# Dynamic Visual Reasoning by Learning Differentiable Physics Models from Video and Language


Mingyu Ding[1] Zhenfang Chen[3] Tao Du[2] Ping Luo[1] Joshua B. Tenenbaum[2] Chuang Gan[3]

[1]The University of Hong Kong [2]MIT CSAIL [3]MIT-IBM Watson AI Lab

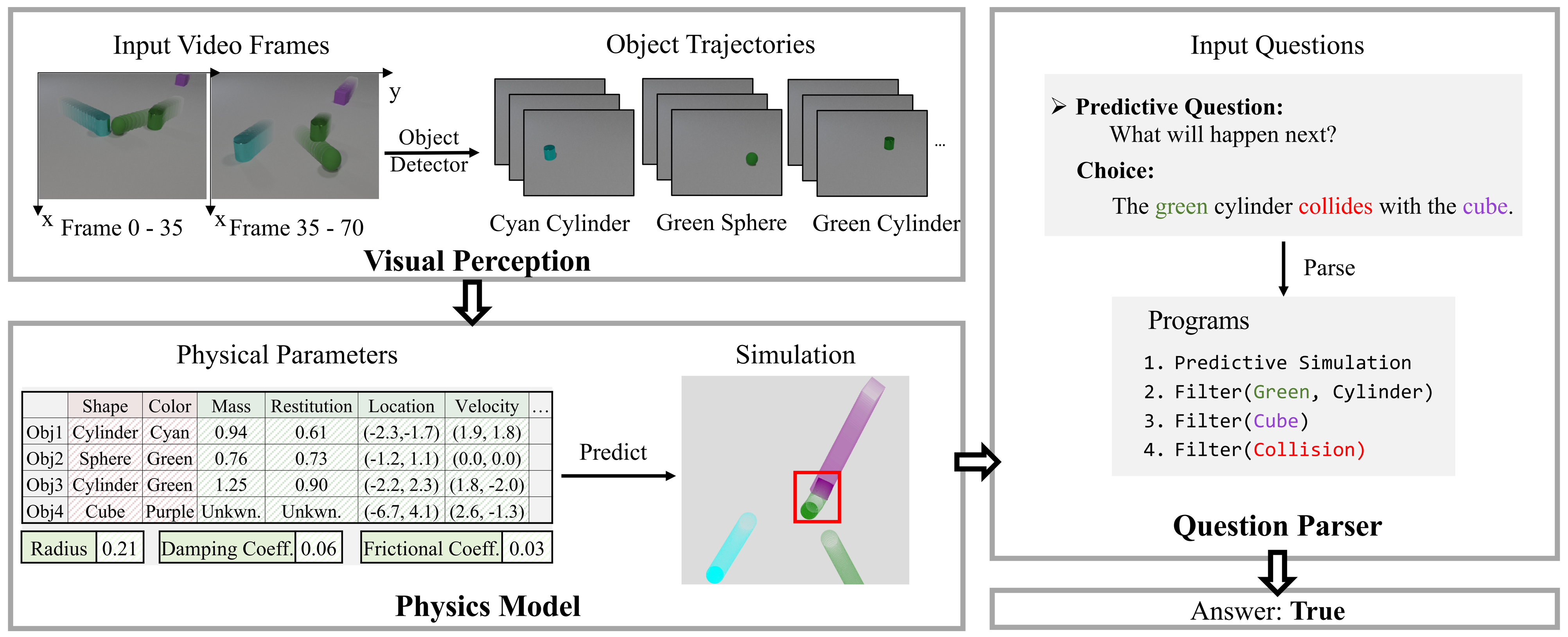

## Problem Definition and Contribution

**Goal:** Dynamic visual reasoning about objects, relations, and physics. To explain what has happened, predict what will happen, and infer what would happen in counterfactual situations.

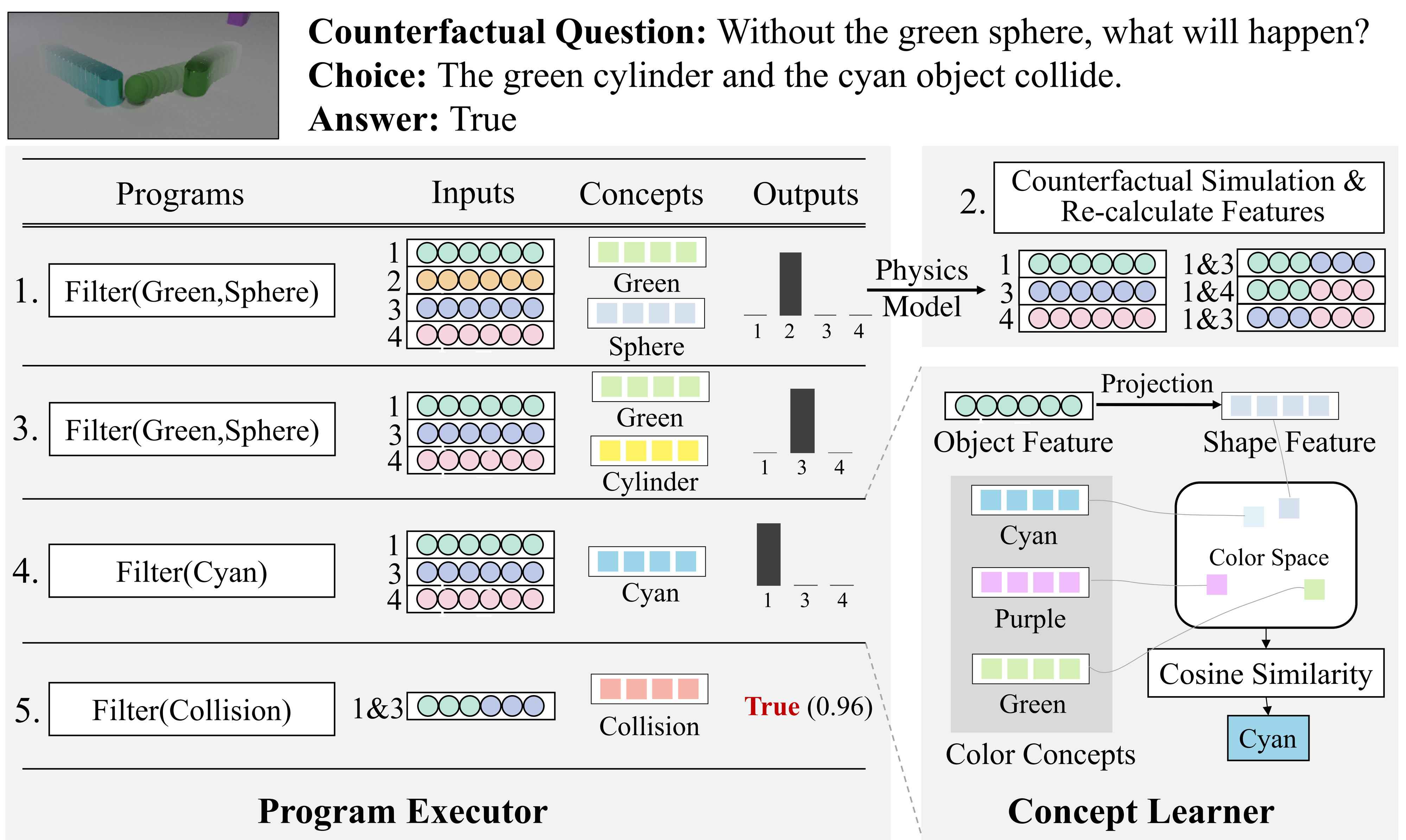
**Solution:** A unified framework VRDP that combines three mutually beneficial components: a visual perception module, a concept learner, and a differentiable physics engine.

- The visual perception module parses the input video into object trajectories and visual representations.
- The concept learner grounds language concepts and object attributes from question-answer pairs and the visual representations.
- With object trajectories and attributes as prior knowledge, the physics model optimizes all physical parameters of the scene and objects by differentiable simulation.
- The physics model reruns the simulation to reason about future motion and causal events, which are then executed by a symbolic program executor to get the answer.

## Visual Reasoning with Differentiable Physics




## Concept Learning and Program Execution

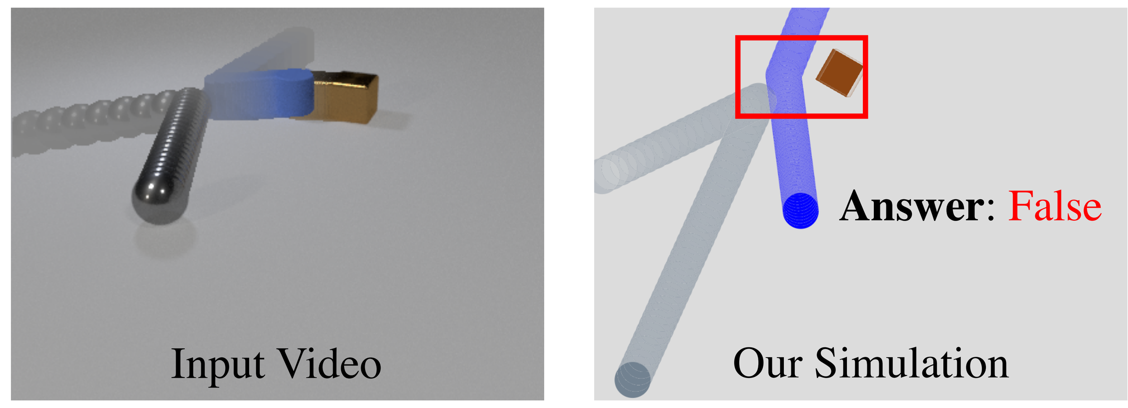



## Experiments & Results

### Learning New Concepts:

**Question**: If the blue sphere were much **heavier**, which of the events that happened would not have happened?

**Choice**: The blue cylinder collides with the cube.

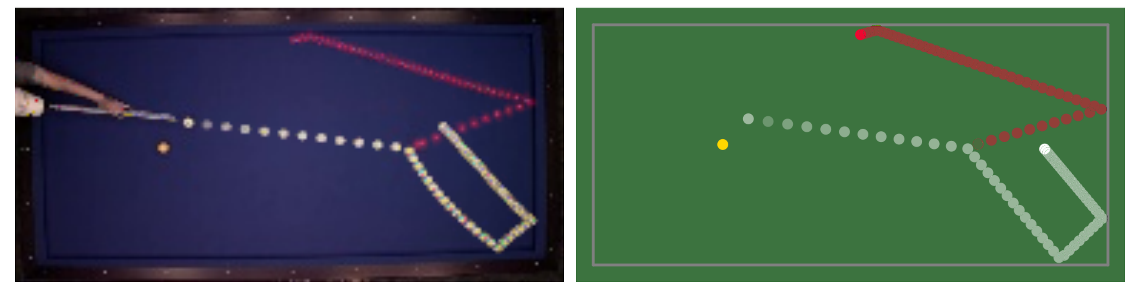



Input Video Our Simulation

### Real-world Examples:

**Question:** Will the red billiard collide with the top side of the billiard table?

**Answer:** True



Ground Truth Our Prediction

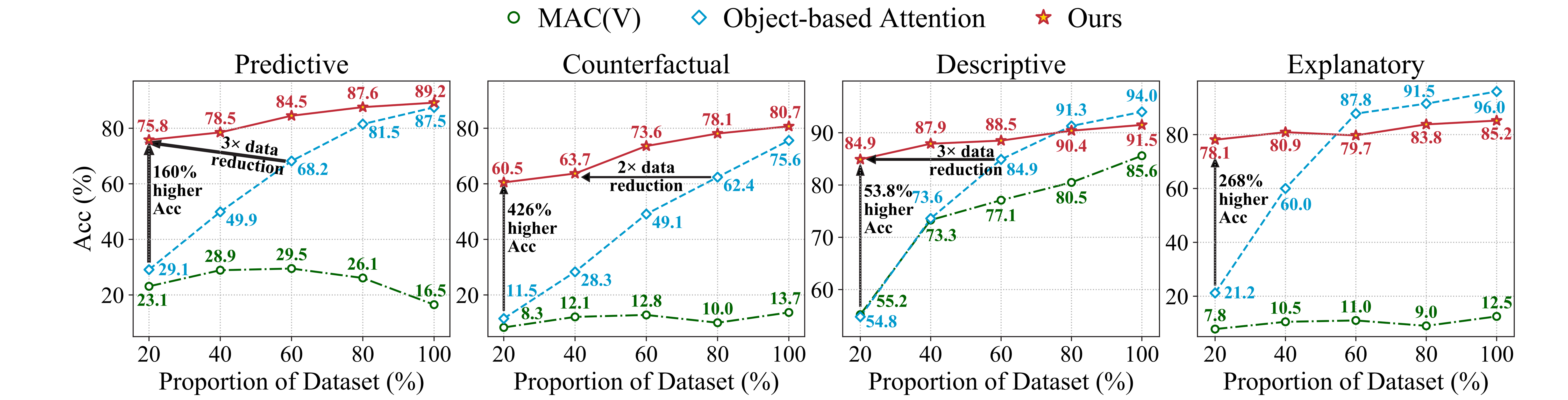
### Results on CLEVRER Benchmark:

| Methods | Overall | | Predictive | | Counterfactual | | Descriptive | Explanatory | |
|---|---|---|---|---|---|---|---|---|---|
| | per task | per ques. | per opt. | per ques. | per opt. | per ques. | | per opt. | per ques. |
| TVQA+ [49] | 37.2 | 57.3 | 70.3 | 48.9 | 53.9 | 4.1 | 72.0 | 63.3 | 23.7 |
| Memory [21] | 27.2 | 43.3 | 50.0 | 33.1 | 54.2 | 7.0 | 54.7 | 53.7 | 13.9 |
| IEP (V) [45] | 20.2 | 40.5 | 50.0 | 9.7 | 53.4 | 3.8 | 52.8 | 52.6 | 14.5 |
| TbD-net (V) [61] | 23.6 | 58.6 | 50.3 | 6.5 | 56.1 | 4.4 | 79.5 | 61.6 | 3.8 |
| HCRN [48] | 27.3 | 44.8 | 54.1 | 21.0 | 57.1 | 11.5 | 55.7 | 63.3 | 21.0 |
| MAC (V) [39] | 32.1 | 65.5 | 51.0 | 16.5 | 54.6 | 13.7 | 85.6 | 59.5 | 12.5 |
| MAC (V+) [39] † | 44.2 | 69.8 | 59.7 | 42.9 | 63.5 | 25.1 | 86.4 | 70.5 | 22.3 |
| NS-DR [81] †‡ | 69.7 | 80.7 | 82.9 | 68.7 | 74.1 | 42.2 | 88.1 | 87.6 | 79.6 |
| NS-DR (NE) [81] †‡ | 64.1 | 77.7 | 75.4 | 54.1 | 76.1 | 42.0 | 85.8 | 85.9 | 74.3 |
| DCL [16] † | 75.5 | 84.1 | 90.5 | 82.0 | 80.4 | 46.5 | 90.7 | 89.6 | 82.8 |
| DCL-Oracle [16] †‡ | 75.6 | 84.5 | 90.6 | 82.1 | 80.7 | 46.9 | 91.4 | 89.8 | 82.0 |
| Object-based Attention [20] | 88.3 | 91.7 | 93.5 | 87.5 | 91.4 | 75.6 | **94.0** | **98.5** | **96.0** |
| VRDP (ours) | 82.9 | 86.9 | 91.7 | 83.8 | 89.9 | **75.7** | 89.8 | 89.1 | 82.4 |
| VRDP (ours) † | 86.6 | 89.4 | **94.5** | **89.2** | **92.5** | **80.7** | 91.5 | 90.9 | 85.2 |
| VRDP (ours) †‡ | **90.3** | **92.0** | **95.7** | **91.4** | **94.8** | **84.3** | 93.4 | 96.3 | 91.9 |

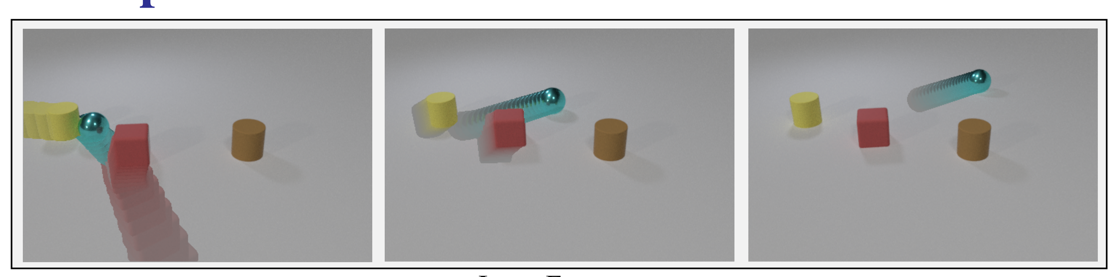
### Ablation Study on the Optimization of Physical Parameters:

| Methods | Overall | | Predictive | | Counterfactual | | Descriptive | Explanatory | |
|---|---|---|---|---|---|---|---|---|---|
| | per task | per ques. | per opt. | per ques. | per opt. | per ques. | | per opt. | per ques. |
| Baseline | 72.6 | 81.6 | 85.1 | 72.4 | 77.6 | 49.6 | 87.8 | 88.0 | 80.6 |
| + Collision-independent First | 81.3 | 87.8 | 86.1 | 72.8 | 89.3 | 74.1 | 91.3 | 91.9 | 86.9 |
| + Curriculum Optimization | 85.6 | 90.2 | 87.6 | 76.5 | 94.8 | 84.3 | 92.2 | 93.3 | 89.2 |
| + Re-optimization for Prediction | **90.3** | **92.0** | **95.7** | **91.4** | **94.8** | **84.3** | **93.4** | **96.3** | **91.9** |

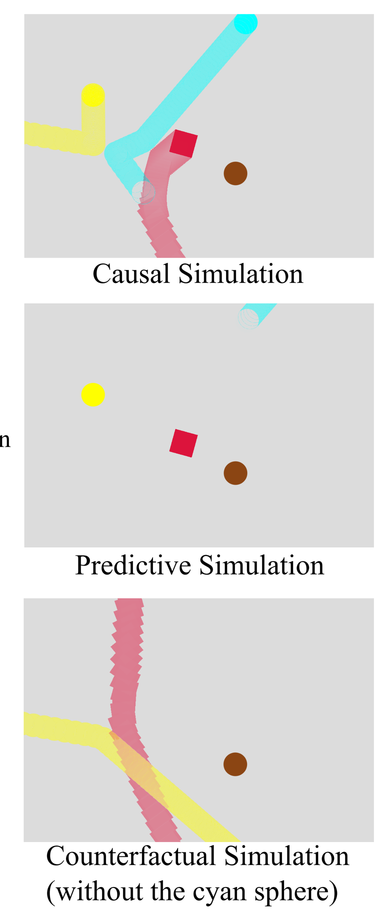
### Data Efficiency Evaluation:



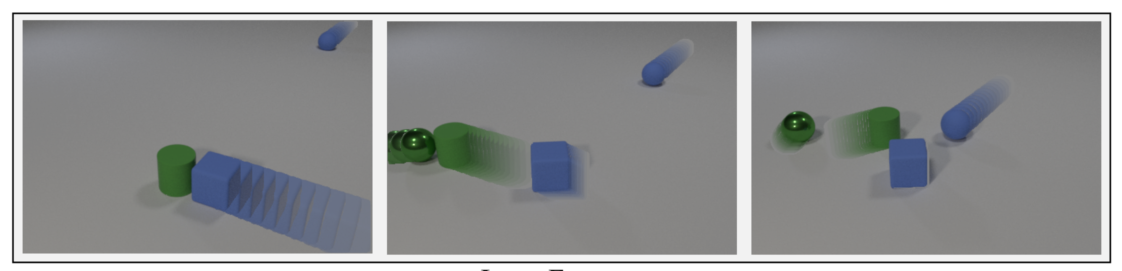

### Examples on the CLEVRER dataset:



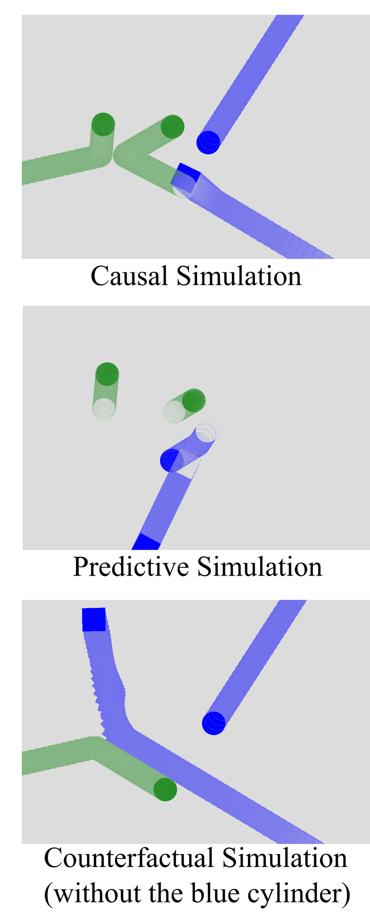
Input Frames




| | | |
|---|---|---|
| **Descriptive Q1:** How many collisions happen? | **A:** 3 | **GT:** 3 |
| **Descriptive Q2:** What shape is the first object that enter the scene? | **A:** cube | **GT:** cube |
| **Descriptive Q3:** How many cylinders enter the scene after the cube enters? | **A:** 1 | **GT:** 1 |
| **Descriptive Q4:** What color is the object that is stationary? | **A:** brown | **GT:** brown |
| **Explanatory Q:** Which of the following is not responsible for the collision between the yellow cylinder and the sphere? | | |
| a) The presence of the red cube. | **A:** True | **GT:** True |
| b) The collision between the cube and the cyan sphere. | **A:** True | **GT:** True |
| c) The presence of the brown cylinder. | **A:** False | **GT:** False |
| **Predictive Q:** Which event will happen next? | | |
| a) The cyan sphere and the brown cylinder collide. | **A:** False | **GT:** False |
| b) The yellow object exits the scene. | **A:** False | **GT:** False |
| **Counterfactual Q:** Without the cyan sphere, what will happen? | | |
| a) The cube collide with the brown object. | **A:** False | **GT:** False |
| b) The cube and the yellow object collide. | **A:** True | **GT:** True |
| c) The brown cylinder collide with the yellow object. | **A:** False | **GT:** False |



Input Frames




| | | |
|---|---|---|
| **Descriptive Q1:** What is the material of the last object that enters the scene? | **A:** metal | **GT:** metal |
| **Descriptive Q2:** What is the material of the last object to collide with the cylinder? | **A:** metal | **GT:** metal |
| **Descriptive Q3:** How many blue cubes enter the scene? | **A:** 1 | **GT:** 1 |
| **Descriptive Q4:** How many moving green objects are there when the video ends? | **A:** 2 | **GT:** 2 |
| **Explanatory Q:** Which of the following is responsible for the collision between the cylinder and the metal object? | | |
| a) The presence of the blue rubber cube. | **A:** True | **GT:** True |
| b) The cube's colliding with the cylinder. | **A:** True | **GT:** True |
| c) The presence of the blue rubber sphere. | **A:** False | **GT:** False |
| **Predictive Q:** Which event will happen next? | | |
| a) The metal object collides with the blue sphere. | **A:** False | **GT:** False |
| b) The cube and the blue sphere collide. | **A:** True | **GT:** True |
| **Counterfactual Q:** If the cylinder is removed, which event will happen? | | |
| a) The cube collides with the rubber sphere. | **A:** False | **GT:** False |
| b) The cube and the metal sphere collide. | **A:** True | **GT:** True |
| c) The green sphere and the rubber sphere collide. | **A:** False | **GT:** False |