A Visual Question Answering Study In Intuitive Physics

Tayfun ATEŞ, N17230653

Overview

- ✳Introduction
  - ✳ Problem Definition
  - ✳ Our Motivation
- ✳ Related Work
- ✳ Dataset
- ✳ Methodology
- ✳Experiments & Initial Results
- ✳Conclusion

Problem Definition - What is Intuitive Physics?

The knowledge underlying the human ability to understand the physical environment and interact with objects and substances that undergo dynamic state changes, making at least approximate predictions about how observed events will unfold.

*Intuitive Physics: Current Research and Controversies

Problem Definition - What is Intuitive Physics?

The knowledge underlying the human ability to understand the physical environment and interact with objects and substances that undergo dynamic state changes, making at least approximate predictions about how observed events will unfold.

*Intuitive Physics: Current Research and Controversies

Cognitive Science

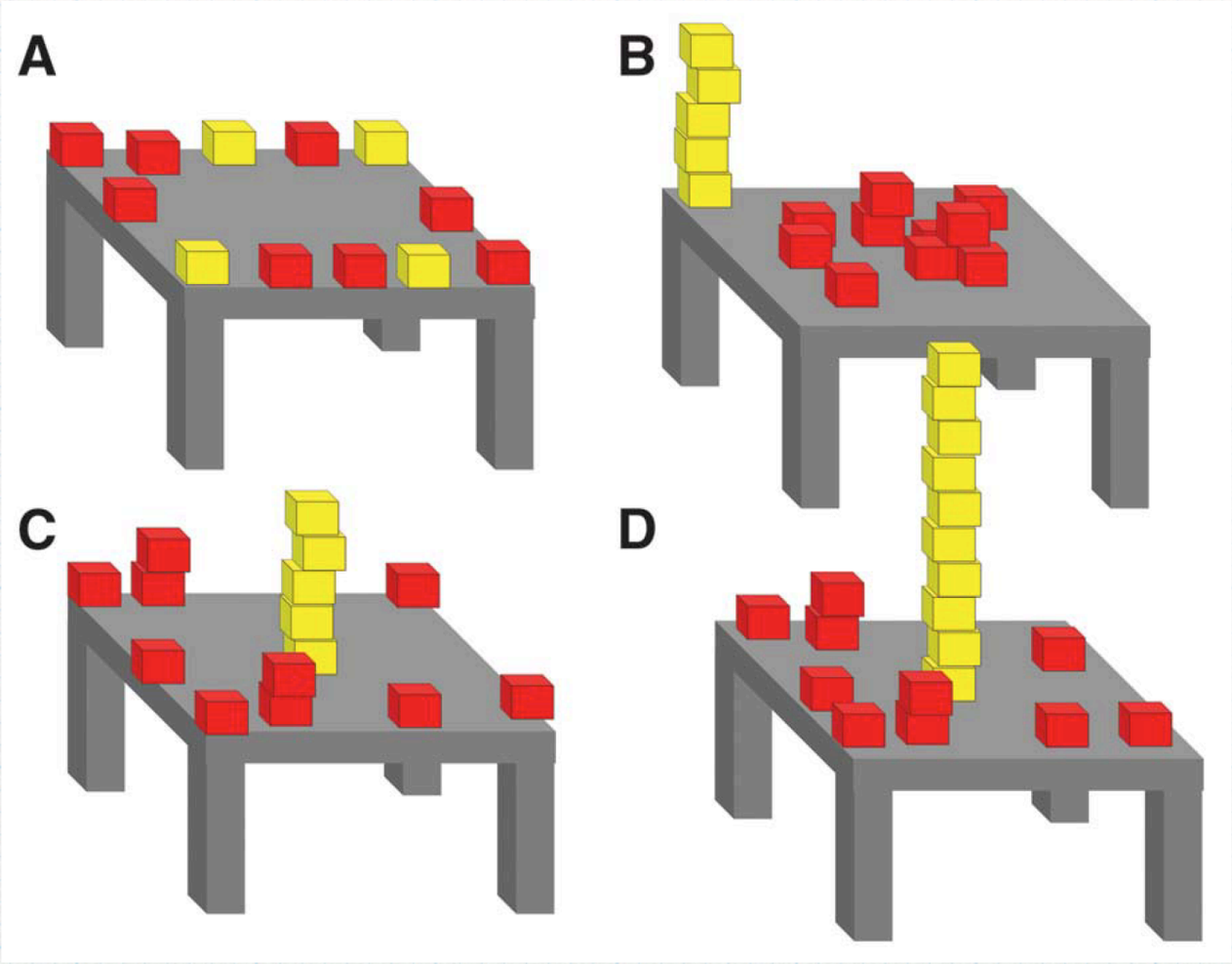
Problem Definition - What is Intuitive Physics?

The knowledge underlying the human ability to understand the physical environment and interact with objects and substances that undergo dynamic state changes, making at least approximate predictions about how observed events will unfold.

*Intuitive Physics: Current Research and Controversies

Cognitive Science and Machine Learning

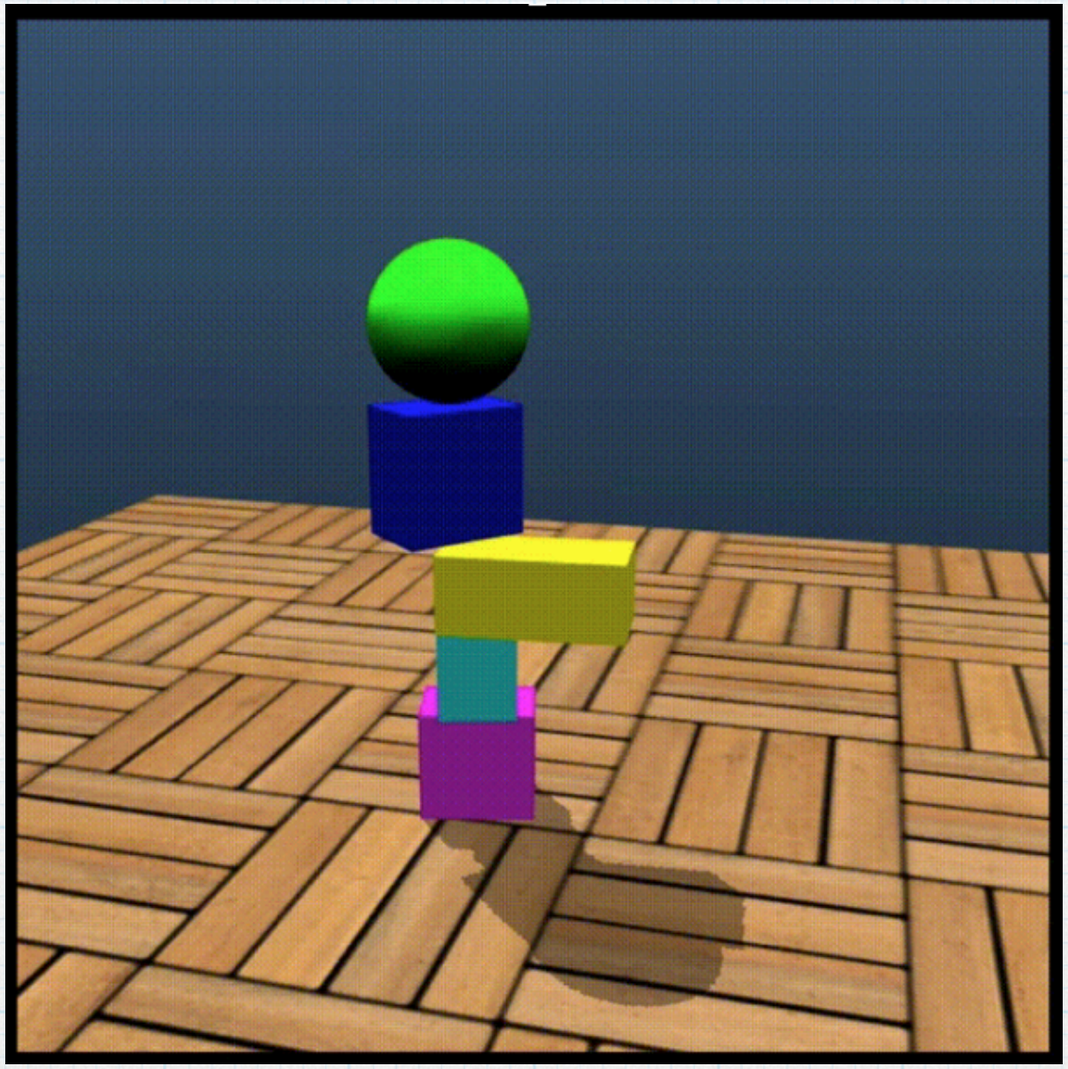
An Example From Cognitive Science



If the table in this scene is bumped so that one block falls off the table onto the floor, is it more likely to be a red or a yellow block?

*Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference

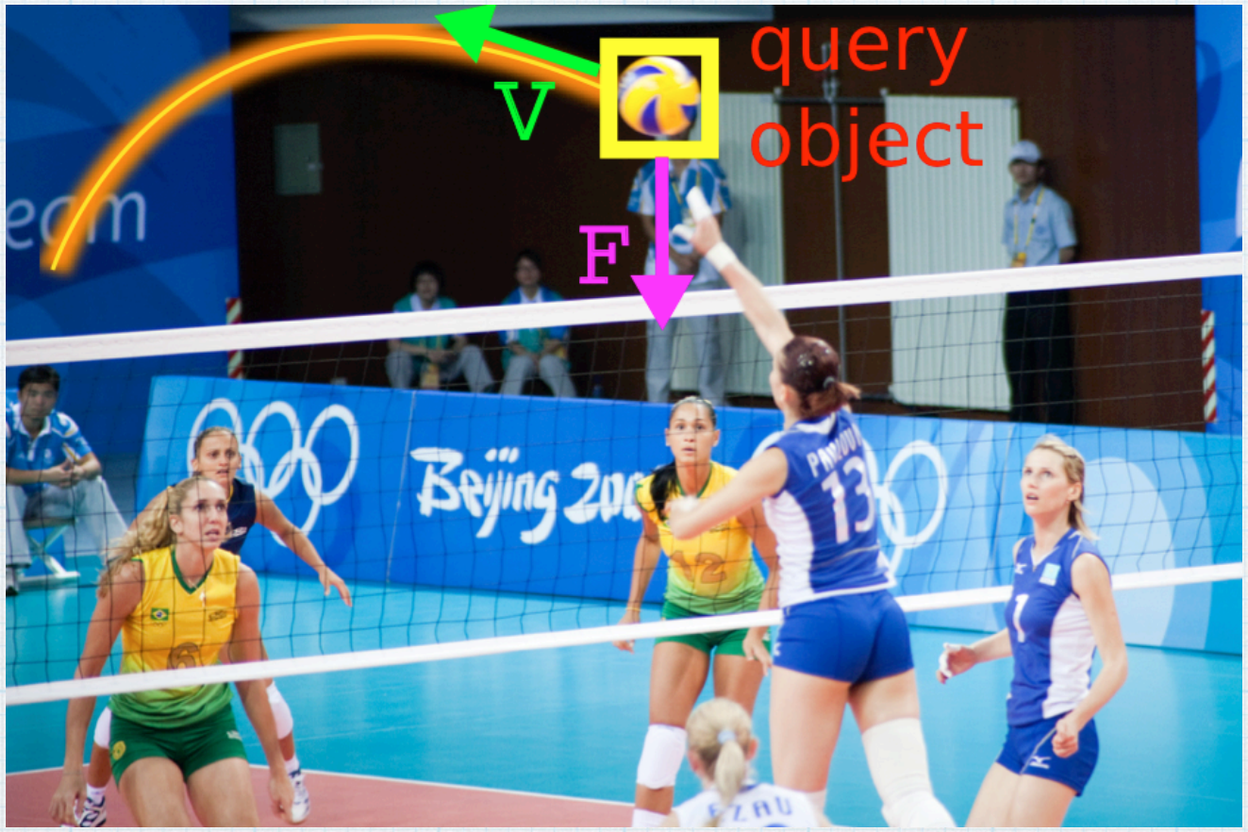
An Example From Computer Vision



Is the scene is stable or not?

*ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking

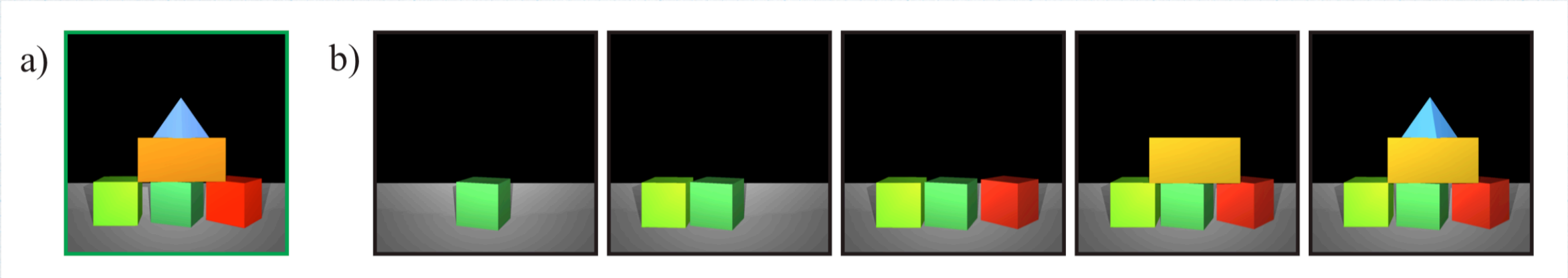
Another Example



Can you estimate the motion path of a specific object under seen in a single image under different forces?

*Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images

Another Example

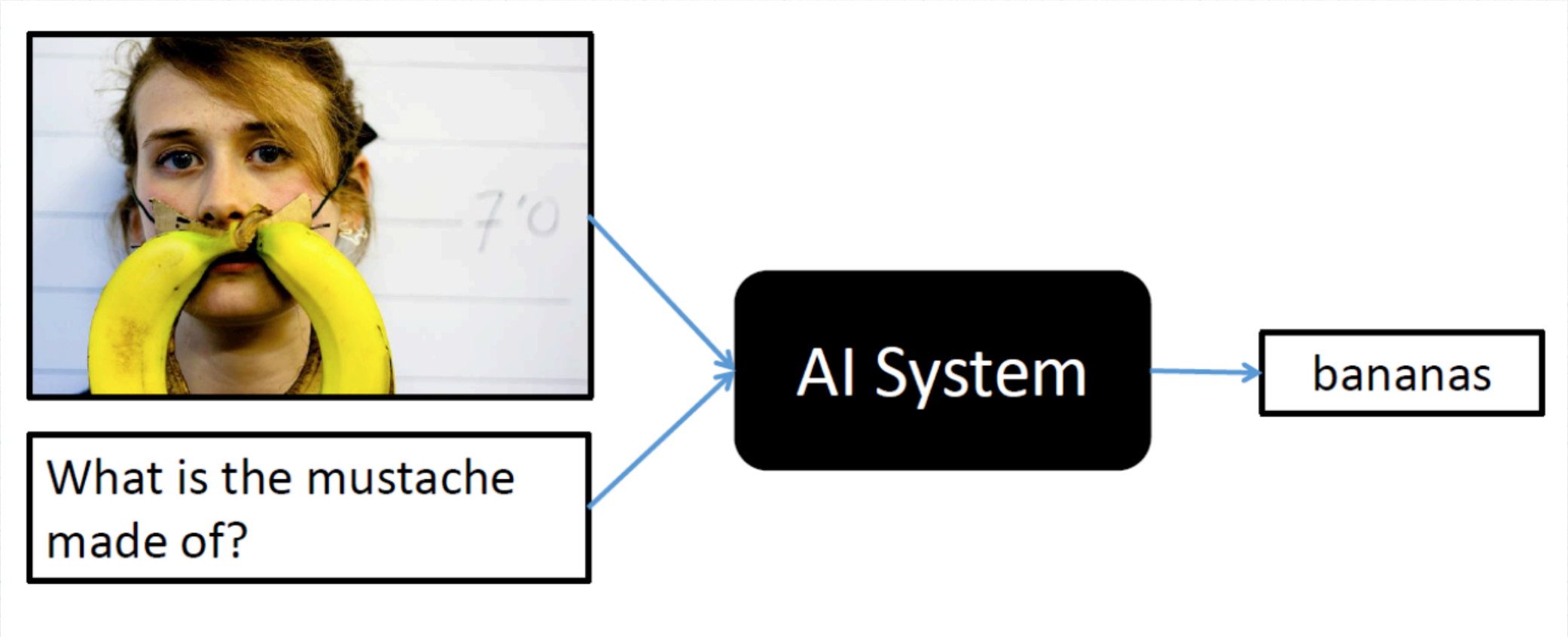


Given an initial observation can you predict how it is constructed?

*Reasoning About Physical Interactions With Object-Oriented Prediction and Planning

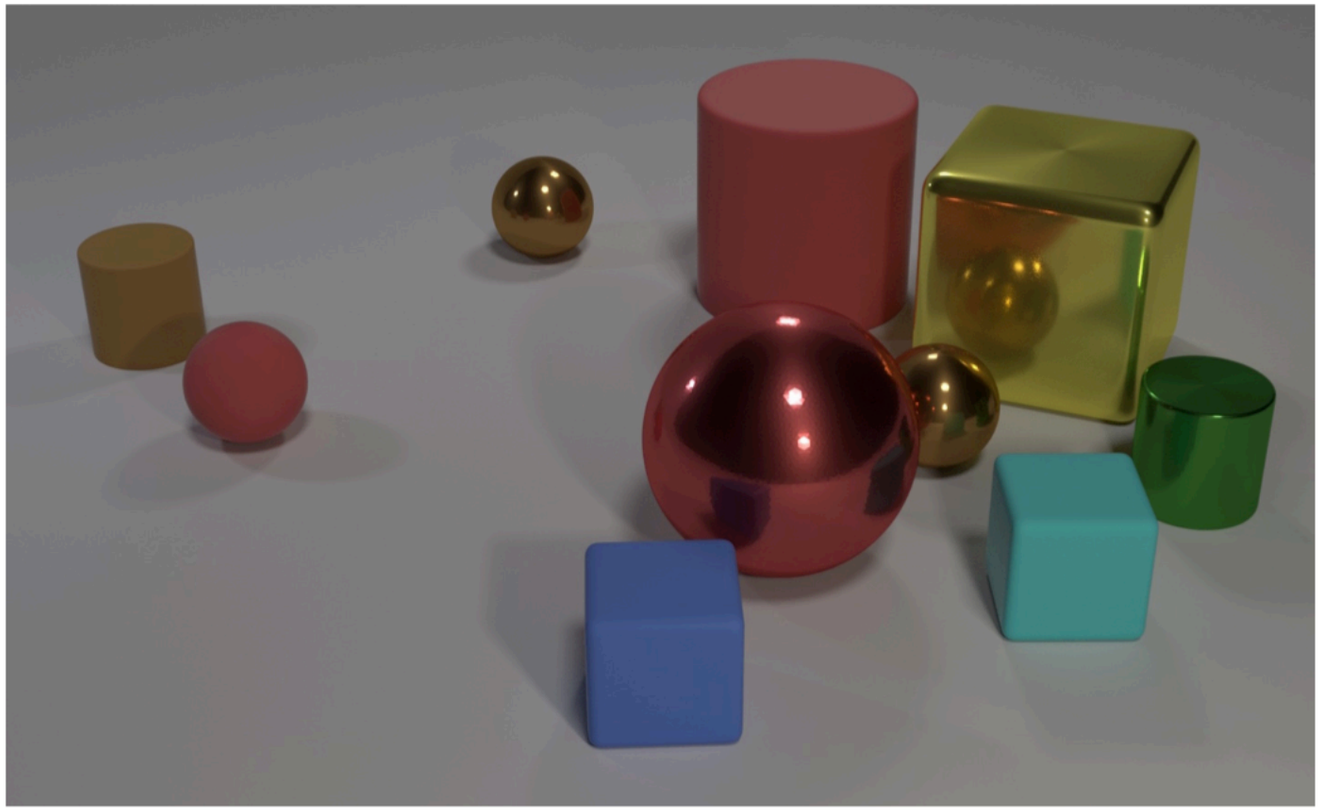
Visual Question Answering

- Can a model generate an answer to a given question based on an image?



*Vqa: Visual Question Answering

CLEVR



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

*CLEVR: A Diagnostic Dataset For Compositional Language and Elementary Visual Reasonings

What Are Missing Components?

- From VQA point of view, mostly a single state is under consideration and future states are ignored.

What Are Missing Components?

- From VQA point of view, mostly a single state is under consideration and future states are ignored.

- From intuitive physics point of view, model is expected simulate a single action for which it does not have to understand what the action is (i.e. for example from a textual domain) and its dynamics.

What Are Missing Components?

- From VQA point of view, mostly a single state is under consideration and future states are ignored.

- From intuitive physics point of view, model is expected simulate a single action for which it does not have to understand what the action is (i.e. for example from a textual domain) and its dynamics.

- In that sense, our aim is to integrate VQA to intuitive physics where the model has to understand the question, reason about the scene and then simulate the action to give an answer the question.

What Are Missing Components?

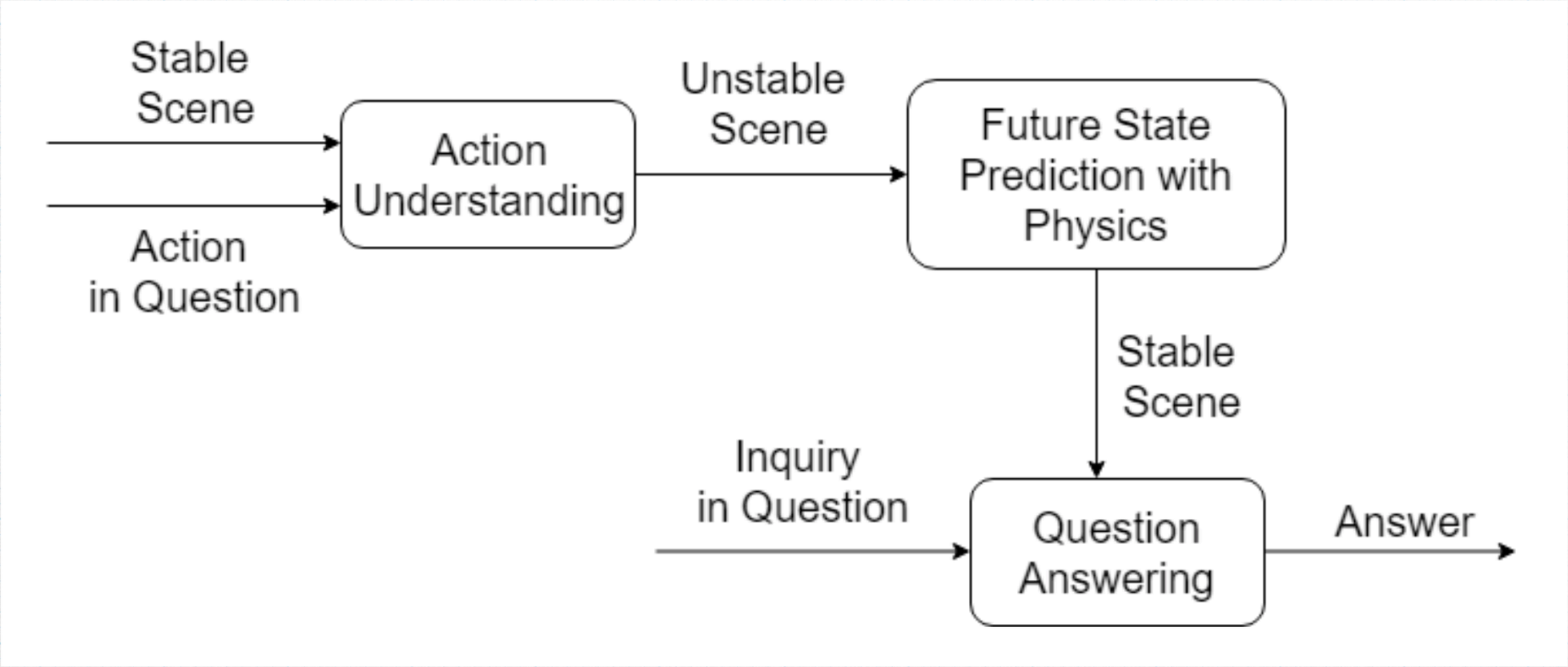
- From VQA point of view, mostly a single state is under consideration and future states are ignored.

- From intuitive physics point of view, model is expected simulate a single action for which it does not have to understand what the action is (i.e. for example from a textual domain) and its dynamics.

- In that sense, our aim is to integrate VQA to intuitive physics where the model has to understand the question, reason about the scene and then simulate the action to give an answer the question.

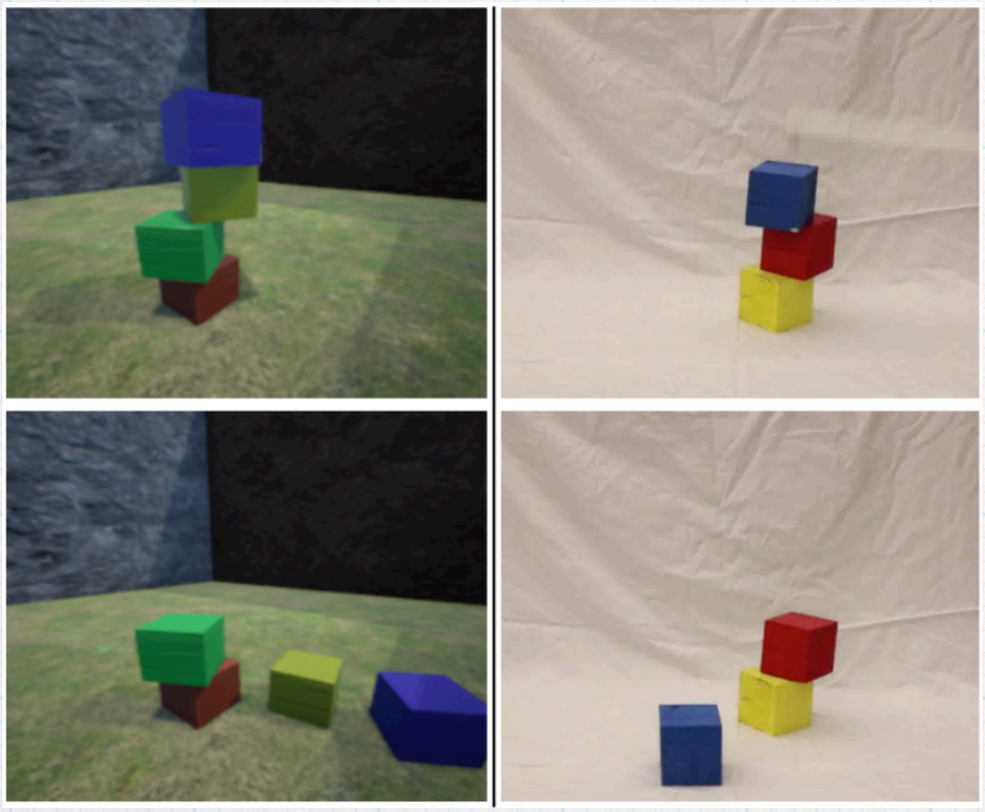
- Besides the model that is capable of simulating such intuitions, we will output a VQA dataset for further improvements.

Method Overview



Related Work (1/5)

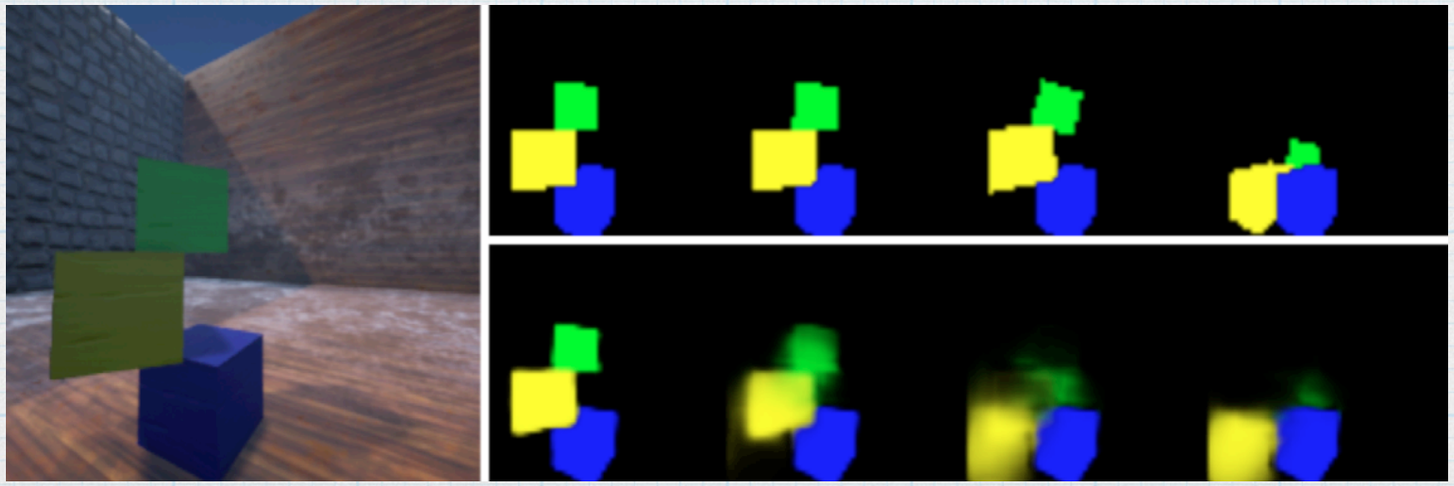
✳Learning Physical Intuition of Block Towers by Example - PhysNet



✳Examines the physical intuition capacity of DNNs

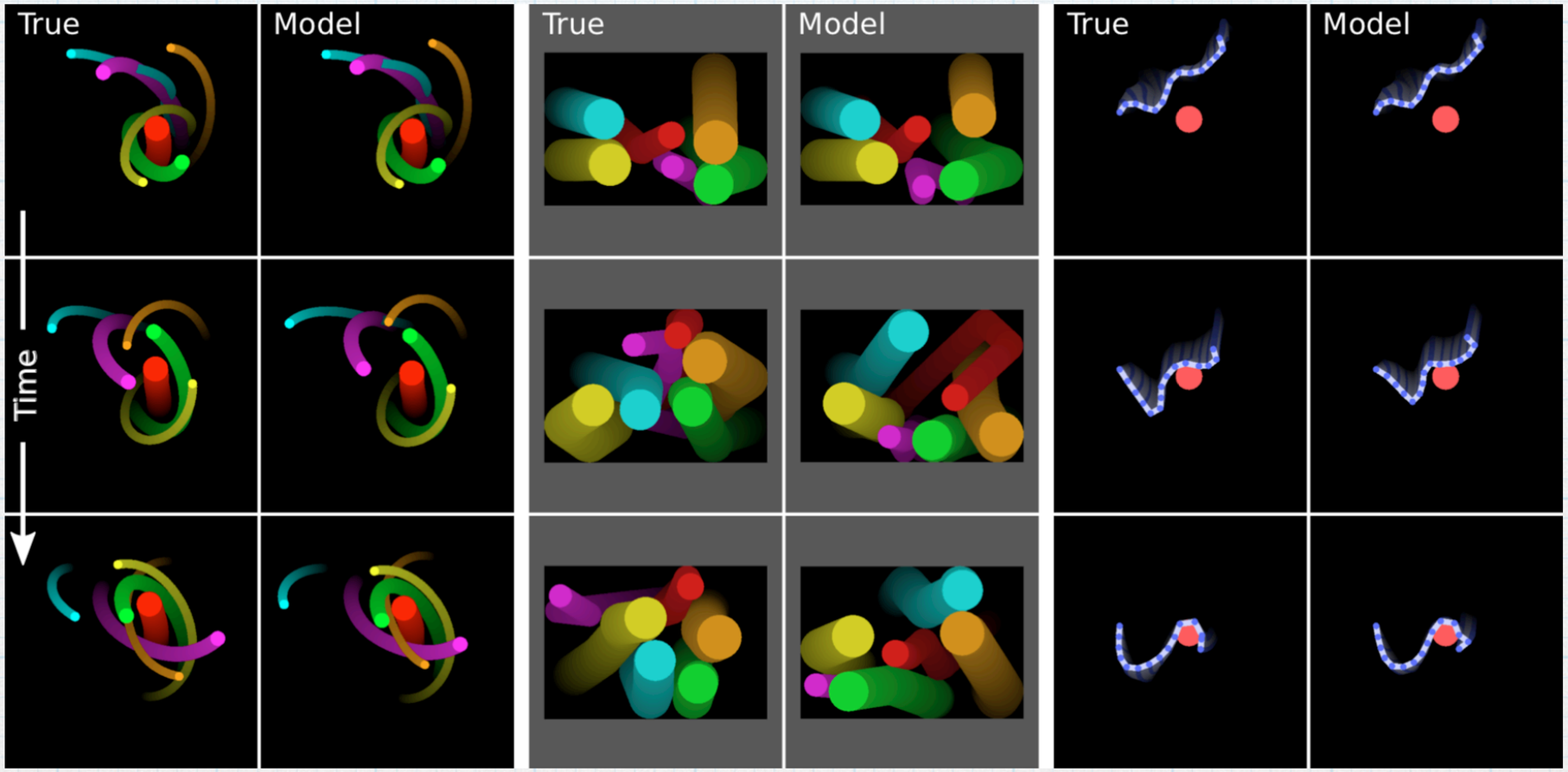
✳Given the initial stack of blocks:

1. Will the blocks fall or not?
2. Where will be the blocks end up?



Related Work (2/5)

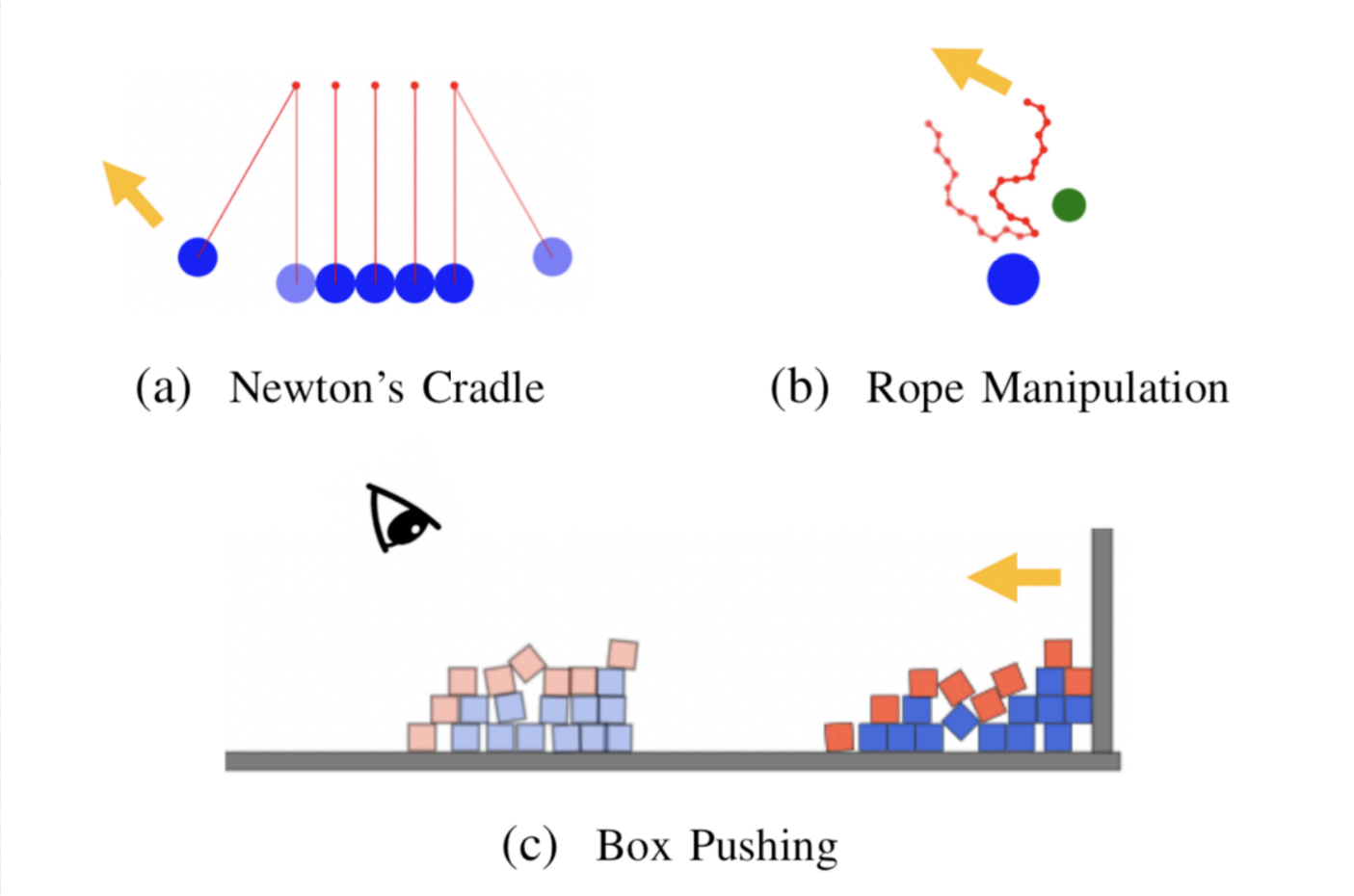
✳Interaction Networks for Learning about Objects, Relations and Physics



✳ Tries to estimate physical interactions between objects using Graph Networks (Static, Single Level, Complete Graph)

Related Work (3/5)

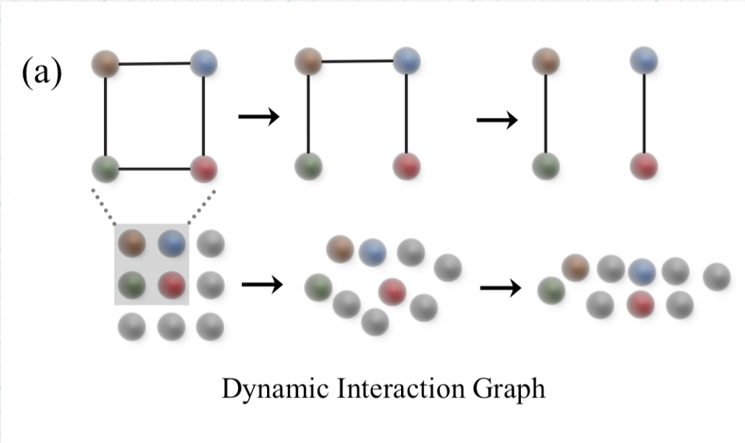
✳Interaction Networks for Learning about Objects, Relations and Physics

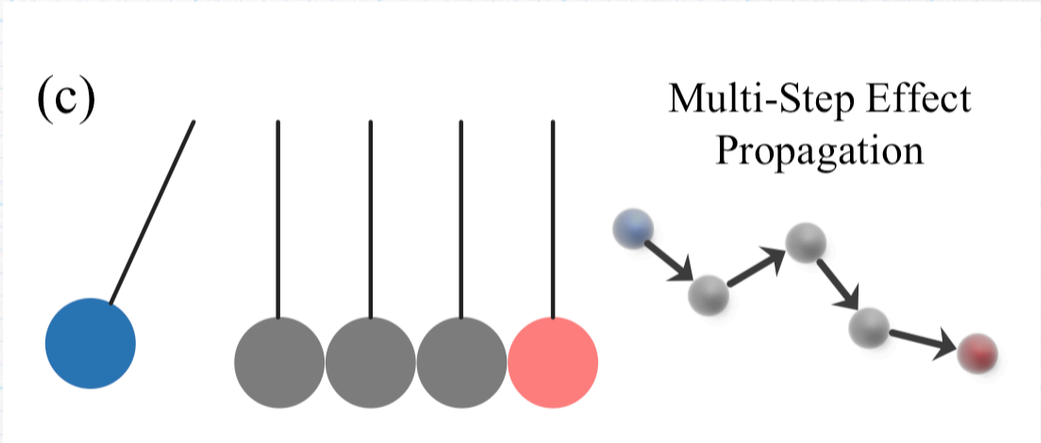


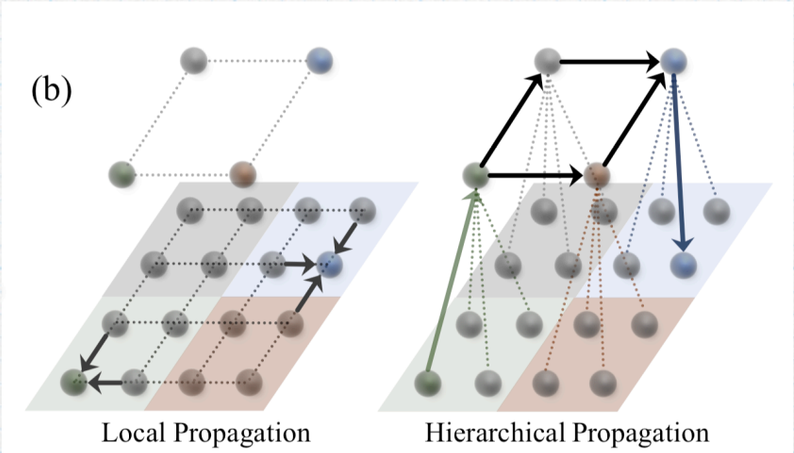
✳At a single time-step t, the effect of an interaction is propagated to other objects

Related Work (4/5)

✳Flexible Neural Representation for Physics Prediction





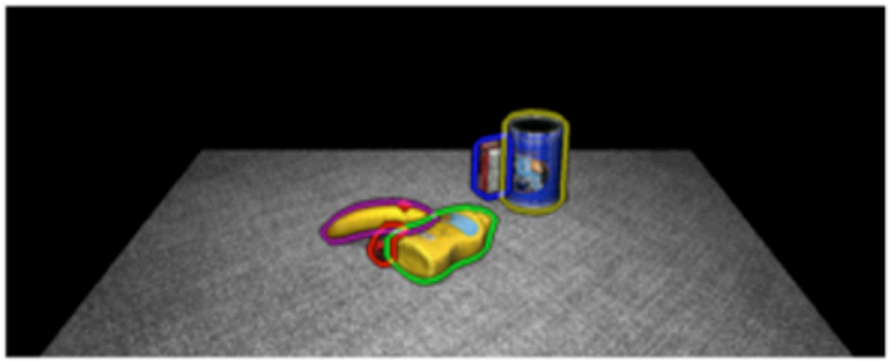


✳Combines three features (dynamic edges, multi step propagation and hierarchical representation)

✳Can represent deformable objects and fluids as well as rigid objects

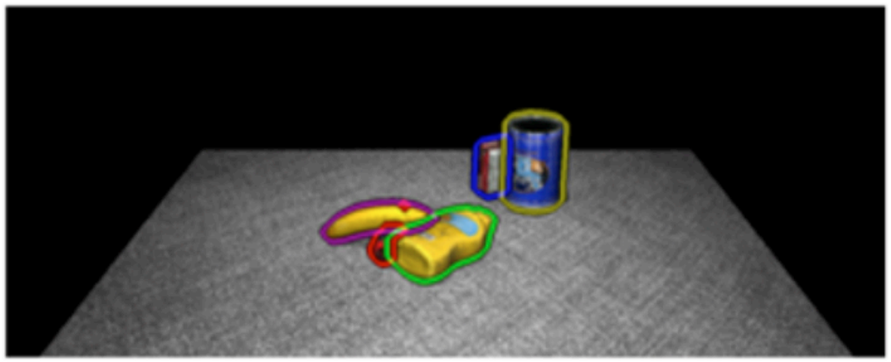
Related Work (5/5)

✳Answering Visual What-If Questions: From Actions to Predicted Scene Descriptions

 2) "the robot spins the screw driver in anti-clowise direction" — what happens to the mustard container?	**Ground Truth** "the screw driver pushes the mustard container" **Human Baseline** "the mustard conatainer moves a little due to the impact of spinning screw driver" **Hybrid Model Prediction** "the screw driver pushes the mustard container" **Data-Driven Prediction** "the mustard container shakes a little from the impact"

Related Work (5/5)

- Answering Visual What-If Questions: From Actions to Predicted Scene Descriptions

 2) "the robot spins the screw driver in anti-clowise direction" — what happens to the mustard container?	**Ground Truth** "the screw driver pushes the mustard container" **Human Baseline** "the mustard conatainer moves a little due to the impact of spinning screw driver" **Hybrid Model Prediction** "the screw driver pushes the mustard container" **Data-Driven Prediction** "the mustard container shakes a little from the impact"

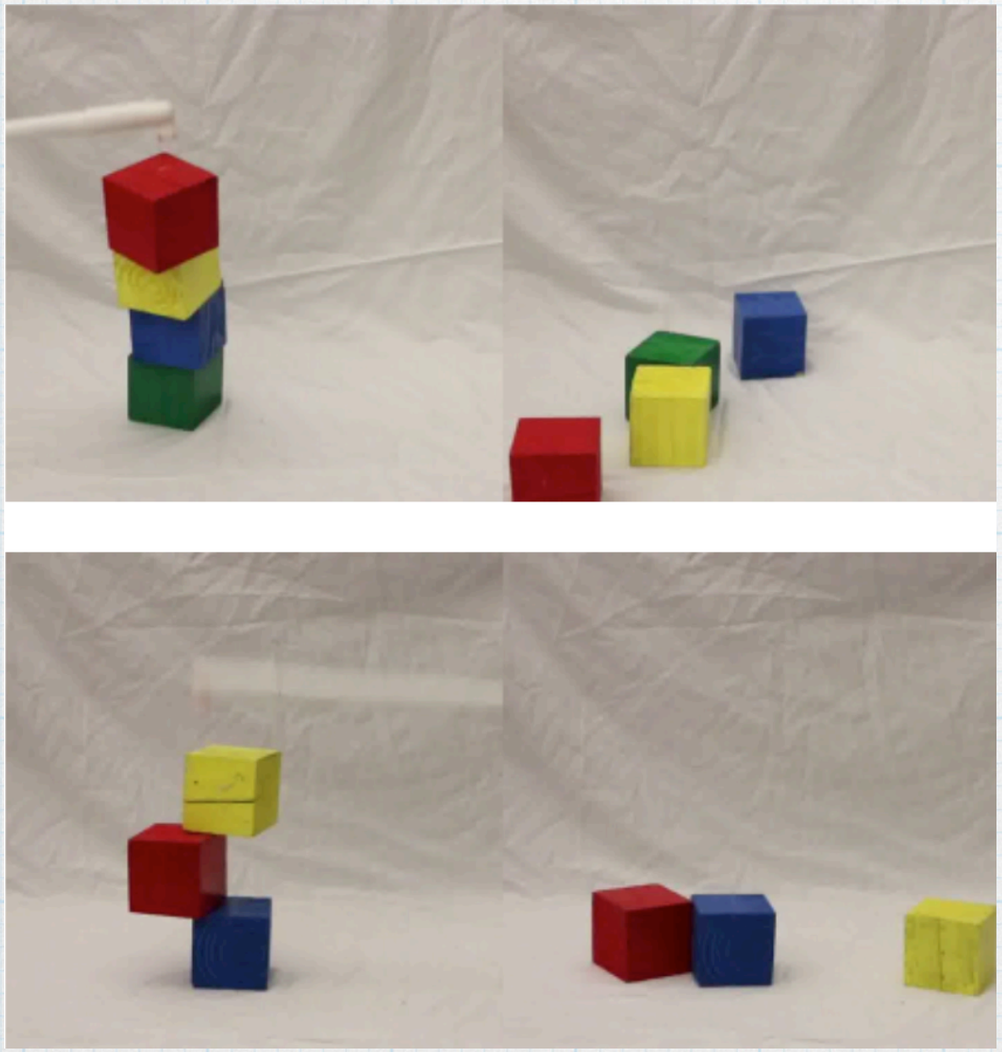
- Limited set of actions and questions (not structured as in CLEVR)

- Uses a physics engine in test time (hybrid model)

Dataset

- We are generating our own data

- In the meantime, we have obtained some initial baseline scores using real images of wooden cubes from PhysNet



- Manually stacked towers of 2, 3 and 4 four blocks

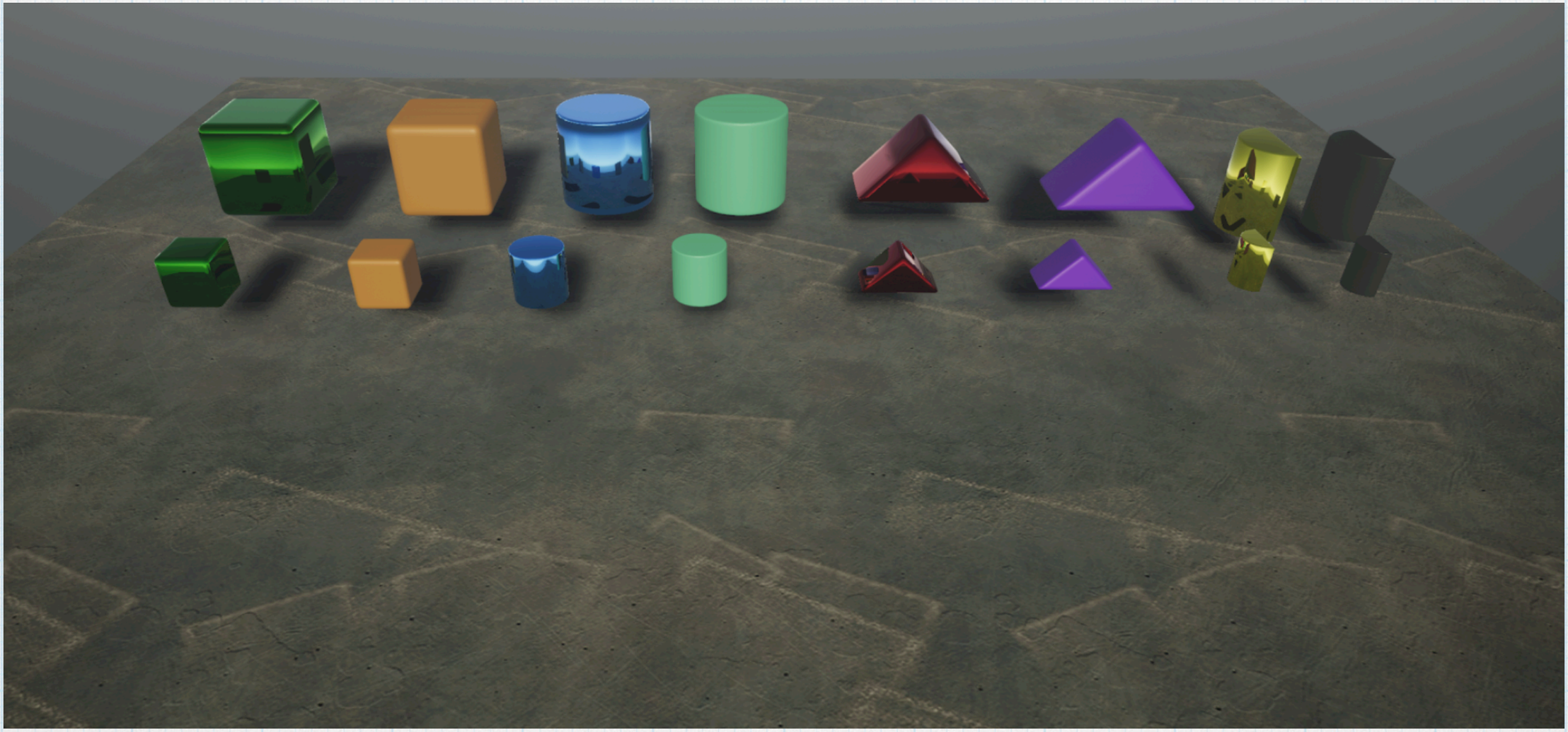
- A white stick holds the tower, and then it is rapidly lifted upwards

- A total of 493 scenes, balanced between stable/unstable configurations

- We are only interested in first and the last frames

*Learning physical intuition of block towers by example.

Dataset Generation



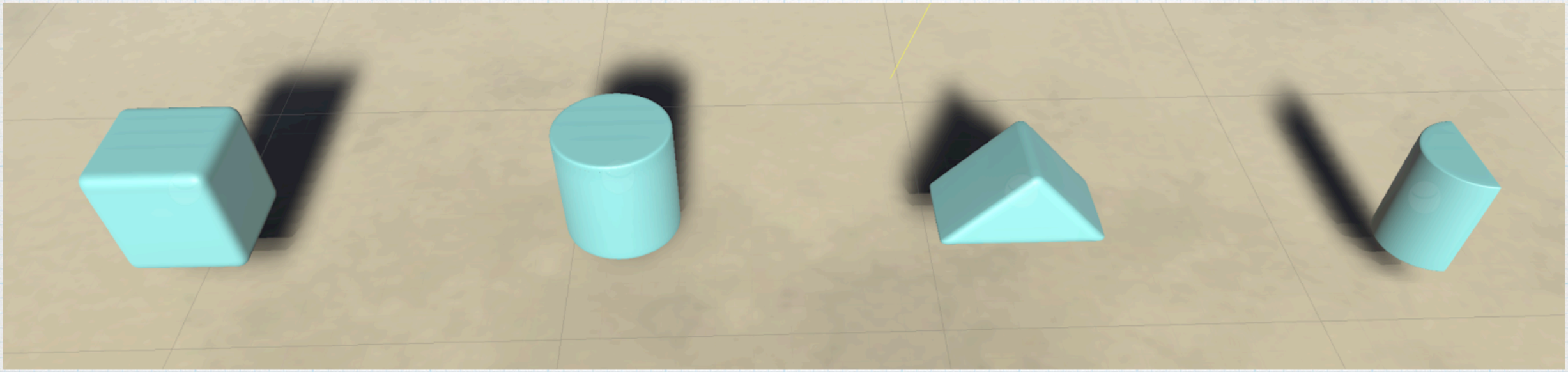
Tools

✳For Simulation:

✳For Mesh Manipulation:

Meshes

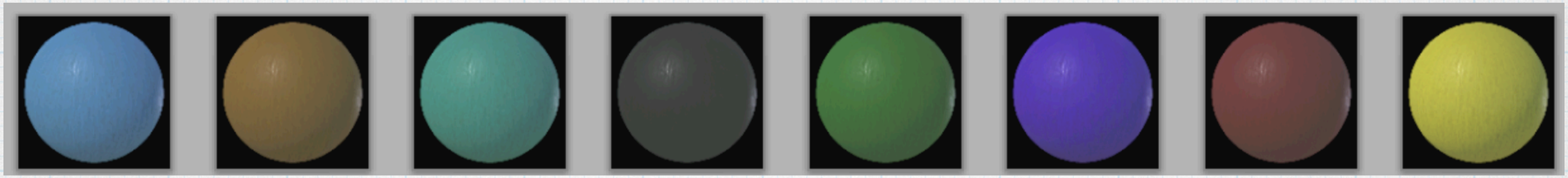
- All are smoothly generated



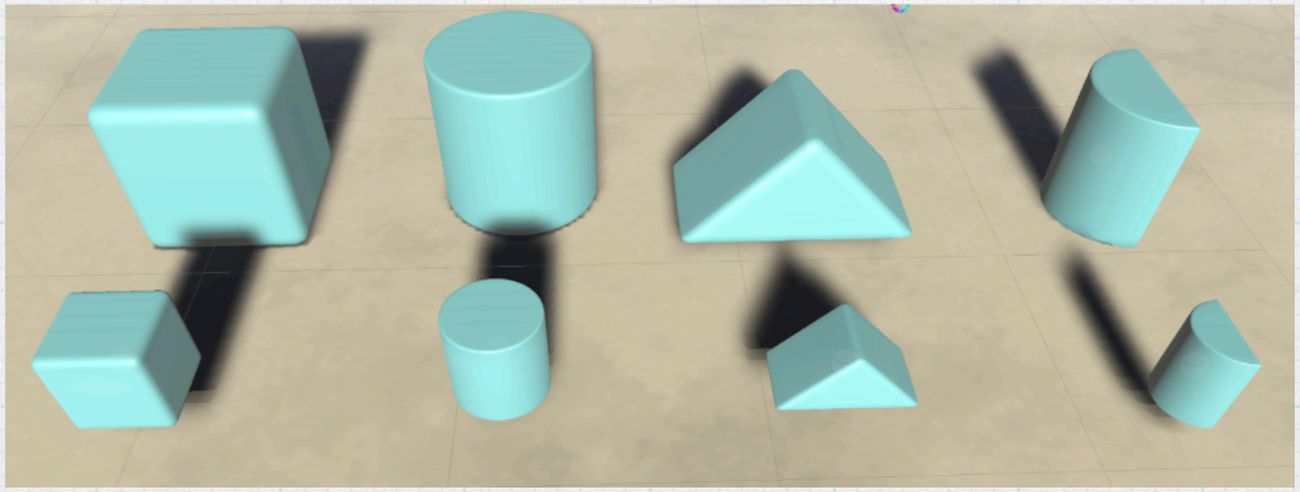
- Cube, Cylinder, Triangle, Half Cylinder

Mesh Properties (Color, Size)

✳8 different colors: Blue, brown, cyan, gray, green, purple, red, yellow

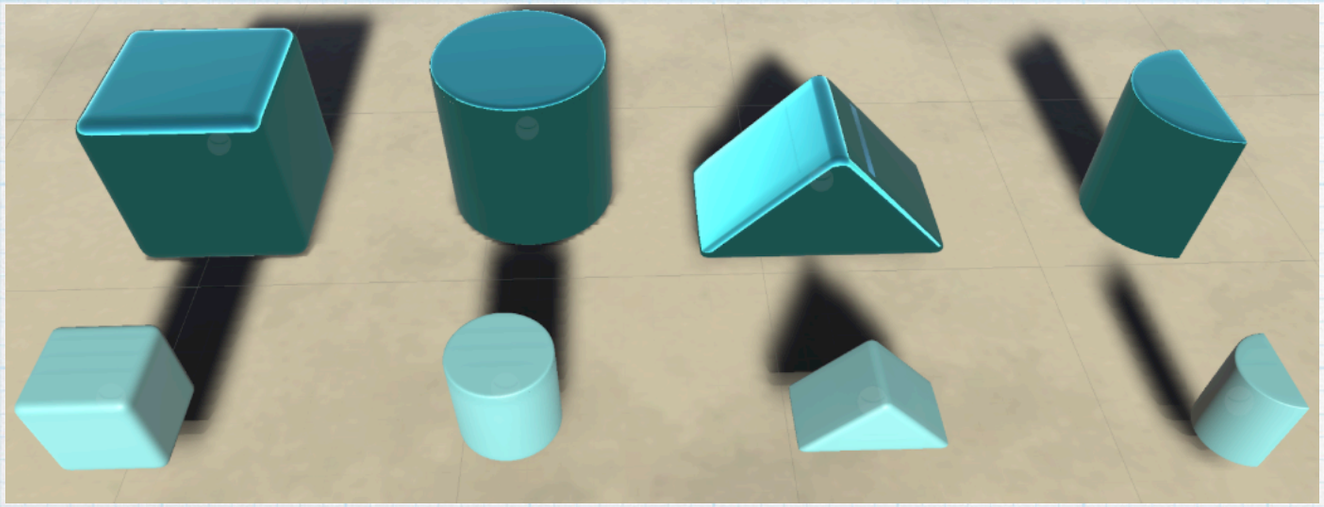


✳2 different Sizes

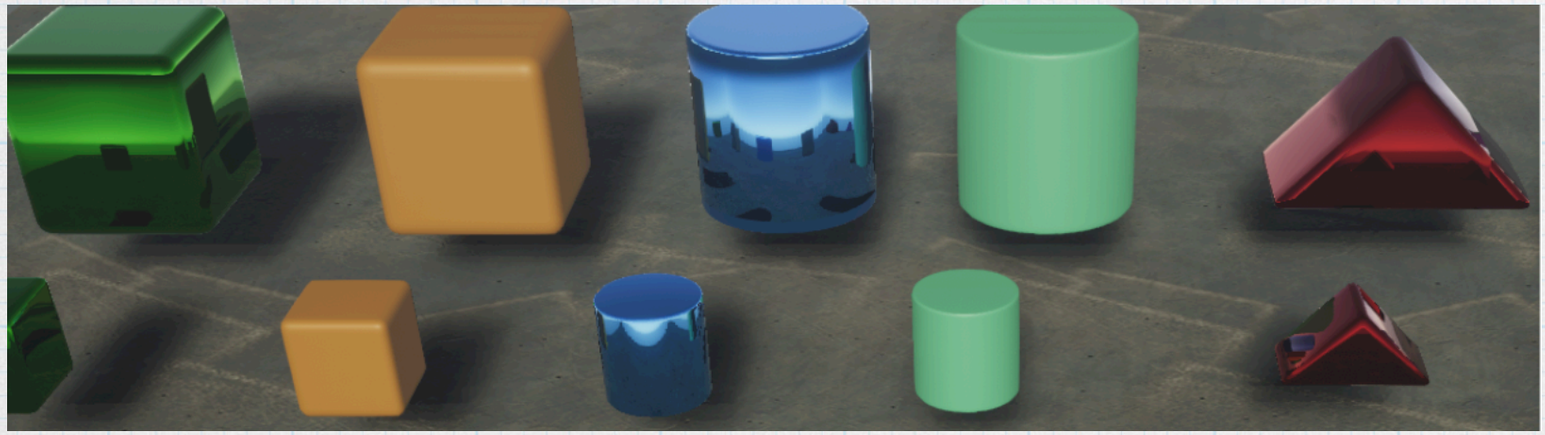


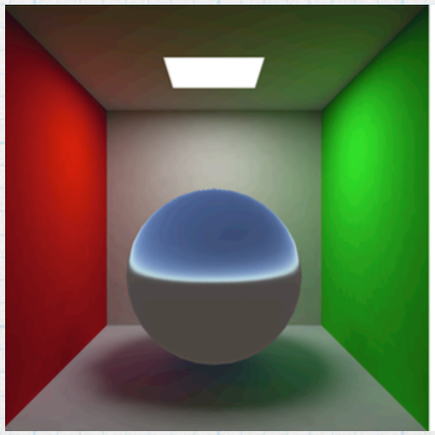
Mesh Properties (Materials)

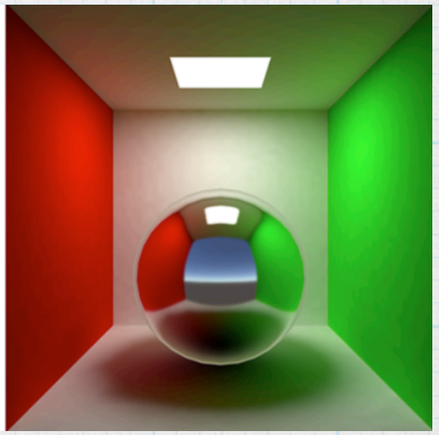
✳2 different materials: Metal, Rubber



✳Metal Reflections: Implemented using Unity's reflection probes.

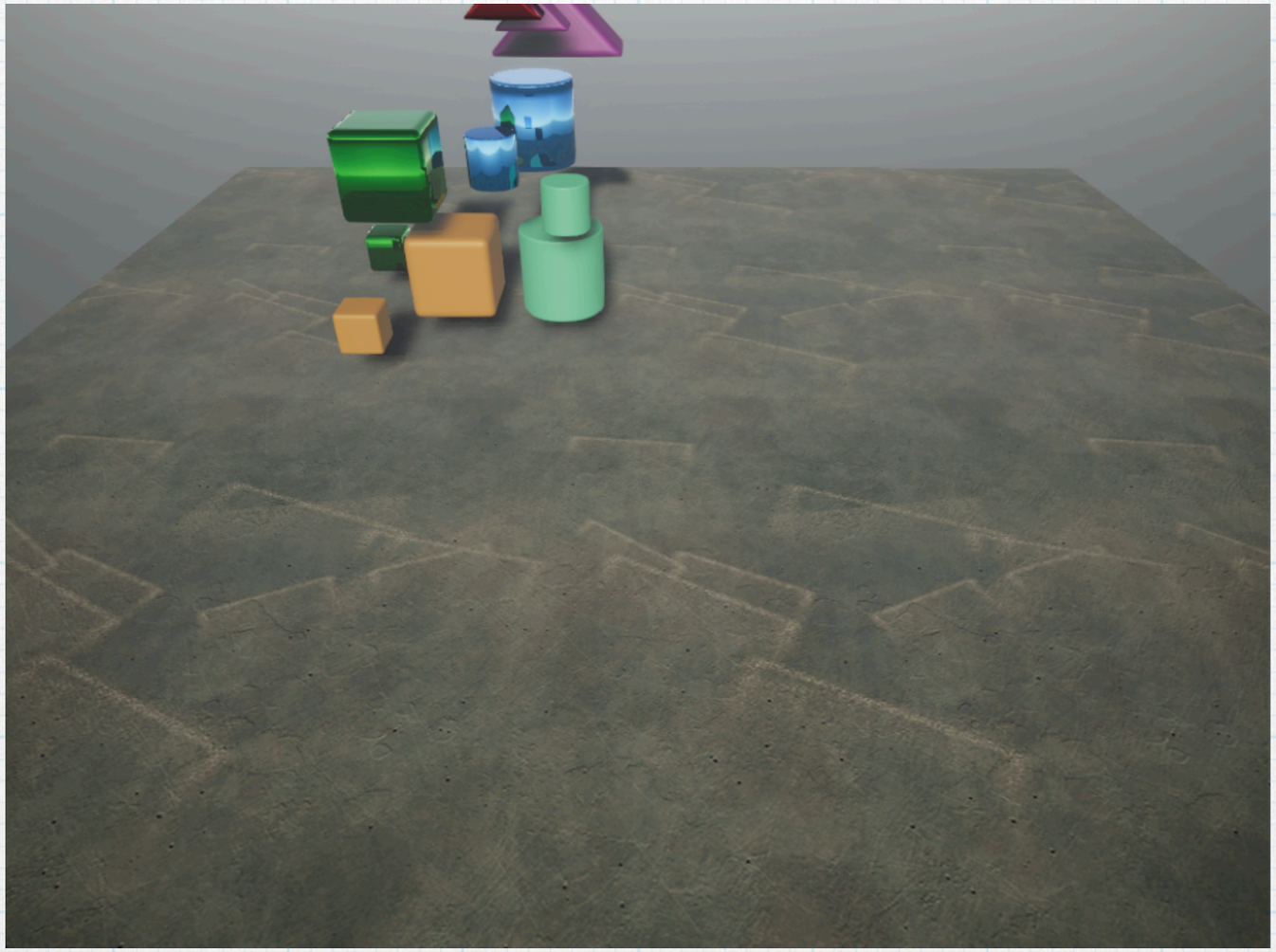






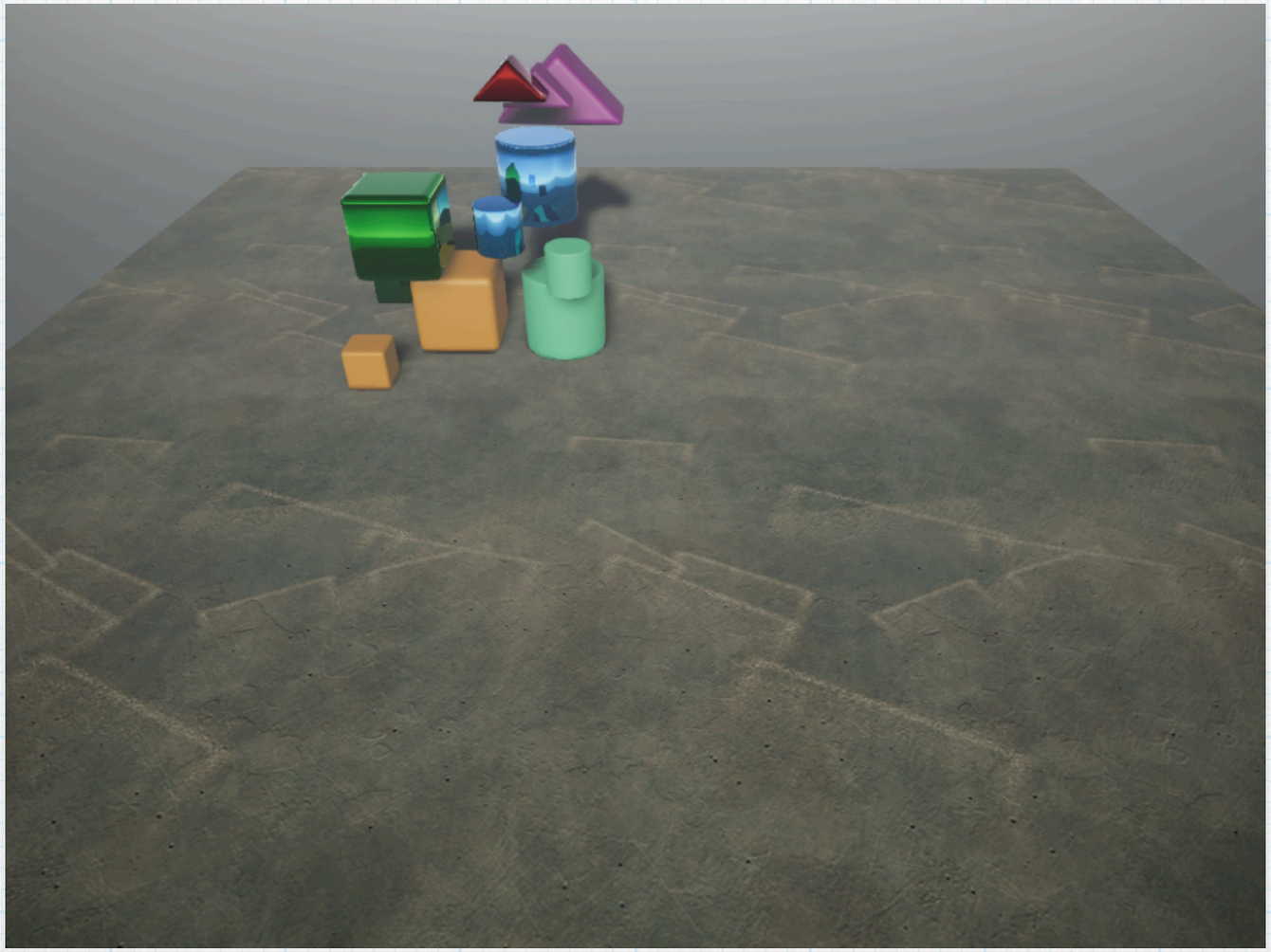
Simulation - Obtaining Initial Stable Configuration

- Initial Unstable



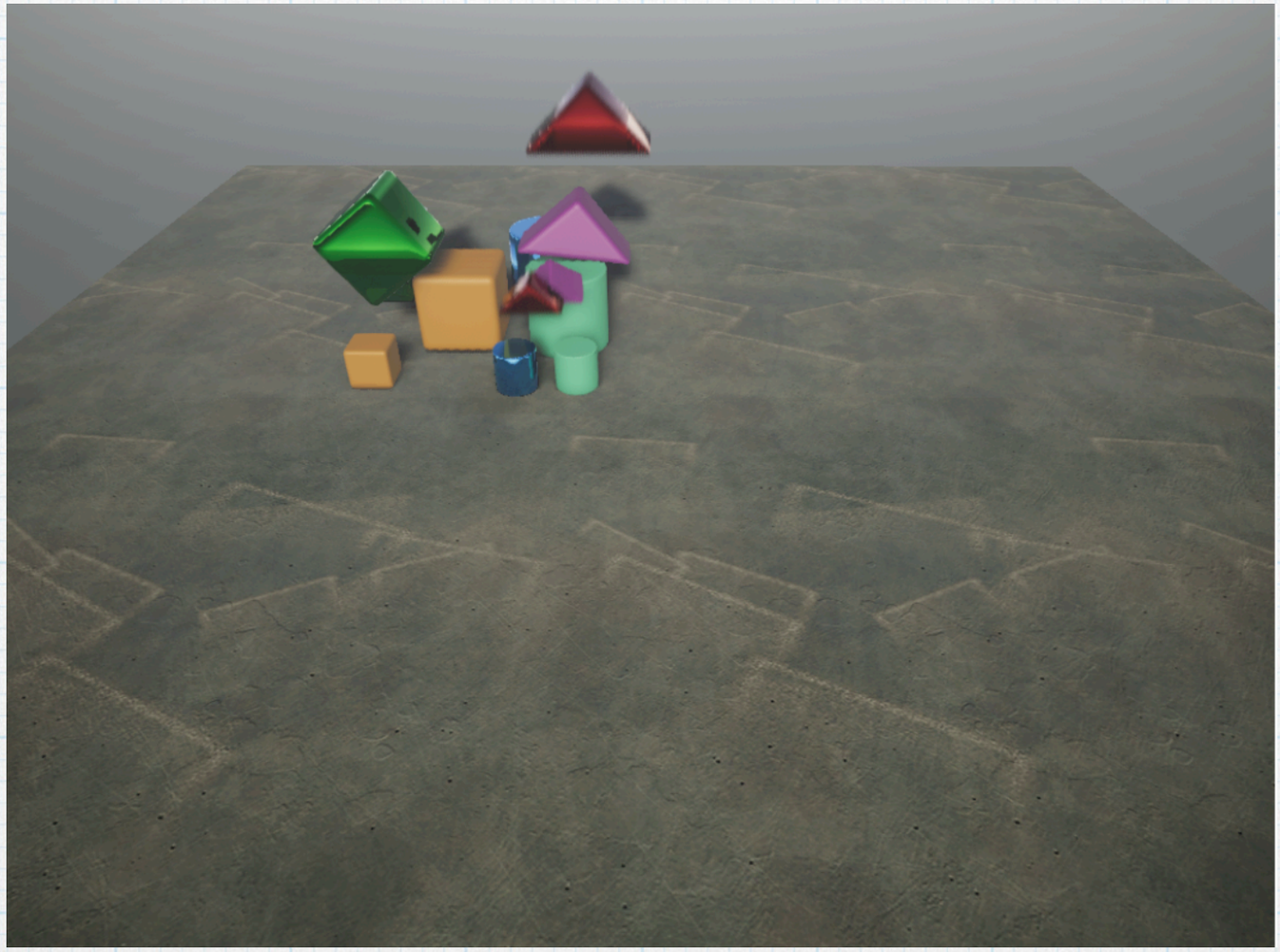
Simulation - Obtaining Initial Stable Configuration

- 20 Frames Later...



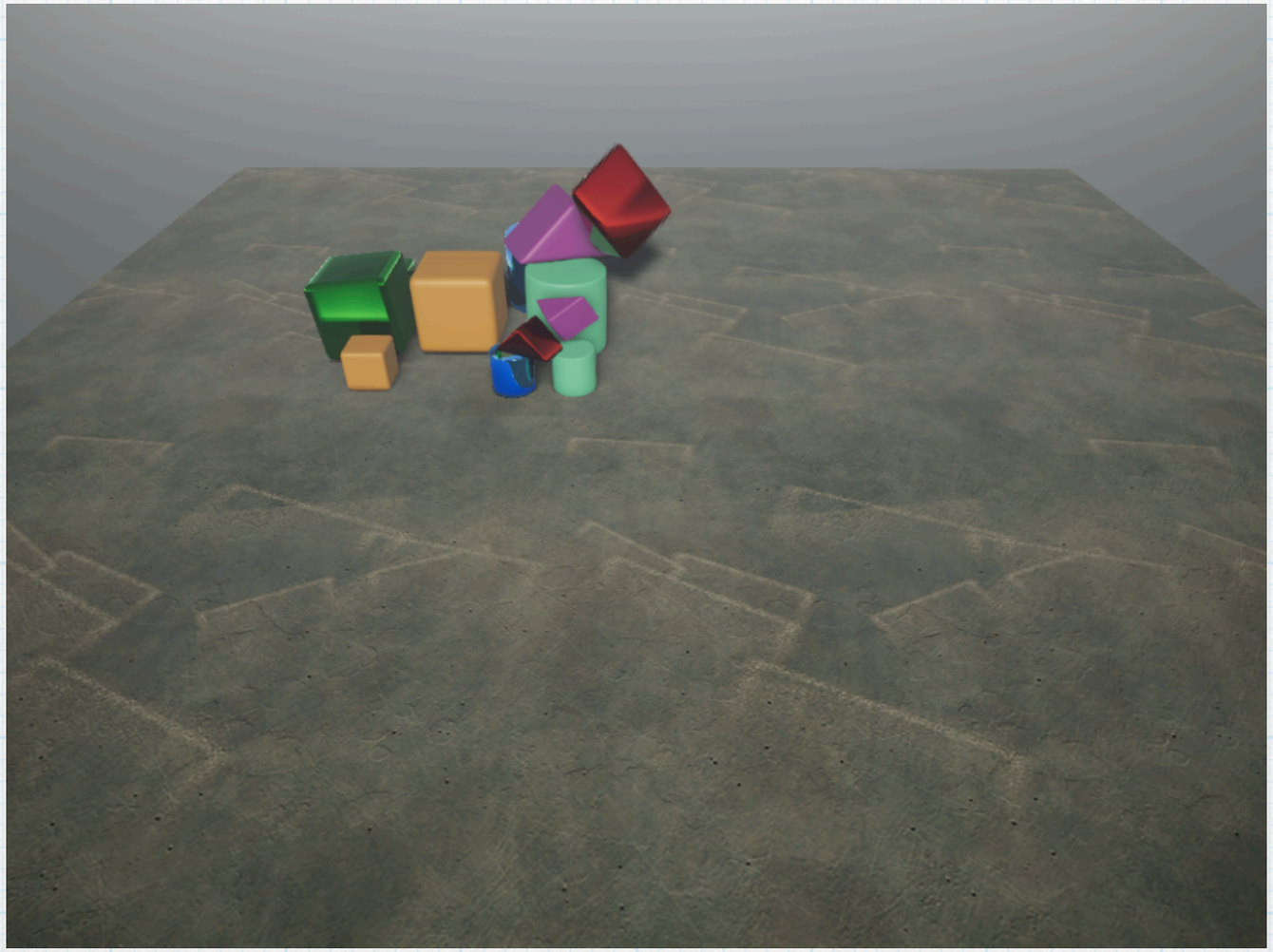
Simulation - Obtaining Initial Stable Configuration

- 40 Frames Later...



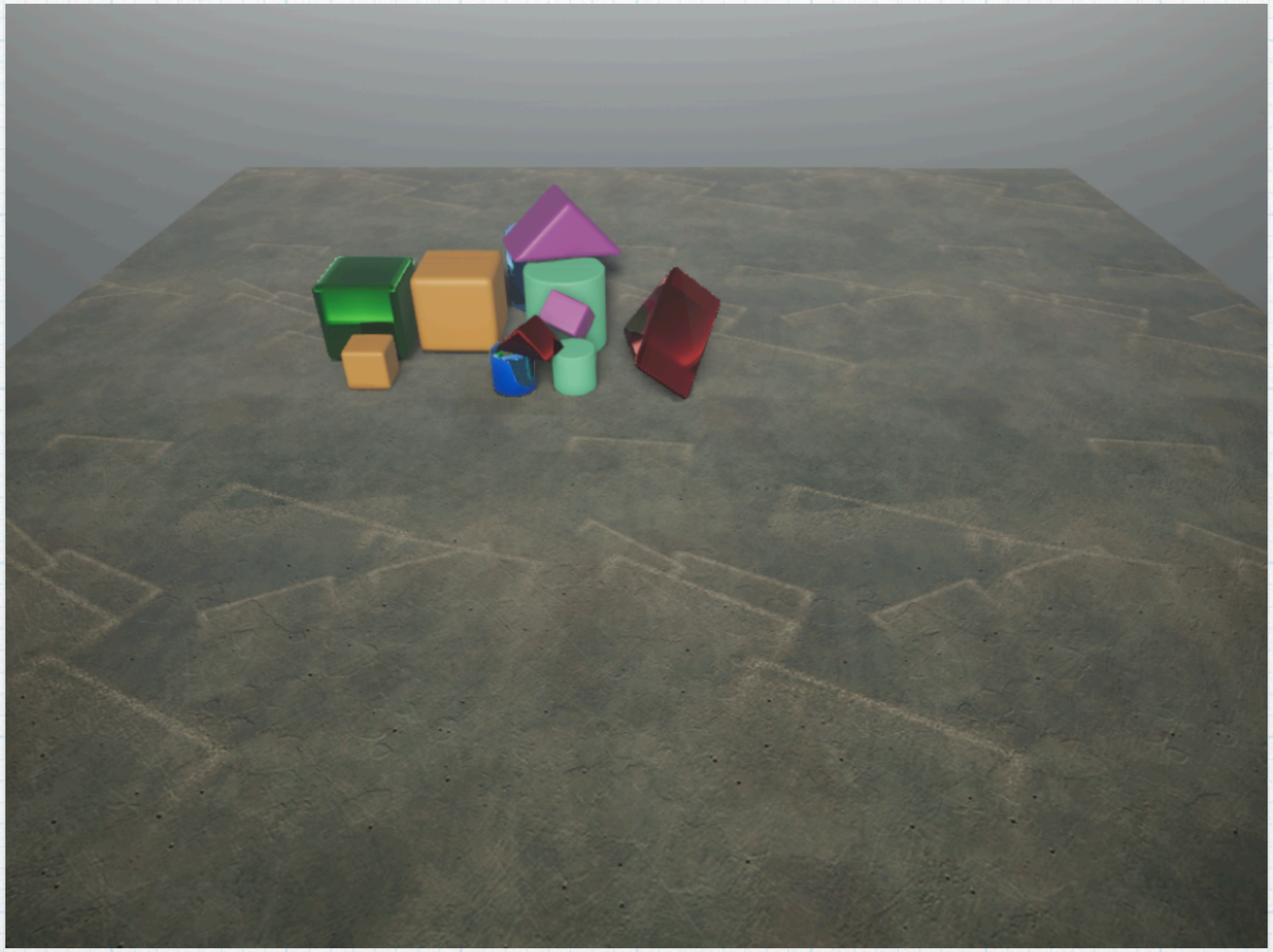
Simulation - Obtaining Initial Stable Configuration

- 60 Frames Later...



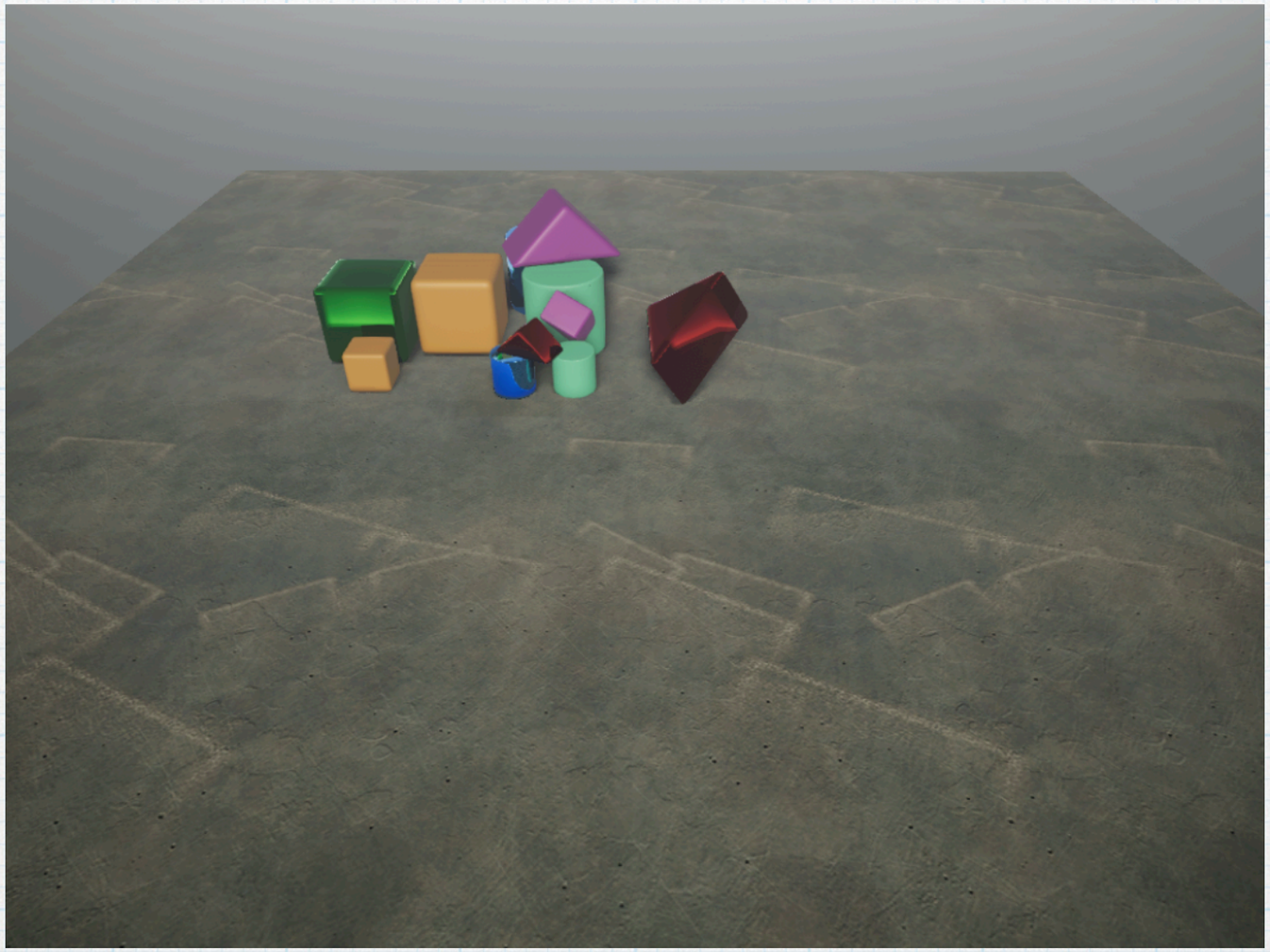
Simulation - Obtaining Initial Stable Configuration

- 80 Frames Later...



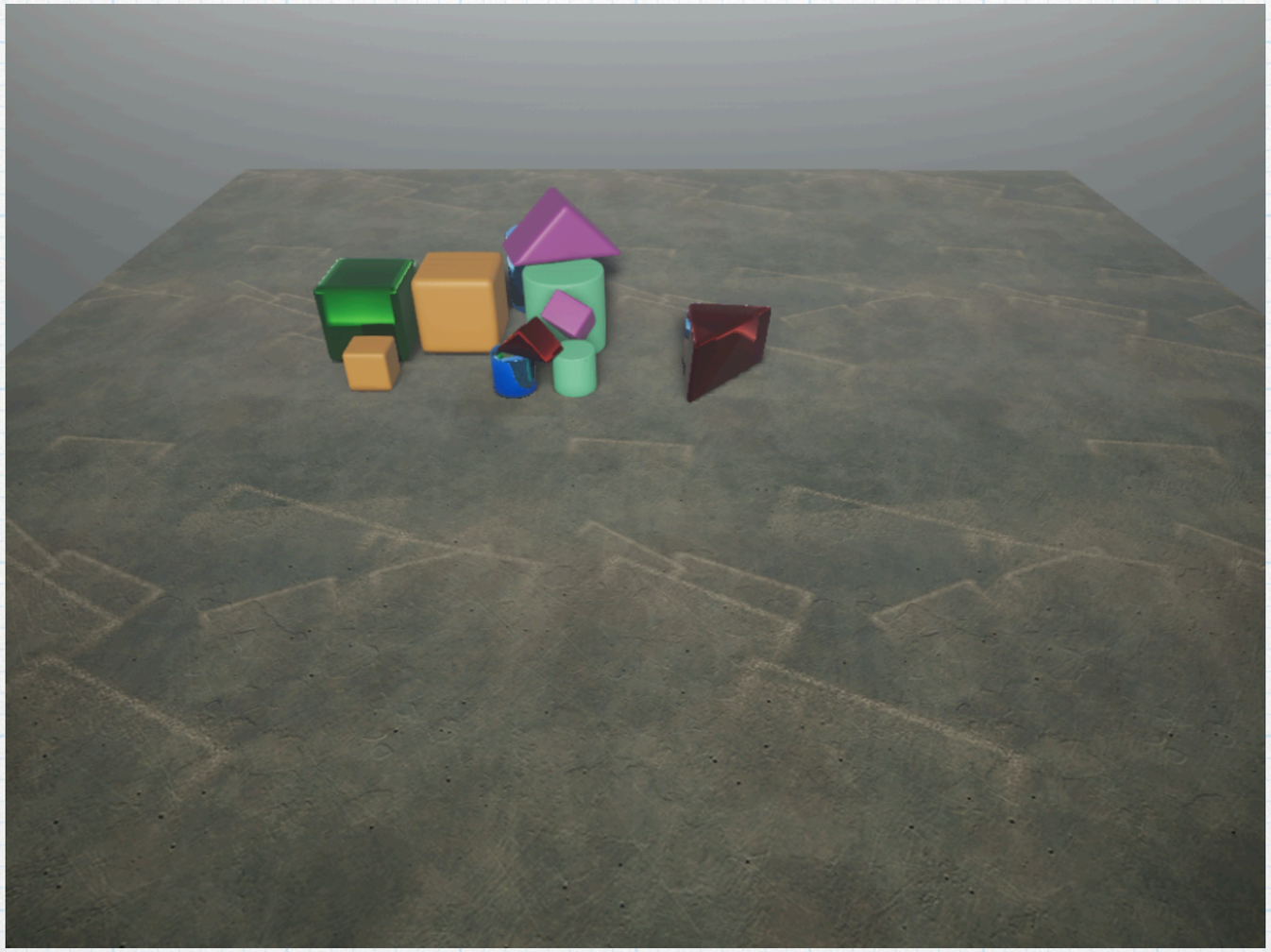
Simulation - Obtaining Initial Stable Configuration

- 100 Frames Later...

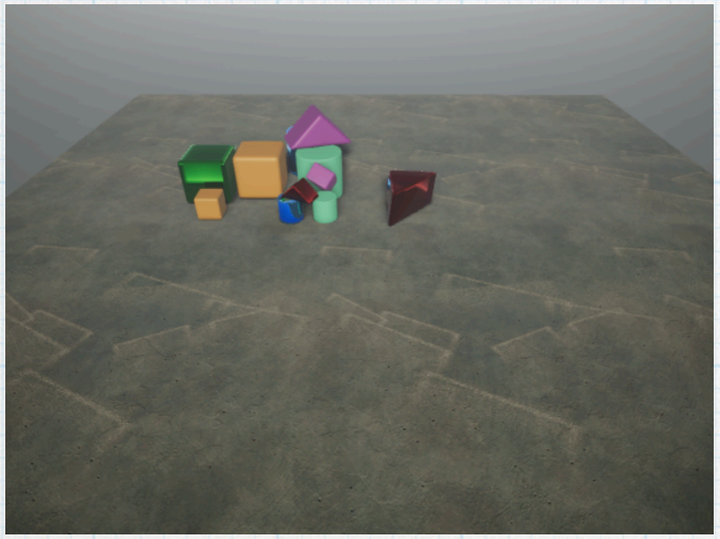
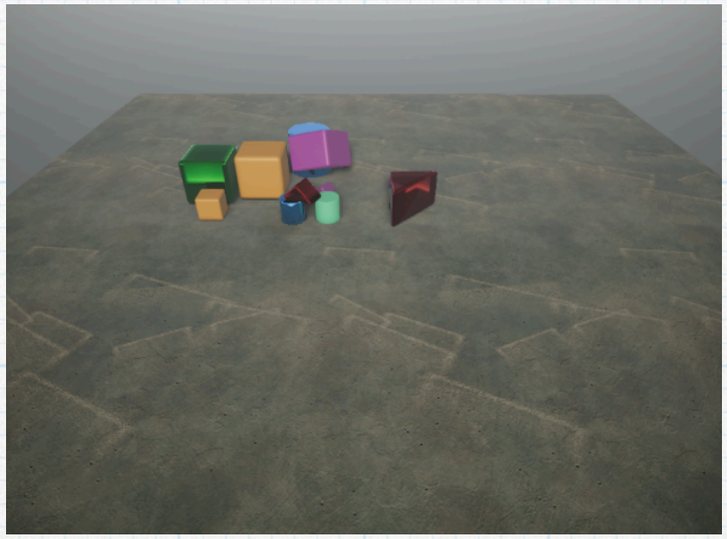
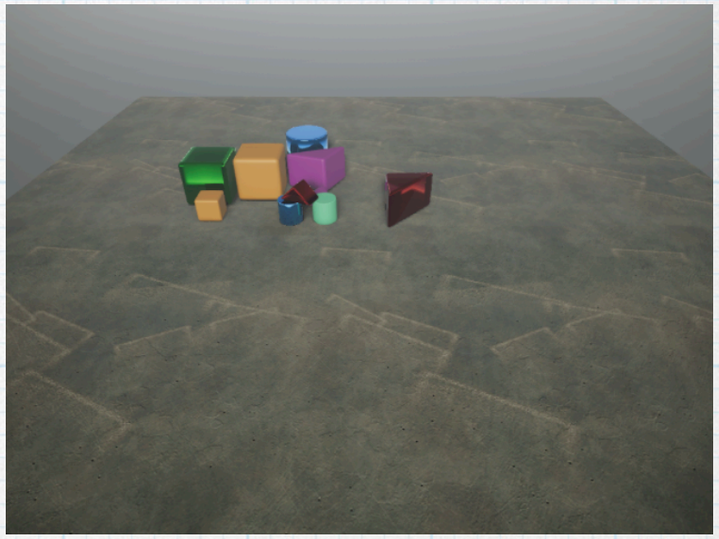


Simulation - Obtaining Initial Stable Configuration

- Initial Stable



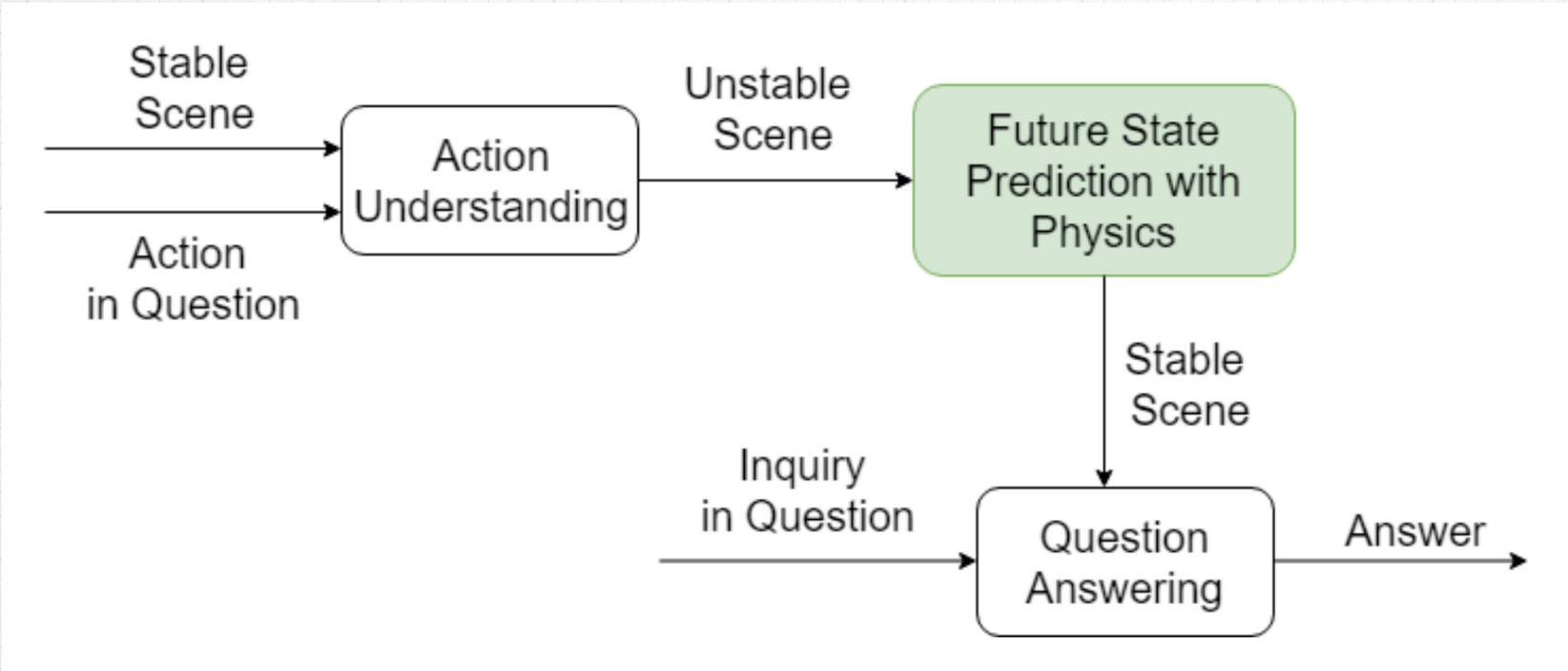
A VQA Example

✳If the big cyan object is removed, do all purple objects move? Answer: YES

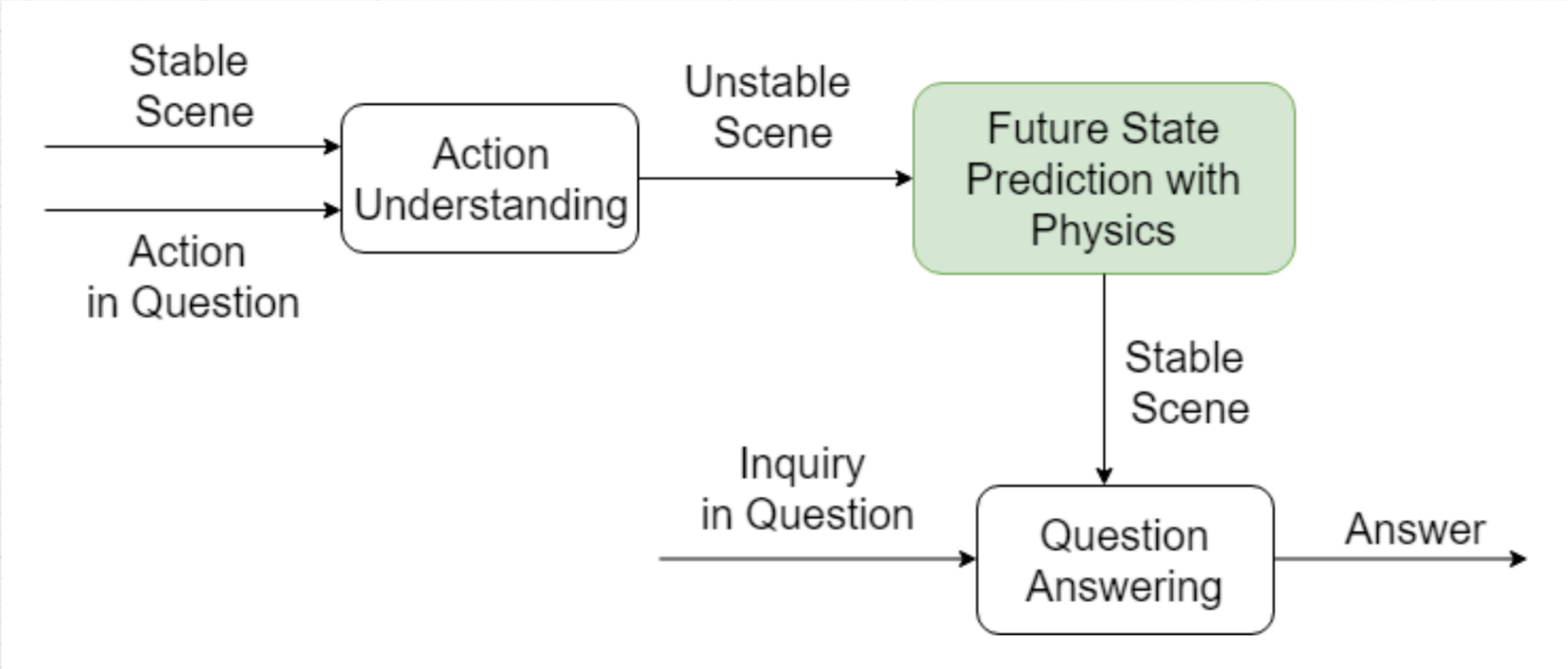
Methodology

✳Need to focus more on future state prediction with Physics



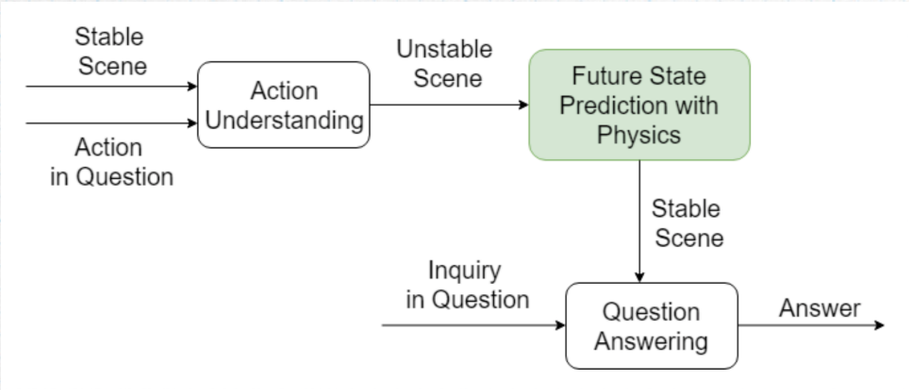
Methodology

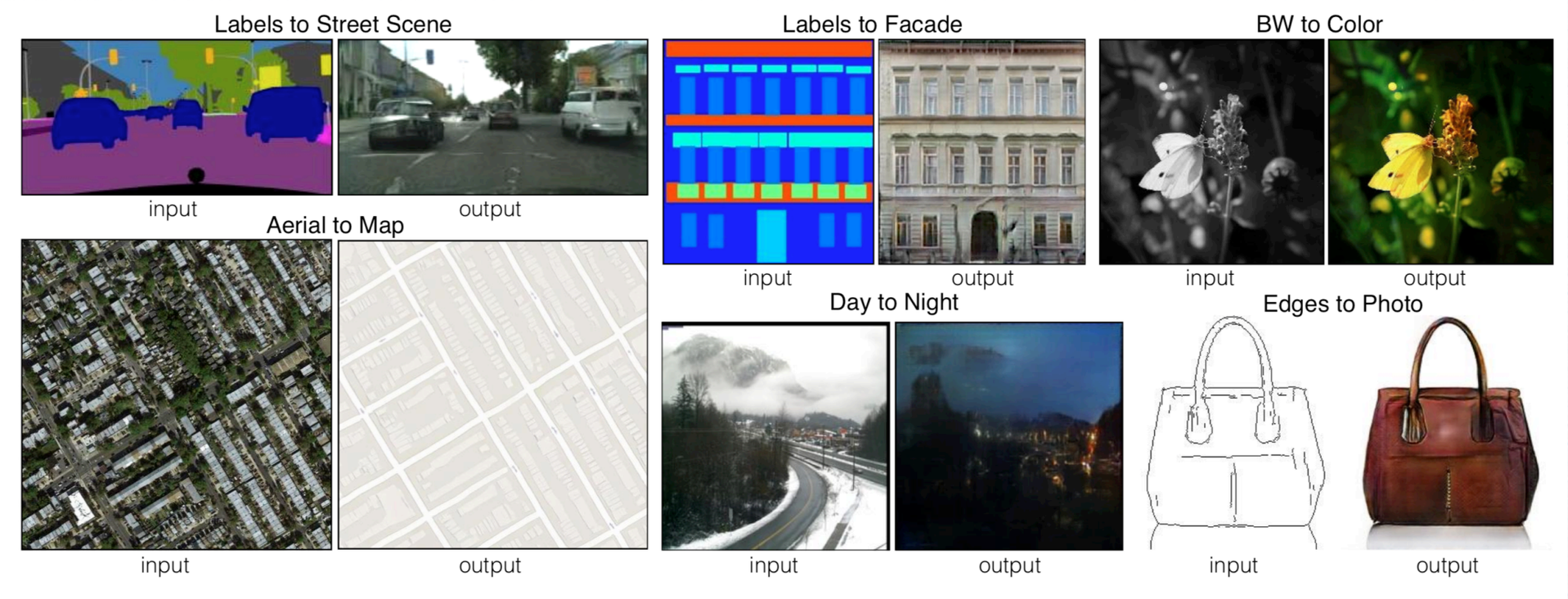
✳Need to focus more on future state prediction with Physics



✳This need can be best understood by observing simple baseline results using a method which does not employ physics

Pix2Pix



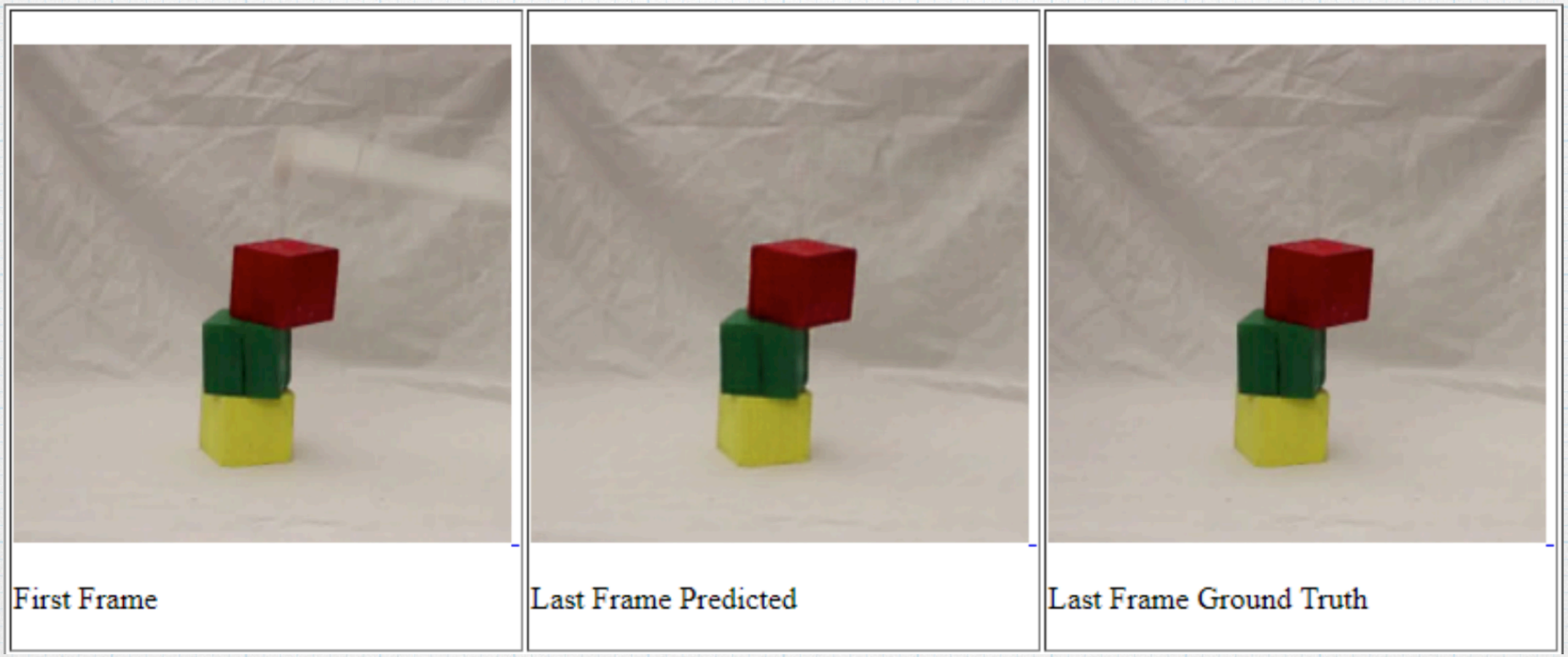


✳A conditional GAN is trained to map first frames to last frames

*Image-to-image translation with conditional adversarial networks.

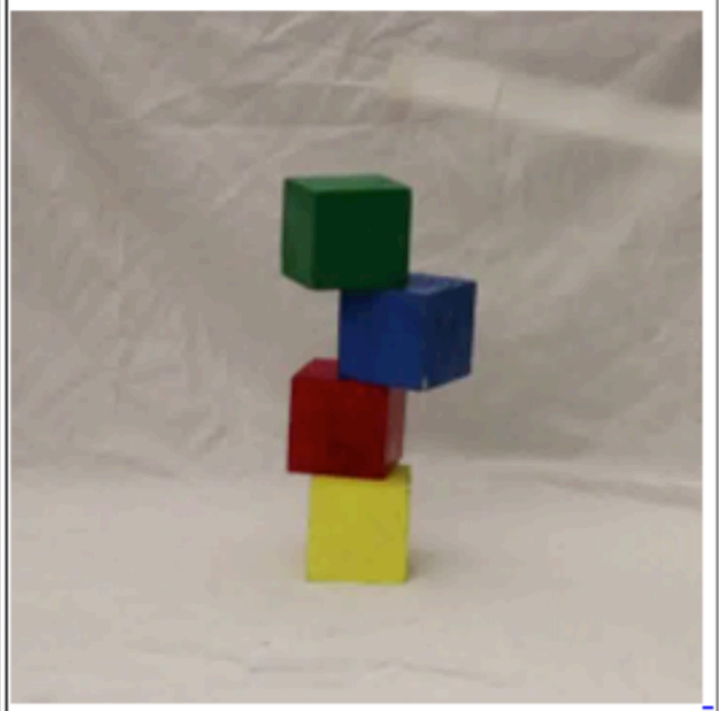
Initial Results

✳Stable scene result when trained with stable and unstable data

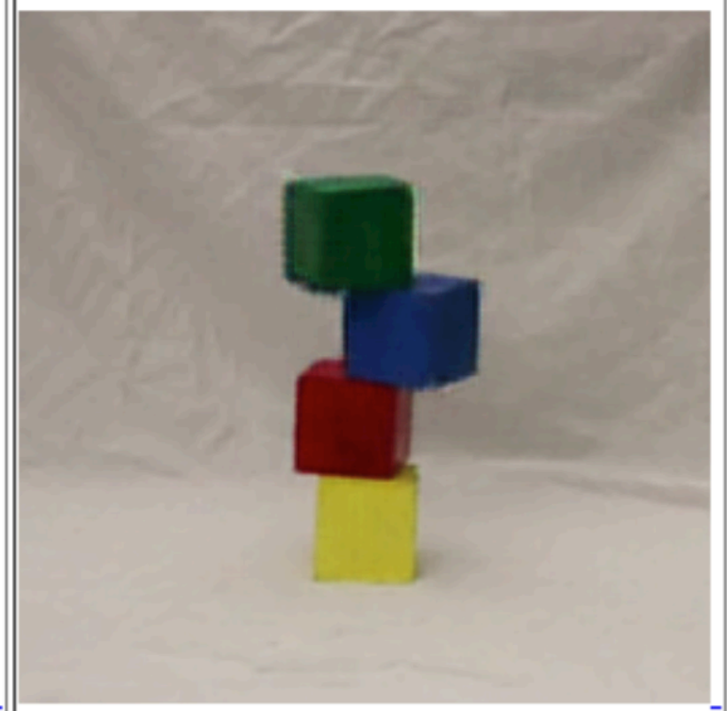


Initial Results

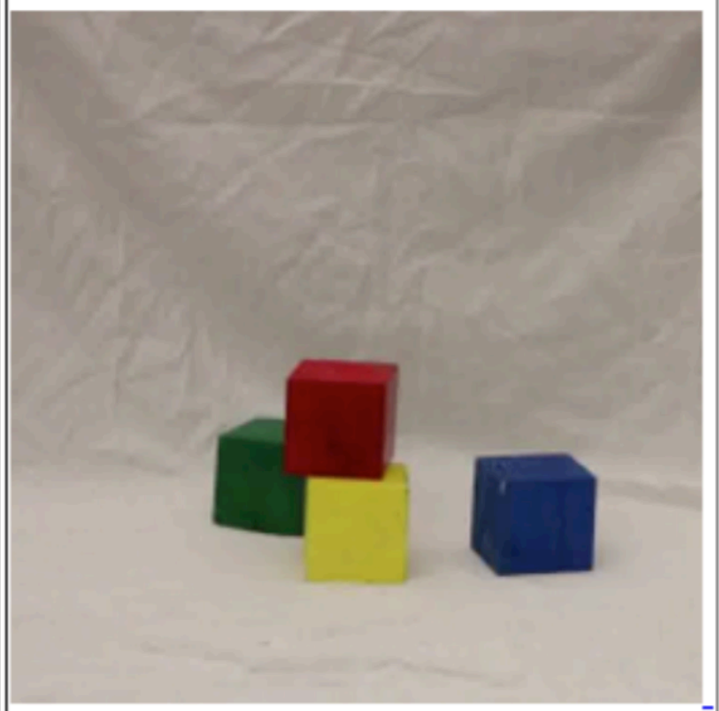
✳Unstable scene result when trained with stable and unstable data



First Frame



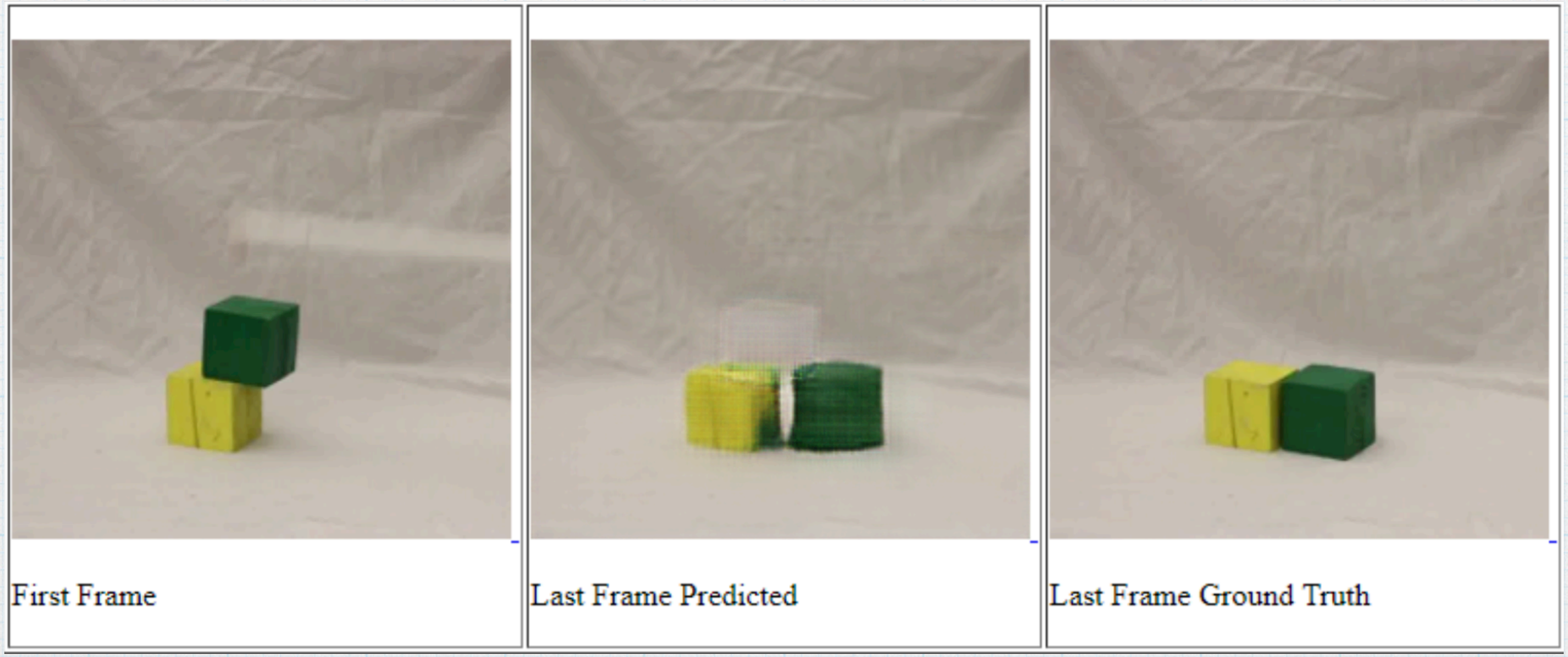
Last Frame Predicted



Last Frame Ground Truth

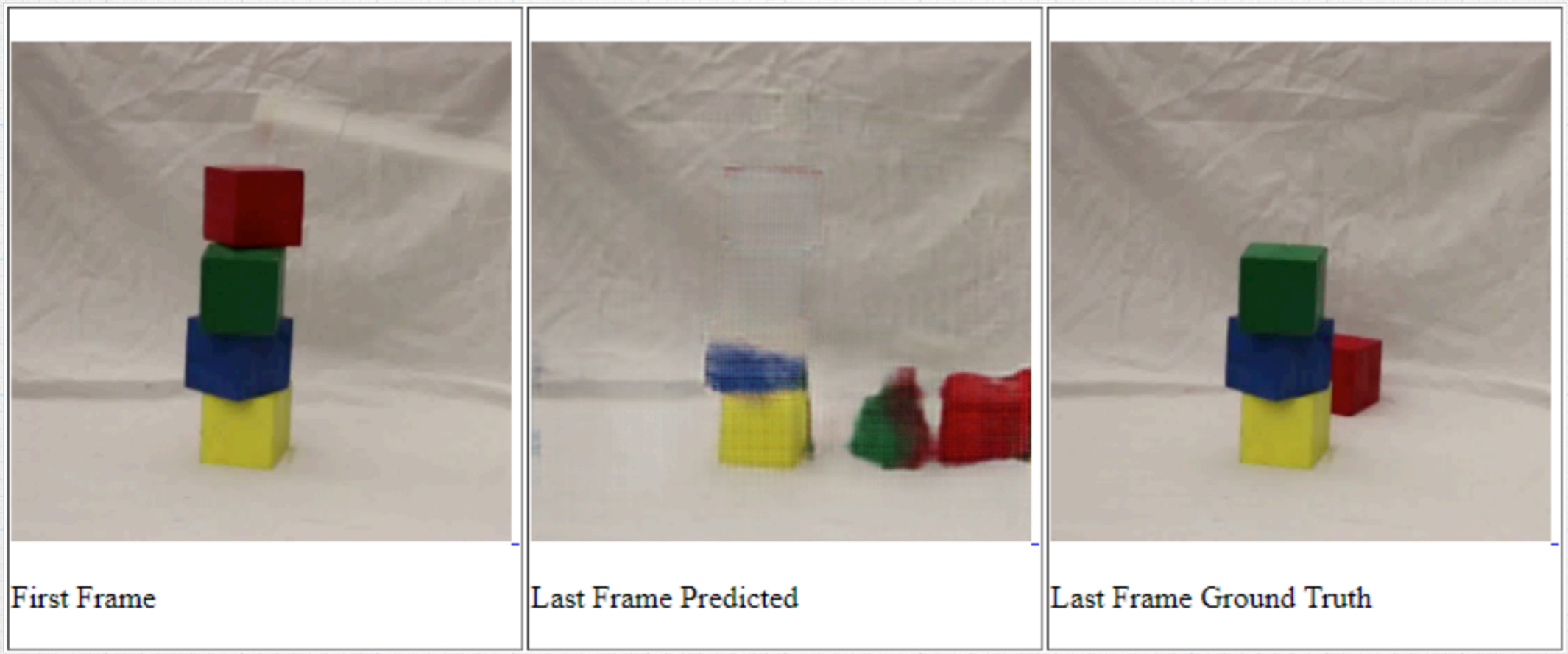
Initial Results

✳Unstable scene result when trained with **only** unstable data



Initial Results

✳Unstable scene result when trained with **only** unstable data



Conclusion

- Intuitive physics has been studied in cognitive science and had started to be studied in computer vision and machine learning
- Visual question answering has also been studied in computer vision, especially for static scenes
- Our aim is to integrate both by creating models understanding such intuitions and a dataset for future improvements
- Existing models (as in the case for Pix2Pix) are not successful understanding the physical interactions
- We will be investigating graph networks for our VQA task

Questions?

Special Thanks To: Aykut Erdem, Erkut Erdem, Burak Ercan