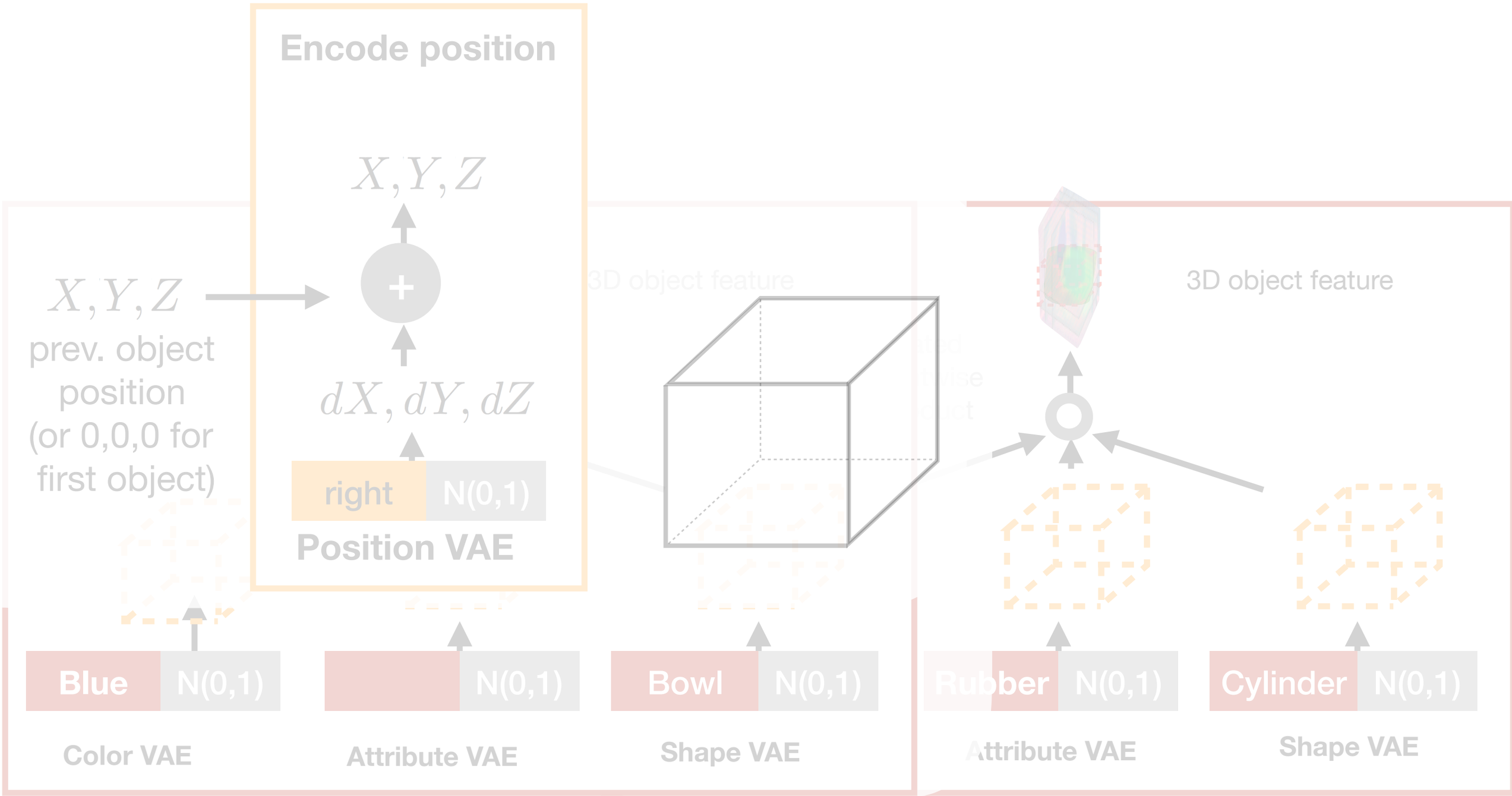
1. Our agent maps noun phrases to object-centric 3D feature maps

"The green rubber cylinder is
***on the right of** the blue bowl"*



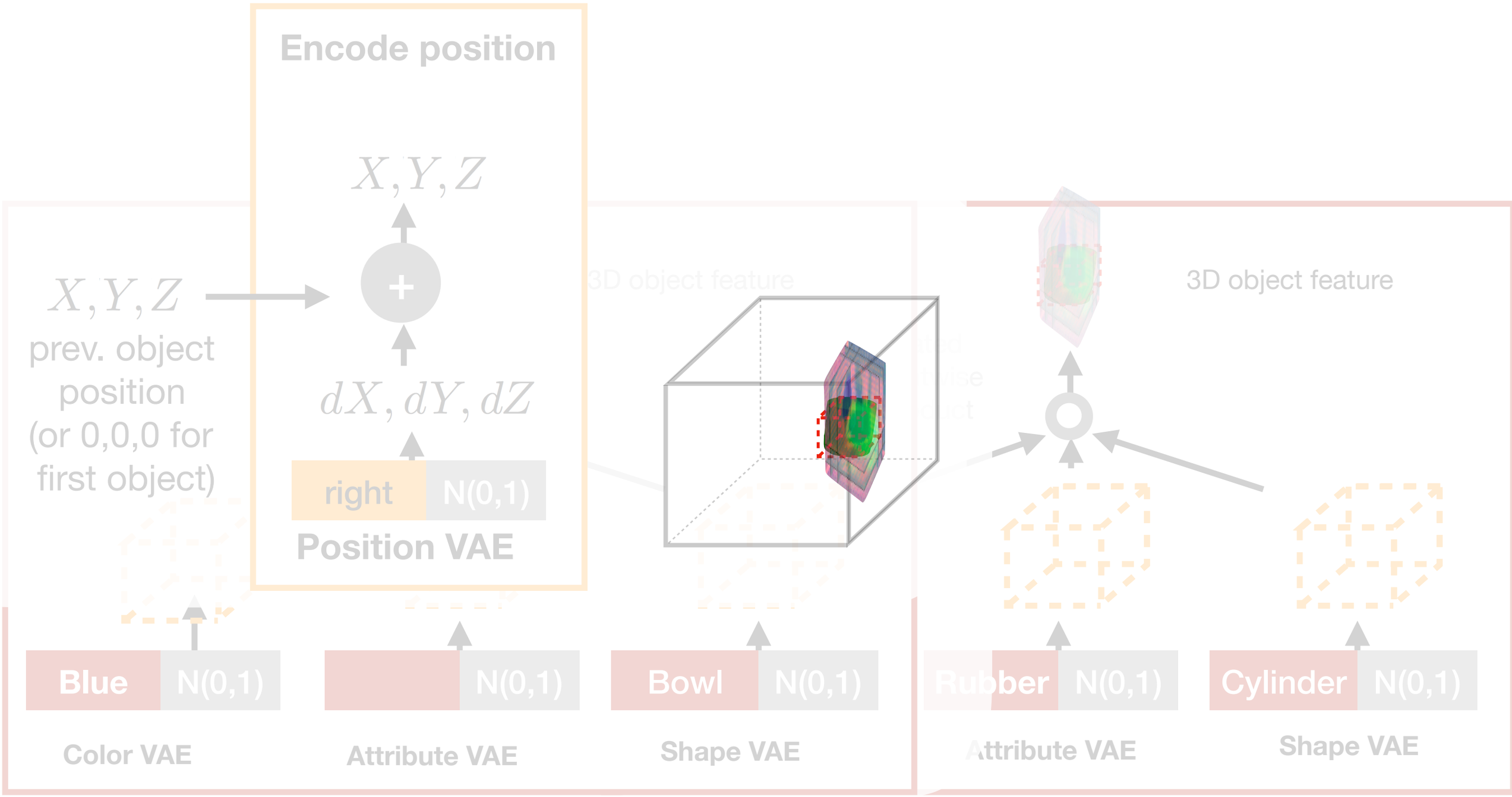
1. Our agent maps spatial expressions to relative 3D offsets

“The green rubber cylinder is on the right of the blue bowl”



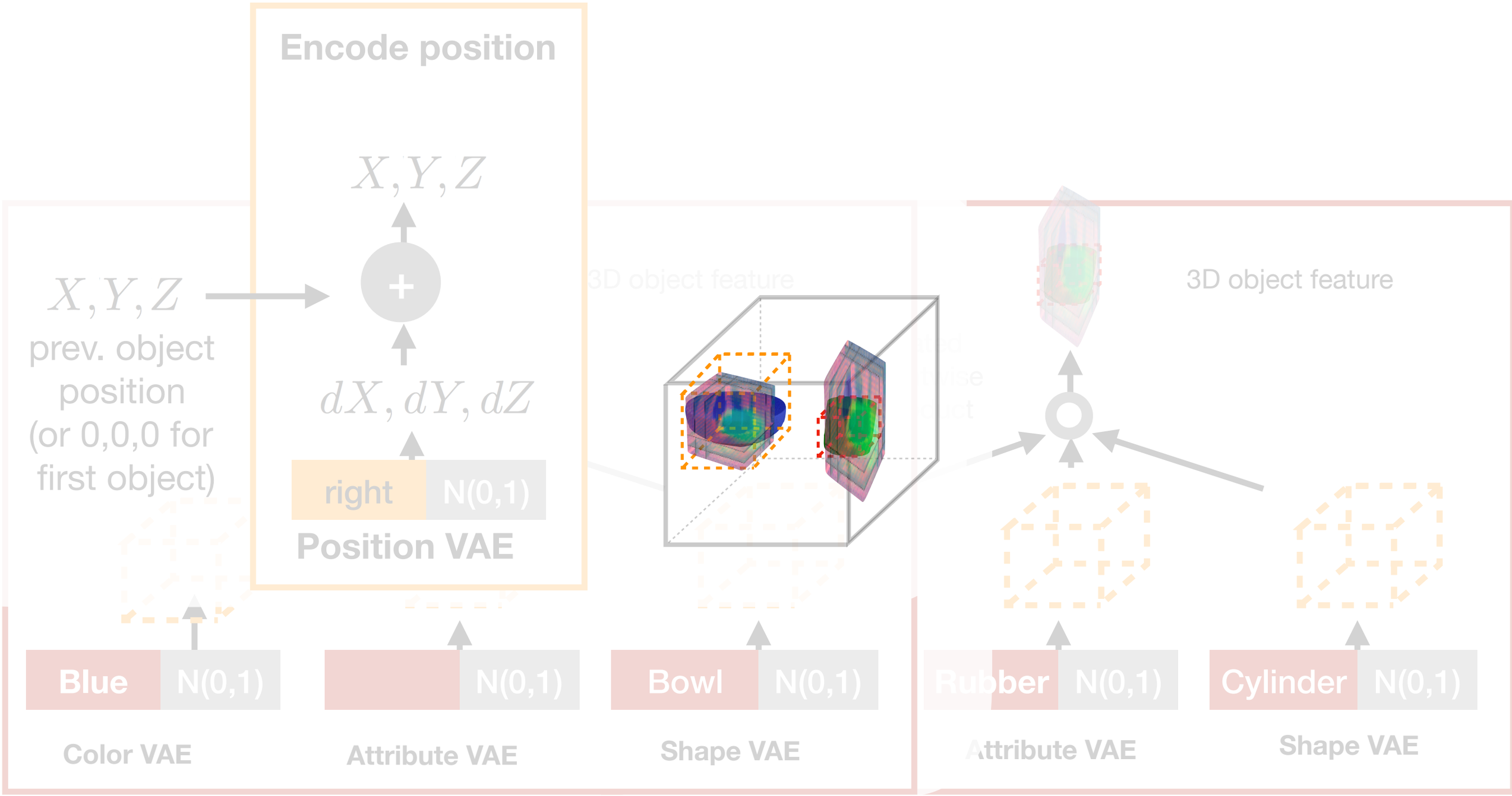
1. Our agent populates a 3D canvas with the predicted object tensors adn their relative offsets

"The green rubber cylinder is on the right of the blue bowl"



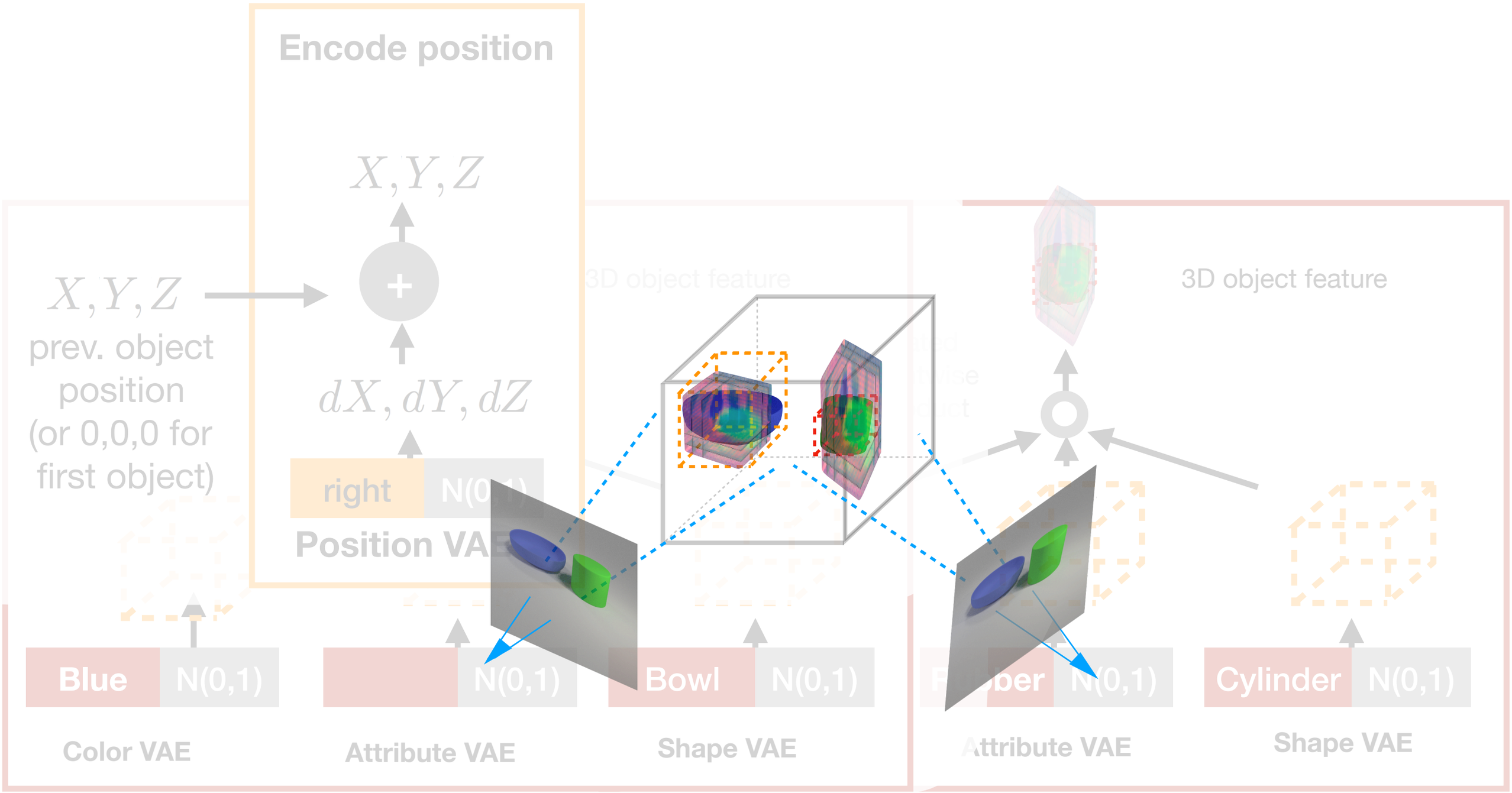
1. Our agent populates a 3D canvas with the predicted object tensors adn their relative offsets

"The green rubber cylinder is on the right of the blue bowl"



1. Our agent populates a 3D canvas with the predicted object feature maps and their relative spatial offsets

"The green rubber cylinder is on the right of the blue bowl"



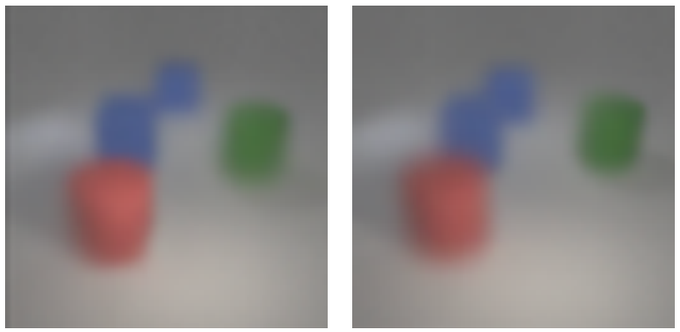
1. The generated canvas when projected should match the RGB image views

Scene imagination

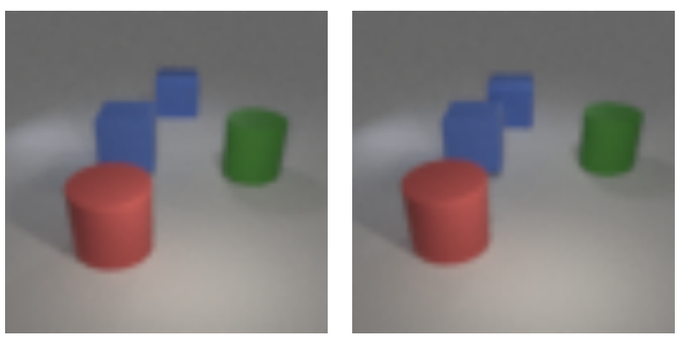
Red Rubber Cylinder to the left front of Blue Rubber Cube to the left front of Green Rubber Cylinder to right front of Blue Rubber Cube

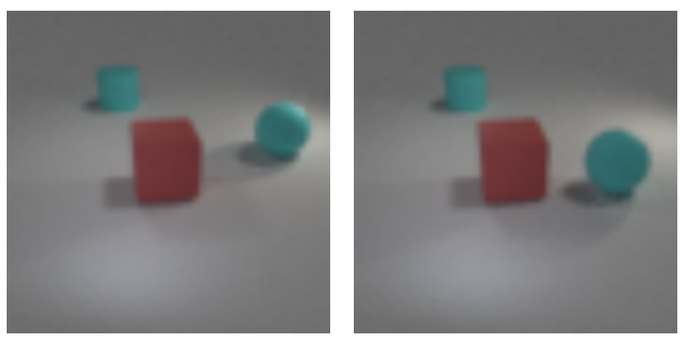
Red Rubber Cube to the left front of the Blue Rubber Sphere to the right front of Cyan Metal Cylinder

Neural render



Blender render





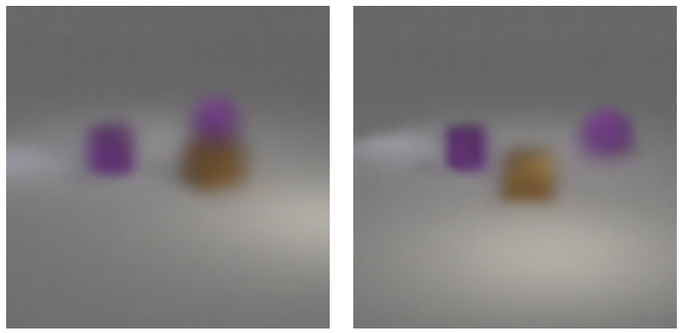
Scene imagination

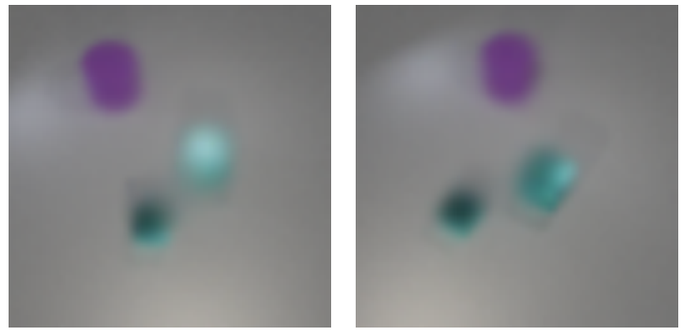
Natural language utterance

Purple Cylinder to the left behind of Brown Cube to the left front of Purple Sphere

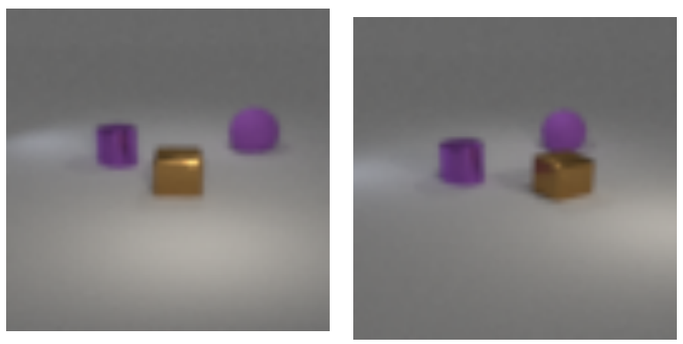
Purple Cylinder to the left behind of Cyan Cube to the left front of Cyan Cube

Neural render





Blender render



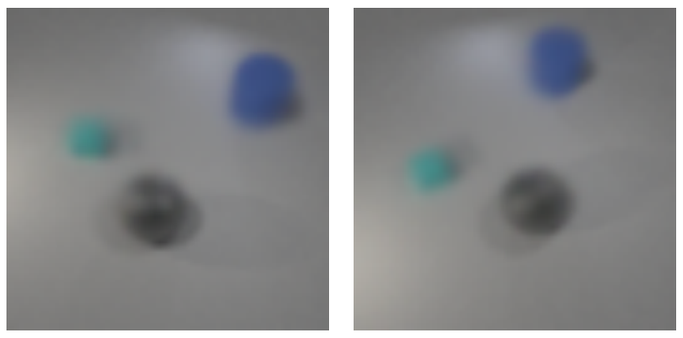
Scene imagination

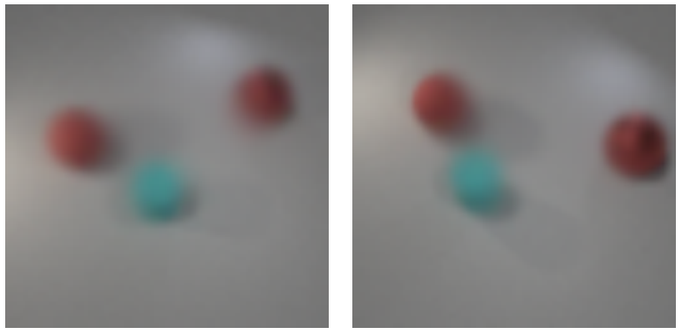
Natural language utterance

Cyan Cube to the left behind of Gray Sphere to the left front of Blue Cube

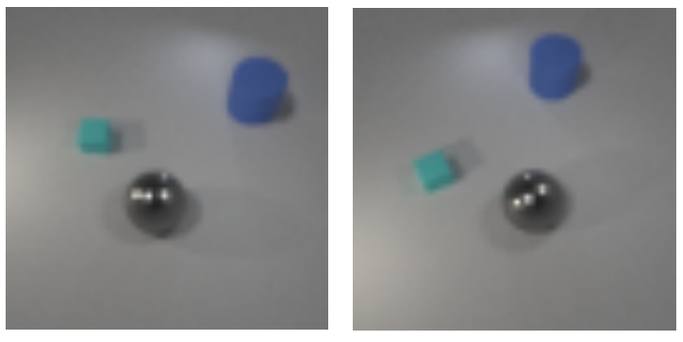
Red Sphere to the left behind of Cyan Cylinder to the left front of Red Sphere

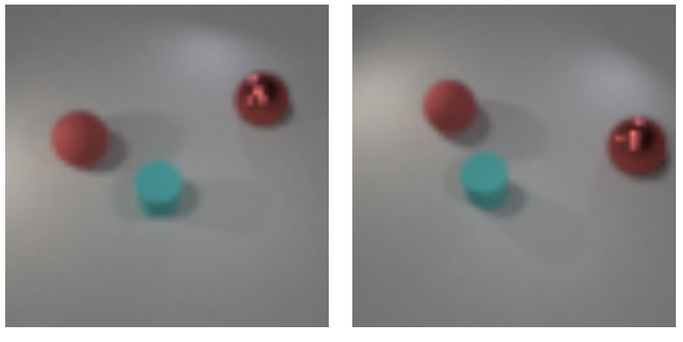
Neural render





Blender render





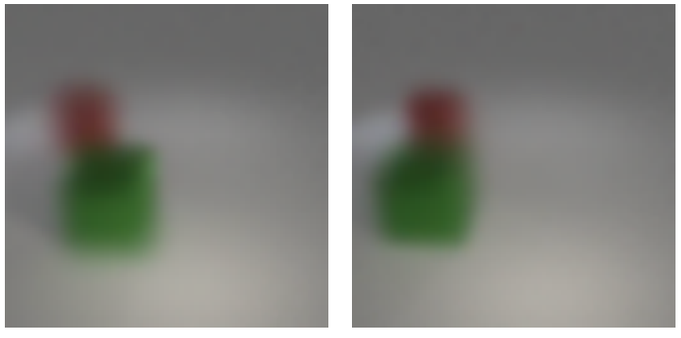
Scene imagination

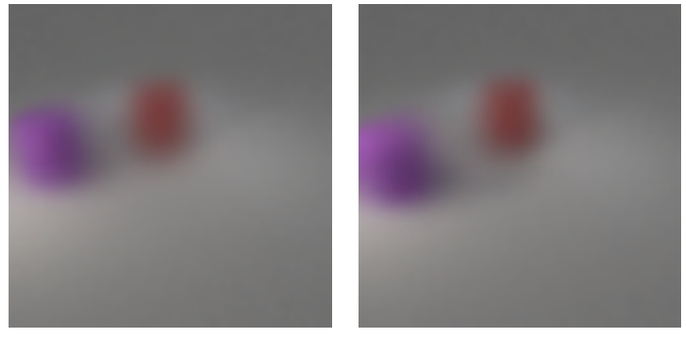
Natural language utterance

“red cylinder to the right behind of green cube”

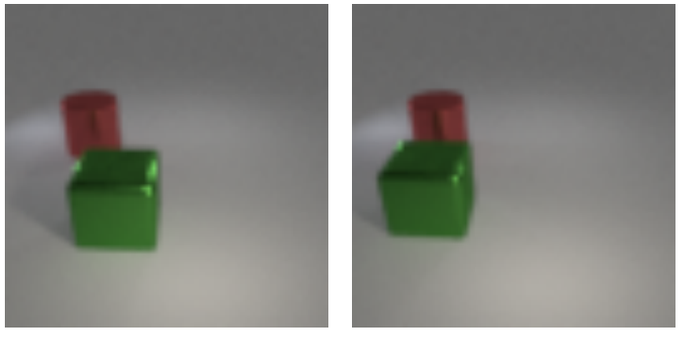
“pink cylinder to the left front of red cylinder”

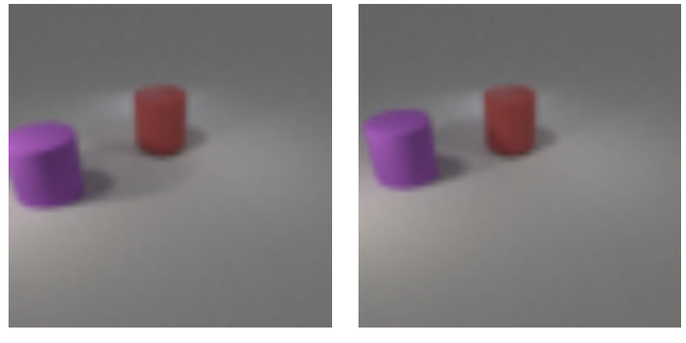
Neural render



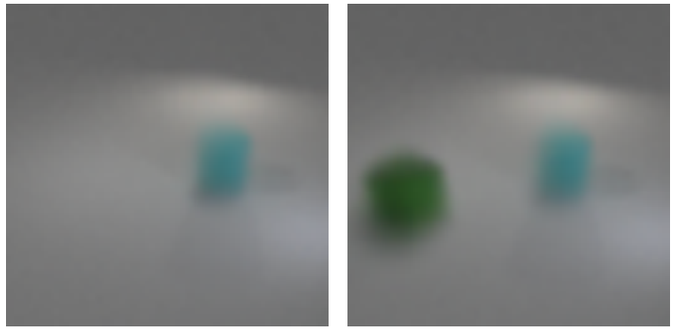
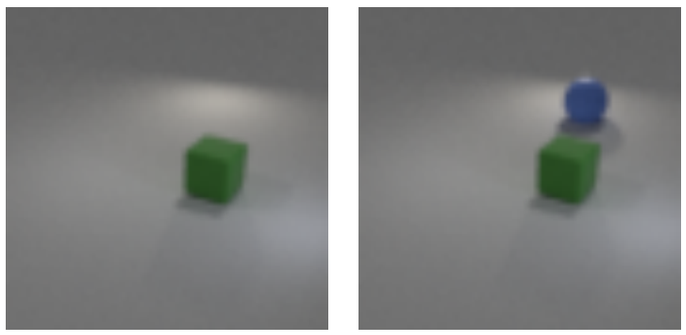
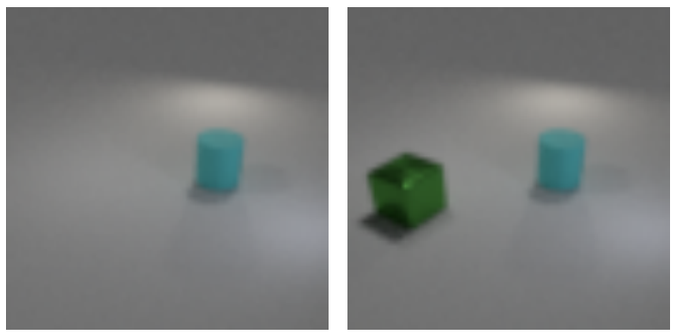


Blender render

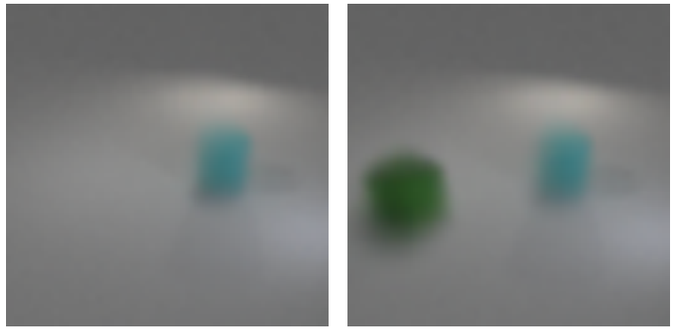
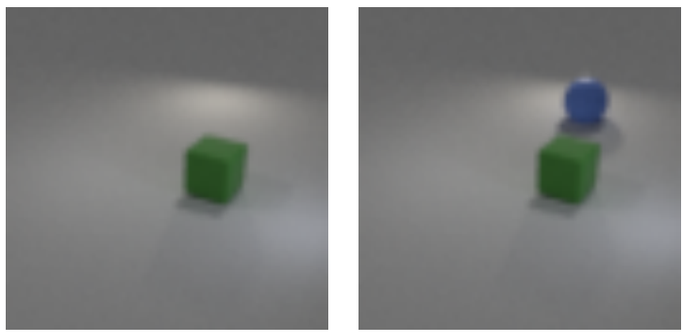
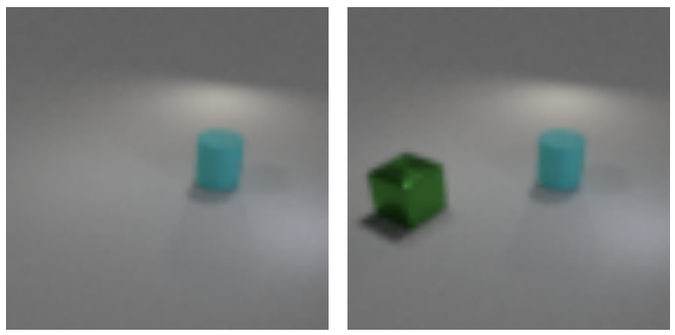




Scene alteration

Natural language utterance	*“blue sphere to the right behind of green cube”*	*“green cube to the left front of cyan cylinder”*
Neural render		
Blender render		

Affordability Inference

Natural language utterance	*“blue sphere to the right behind of green cube”*	*“green cube to the left front of cyan cylinder”*
Neural render		
Blender render		

Grounding Language on 3D visual feature representations

- Objects have regular sizes: object size is disentangled from the camera viewpoint
- Objects have 3D extent
- Objects do not interpenetrate in 3D: during iterative scene generation we can detect 3D intersection and continue sampling valid configurations
- Objects persist over time

Next steps

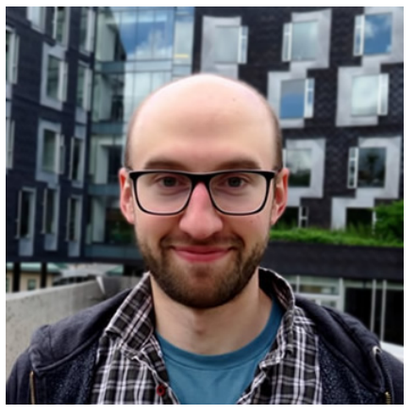
- Grounding action descriptions
- Use intuitive physics and dynamics beyond static spatial constraints

Thank you

Mihir

Fish Tung

Jyed Javed



Adam Harley

Max Sieb

- Embodied language grounding, arxiv
- Reward Learning from Narrated Demonstrations , Tung et al., CVPR 2018