A Visual Question Answering Study In Intuitive Physics

Tayfun ATEŞ, N17230653
Burak Ercan, N17145570

Overview (TODO)

- Introduction
- Related Work
- Dataset
- Proposed Method
  - The Latent Sketch Module
  - The Conditional Sketch Module
  - The Portray Module
  - ClothNet
- Experiments
- Conclusions

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

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

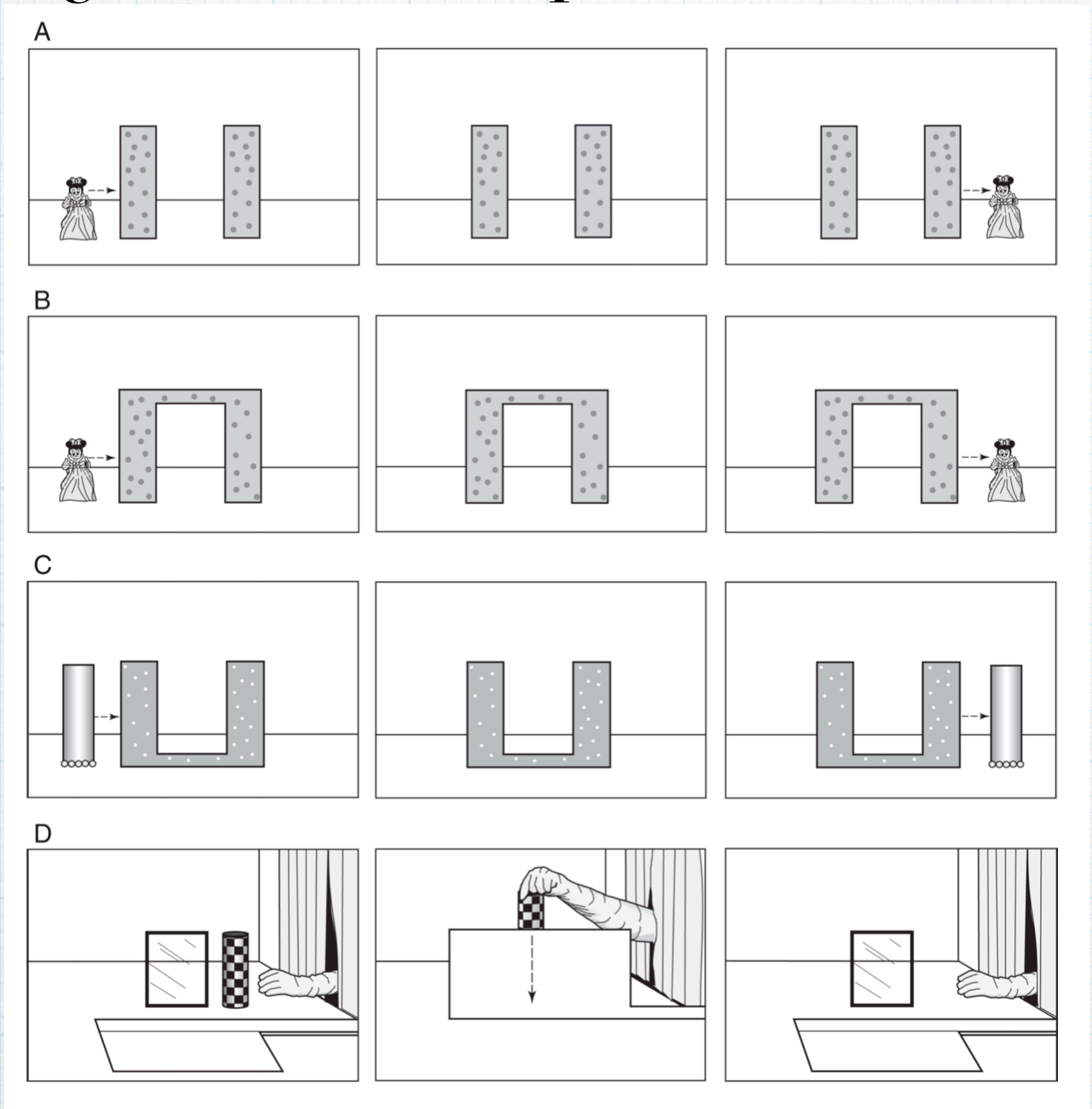
What is Intuitive Physics?

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

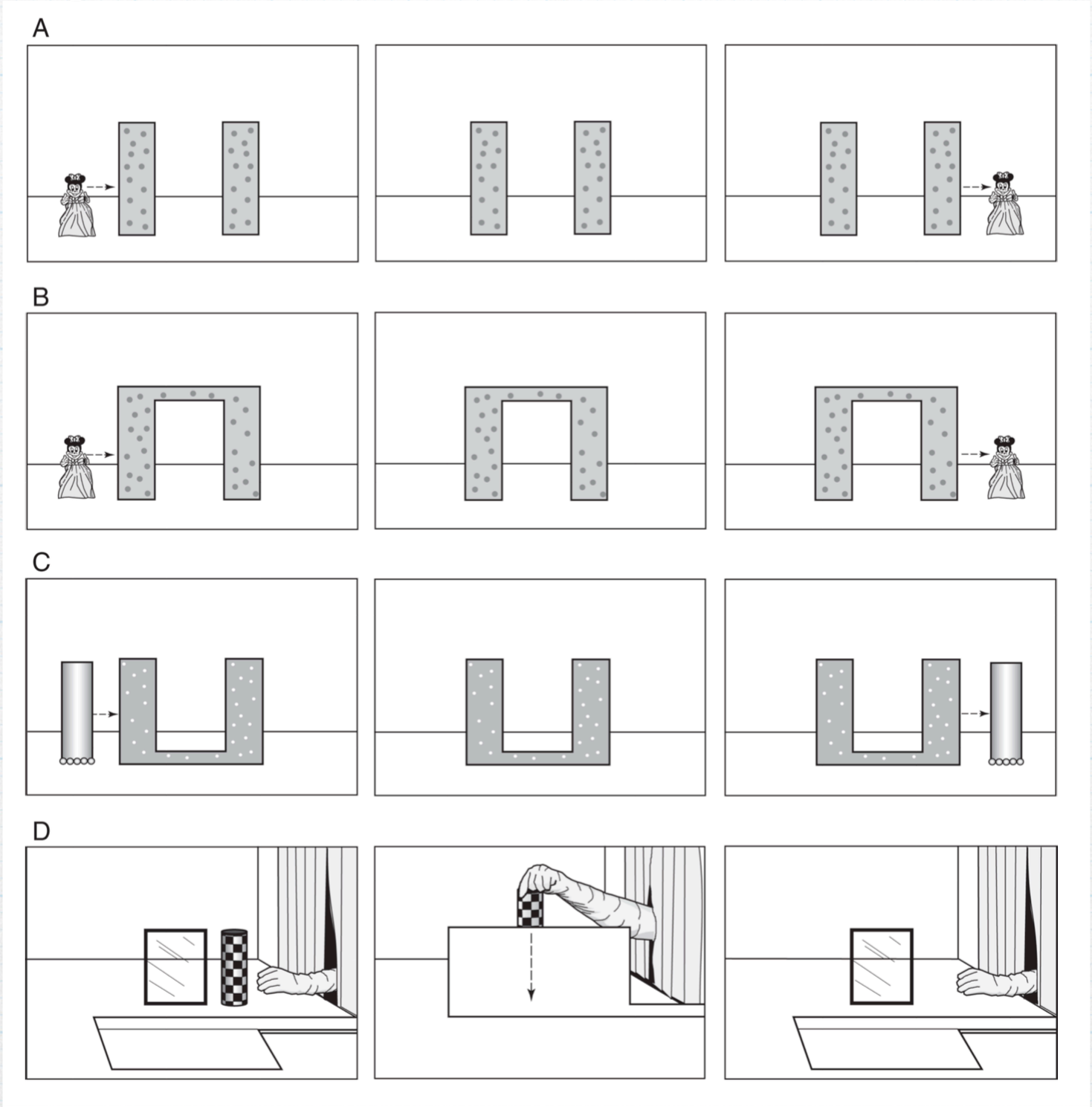
Cognitive Development of Infants



Continuity violation detection of infants for different setups

*Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning

Cognitive Development of Infants

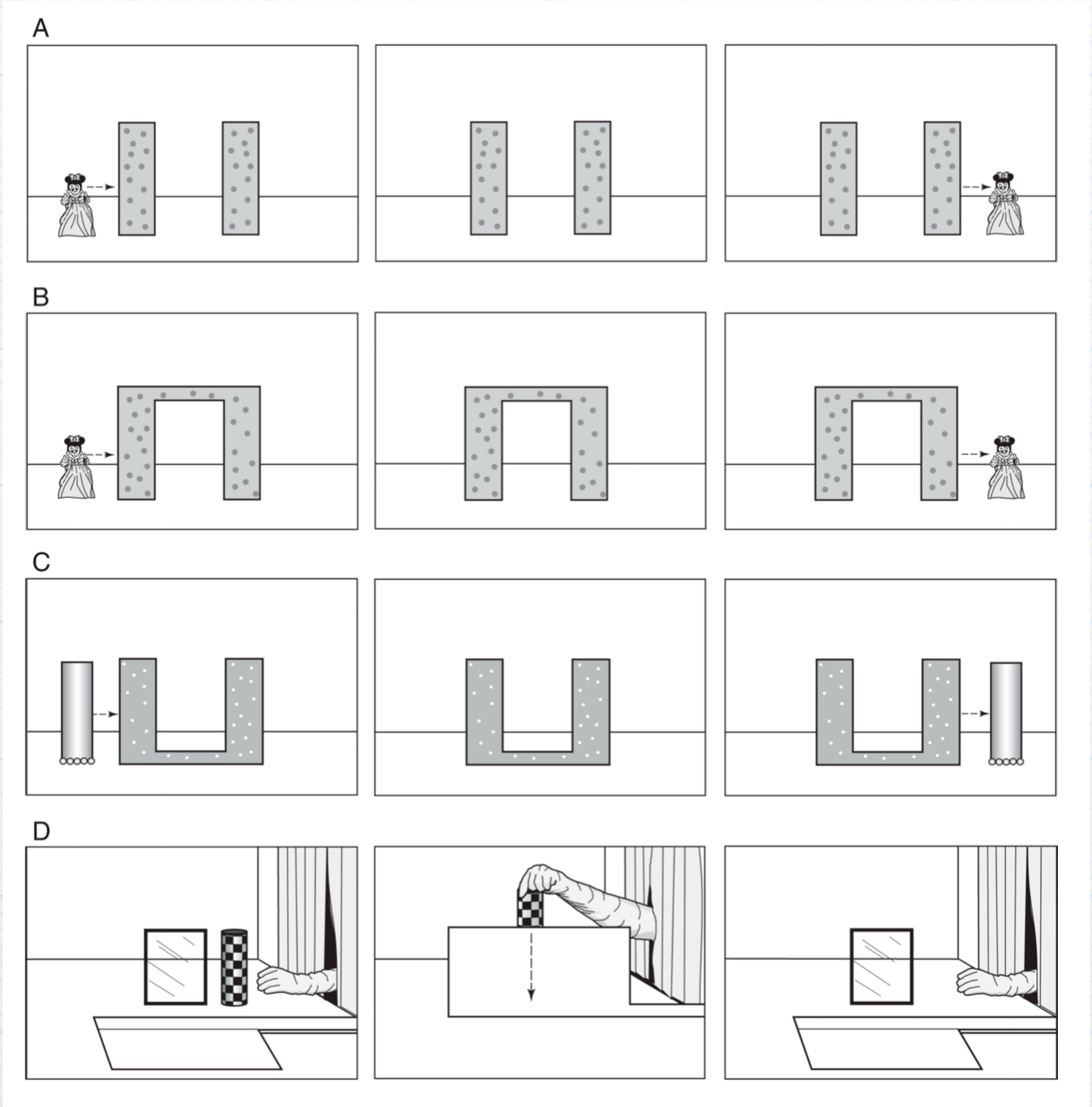


After 2.5 months

Continuity violation detection of infants for different setups

*Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning

Cognitive Development of Infants



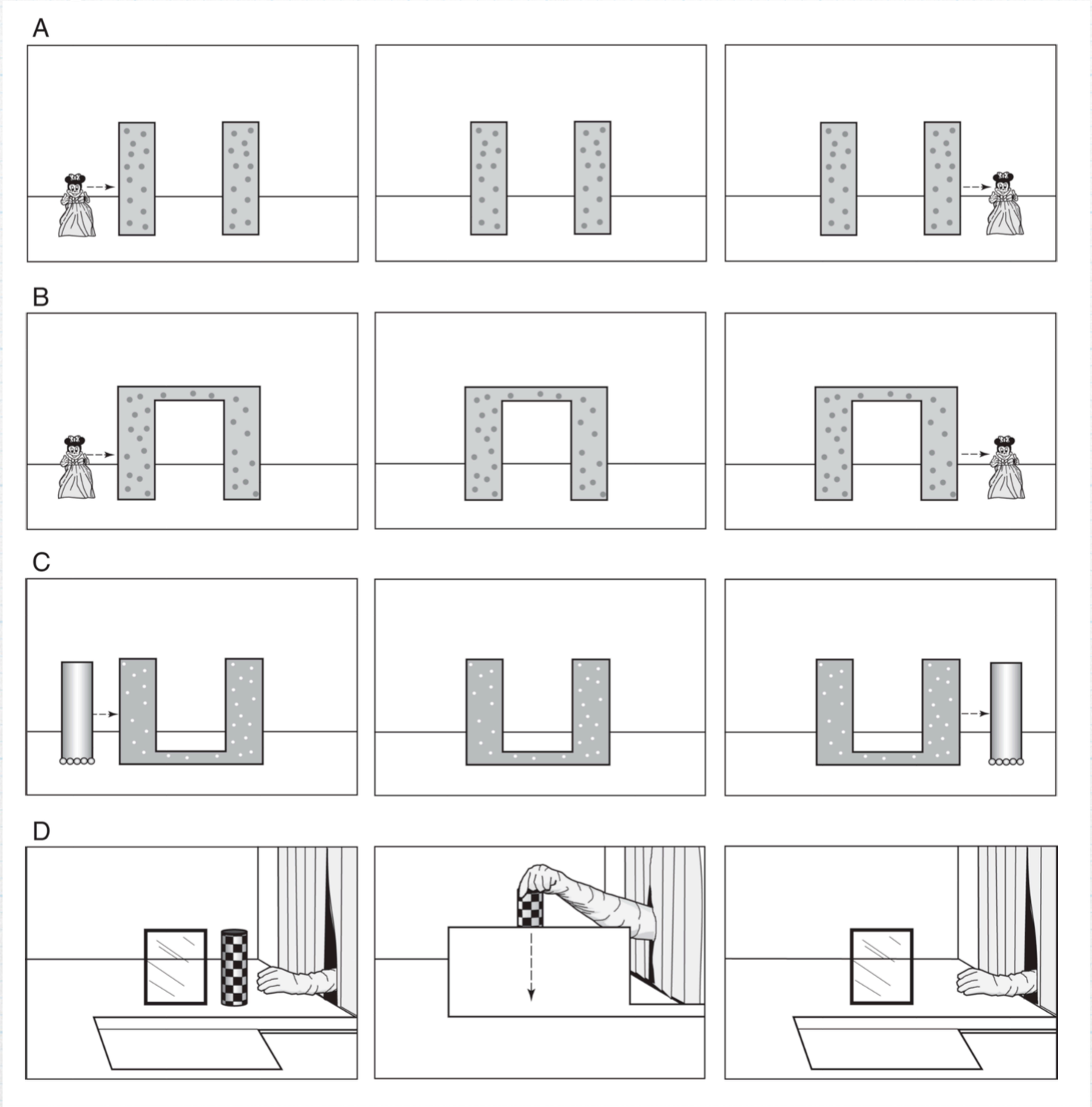
After 2.5 months

After 3 months

Continuity violation detection of infants for different setups

*Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning

Cognitive Development of Infants



After 2.5 months

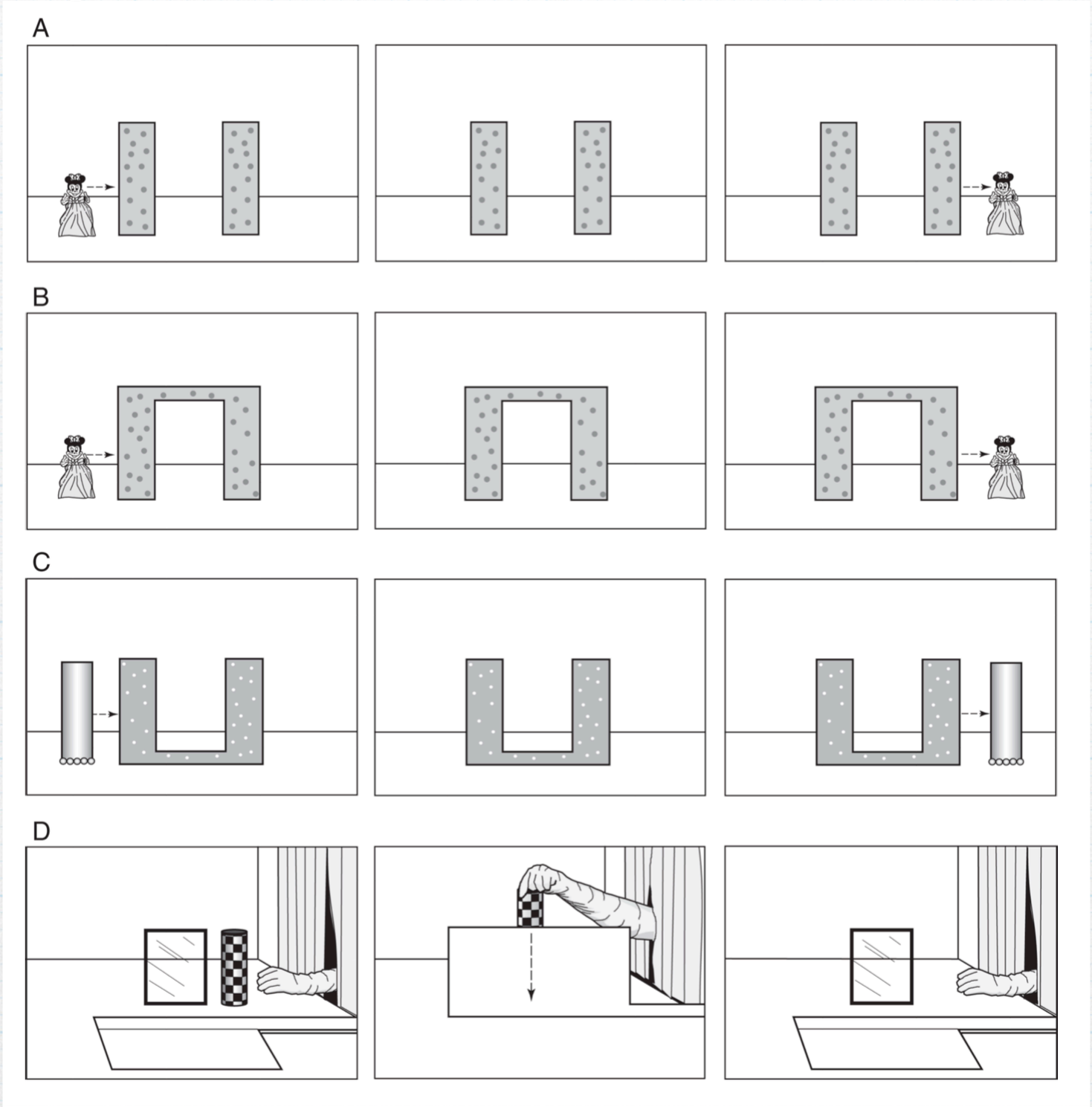
After 3 months

After 3.5 months

Continuity violation detection of infants for different setups

*Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning

Cognitive Development of Infants



After 2.5 months

After 3 months

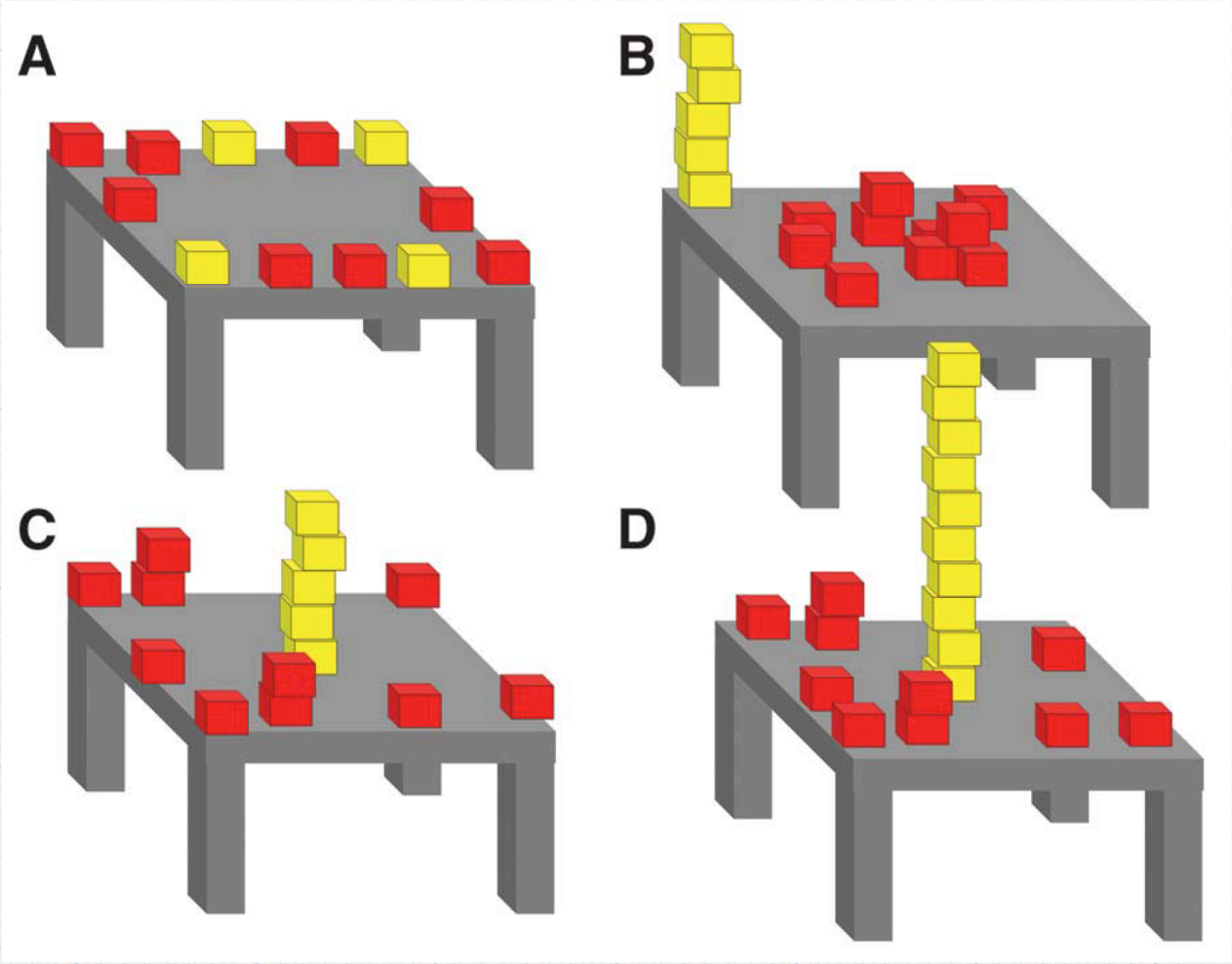
After 3.5 months

After 7.5 months

Continuity violation detection of infants for different setups

*Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning

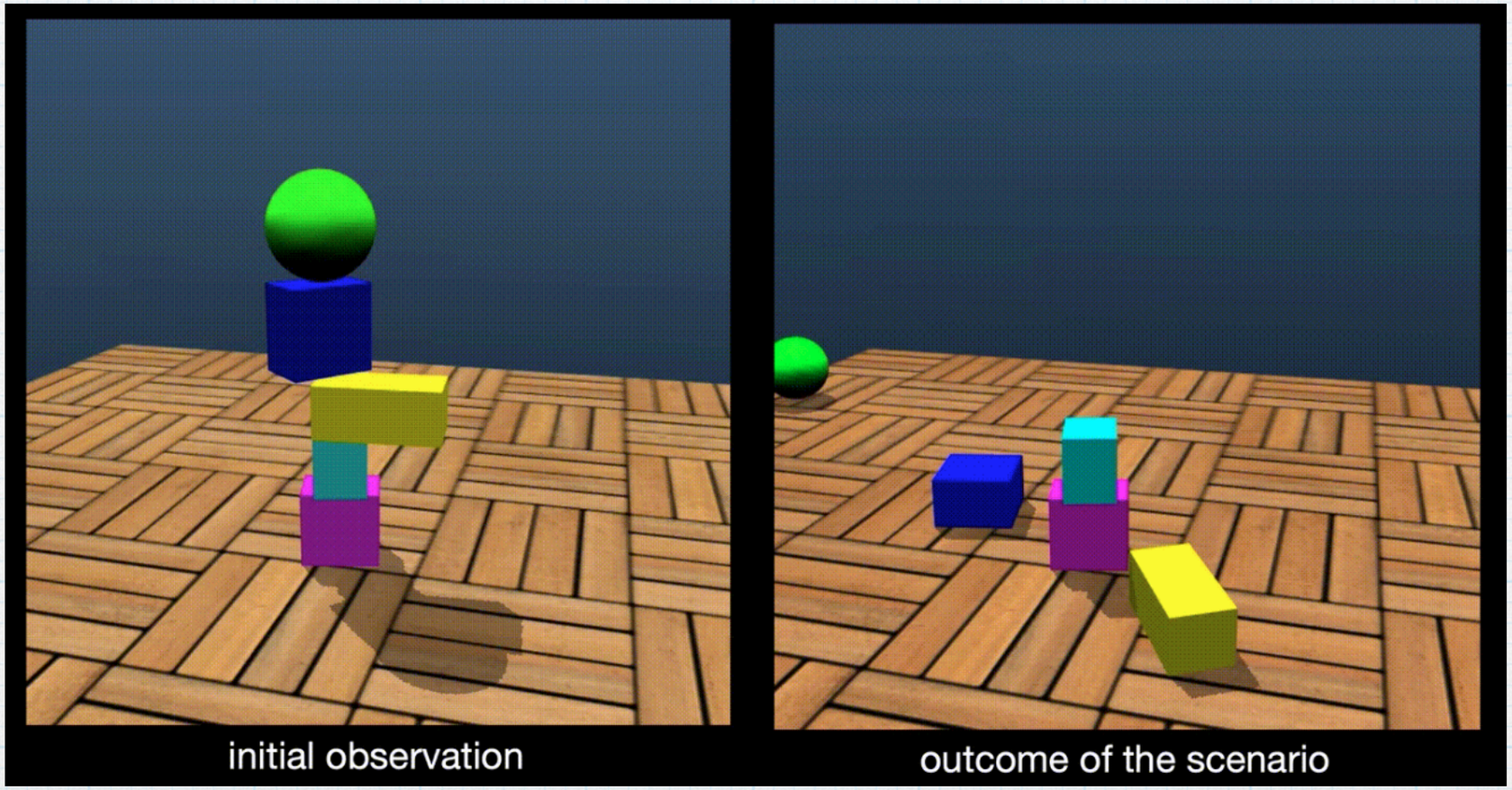
An Example



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

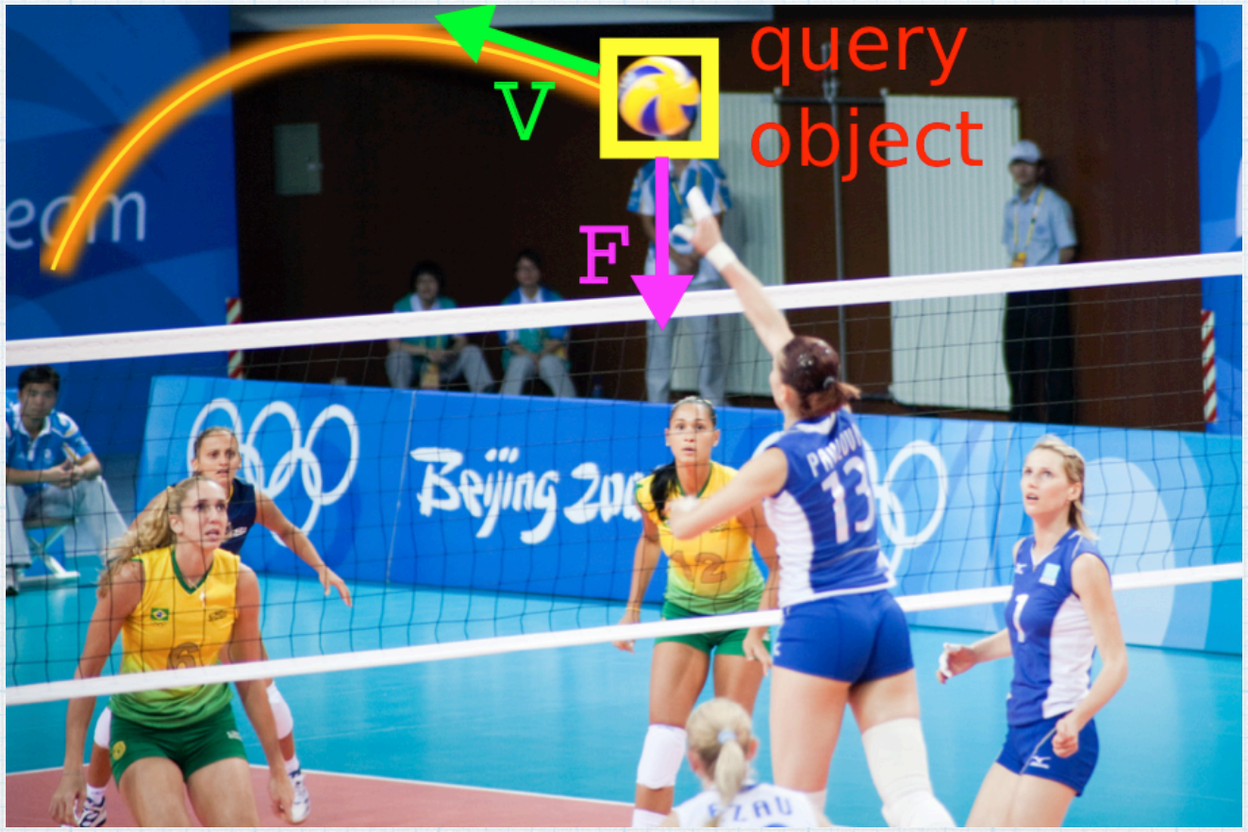
Another Example



Is the scene titled as initial observation is stable or not?

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

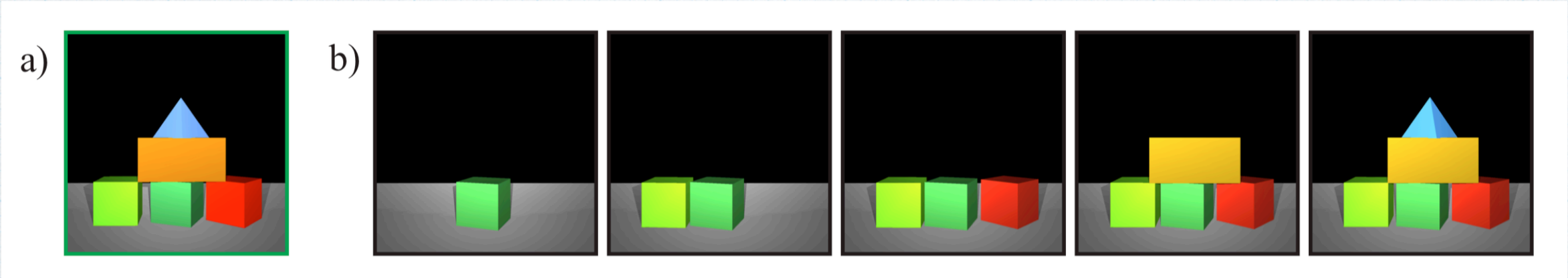
Another Example



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

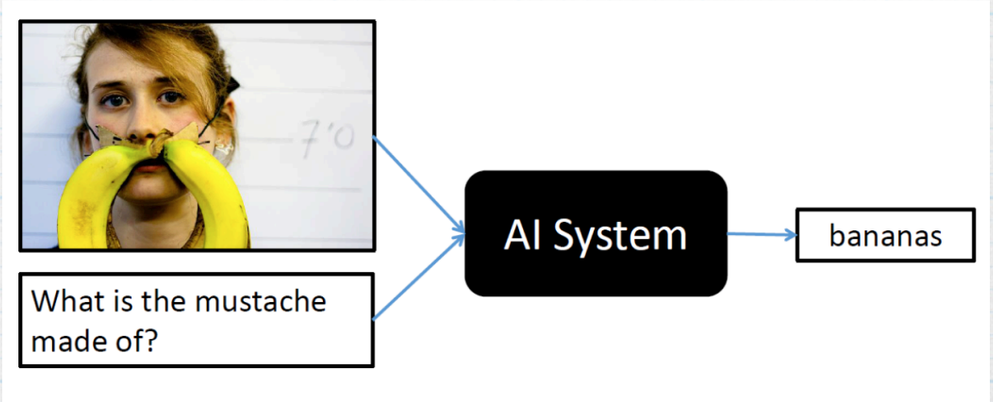


Given an initial observation can you predict how it is constructed?

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

Visual Question Answering

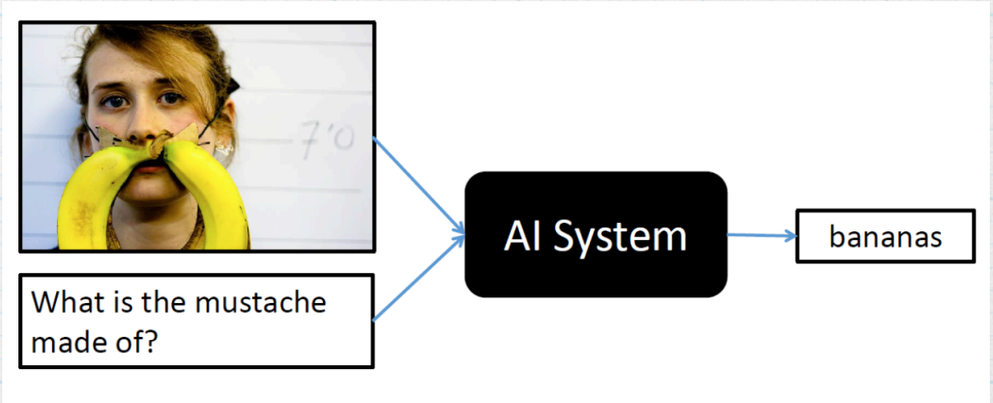
- Given a visual (video or image) and a question in a textual format, model tries to give an answer by understanding the content of the image.



Visual Question Answering

- Given a visual (video or image) and a question in a textual format, model tries to give an answer by understanding the content of the image.

- The model must be able to reason about the scene to give the correct answer. Therefore, extracting object properties and the relations between these objects is crucial.

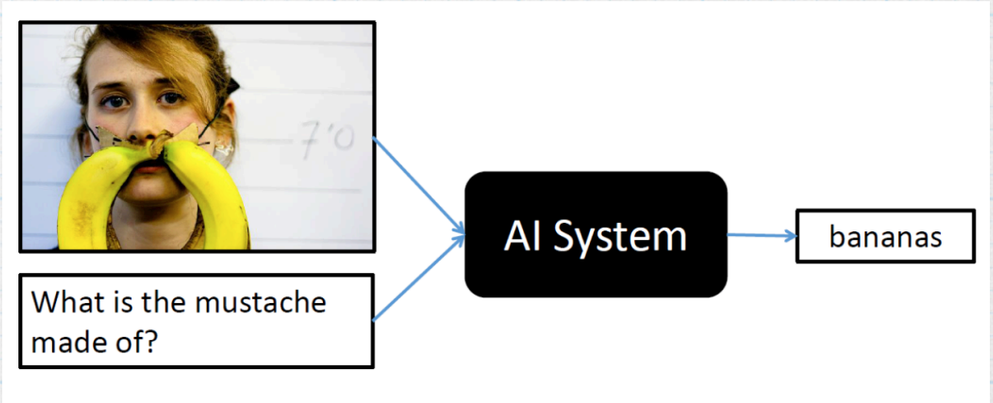


Visual Question Answering

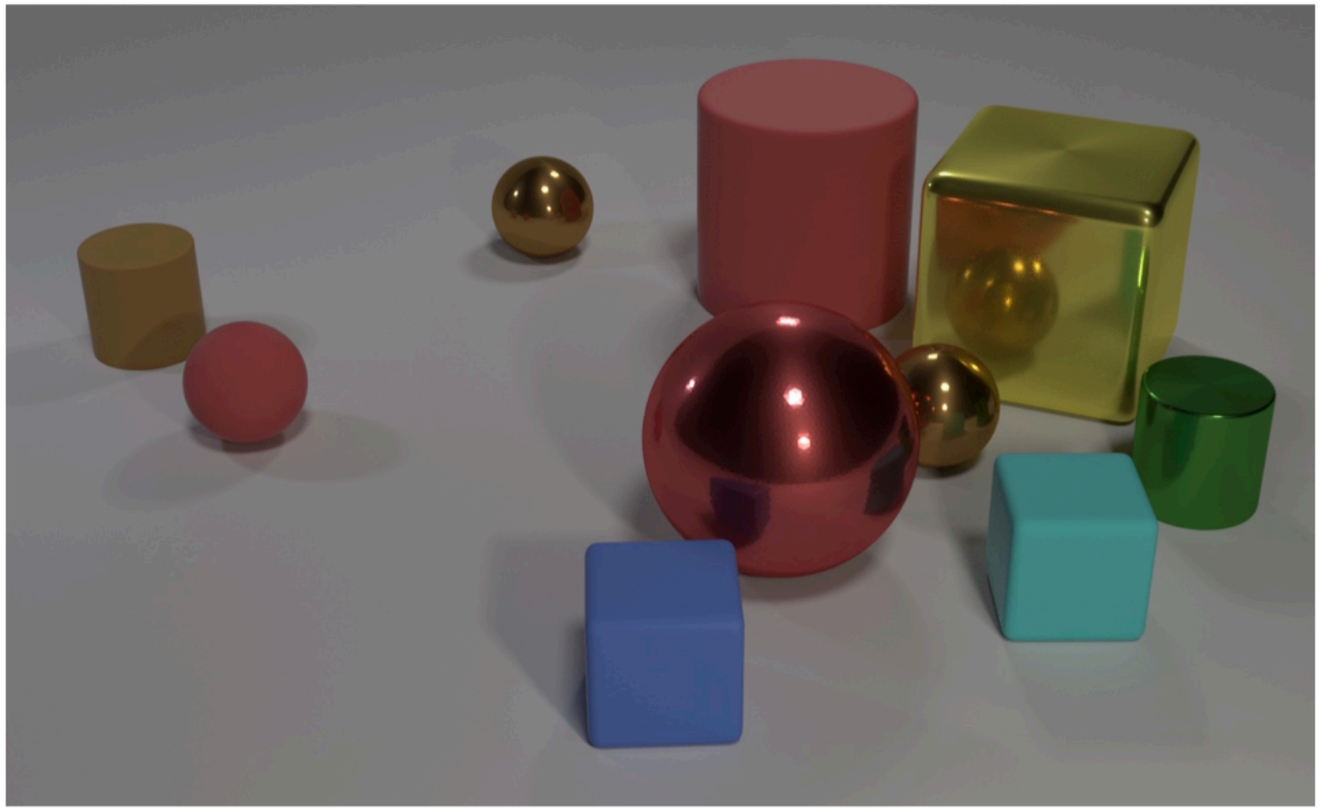
- Given a visual (video or image) and a question in a textual format, model tries to give an answer by understanding the content of the image.

- The model must be able to reason about the scene to give the correct answer. Therefore, extracting object properties and the relations between these objects is crucial.

- Several datasets with different question types are introduced to evaluate the performances of current state-of-the-art.



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.

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?

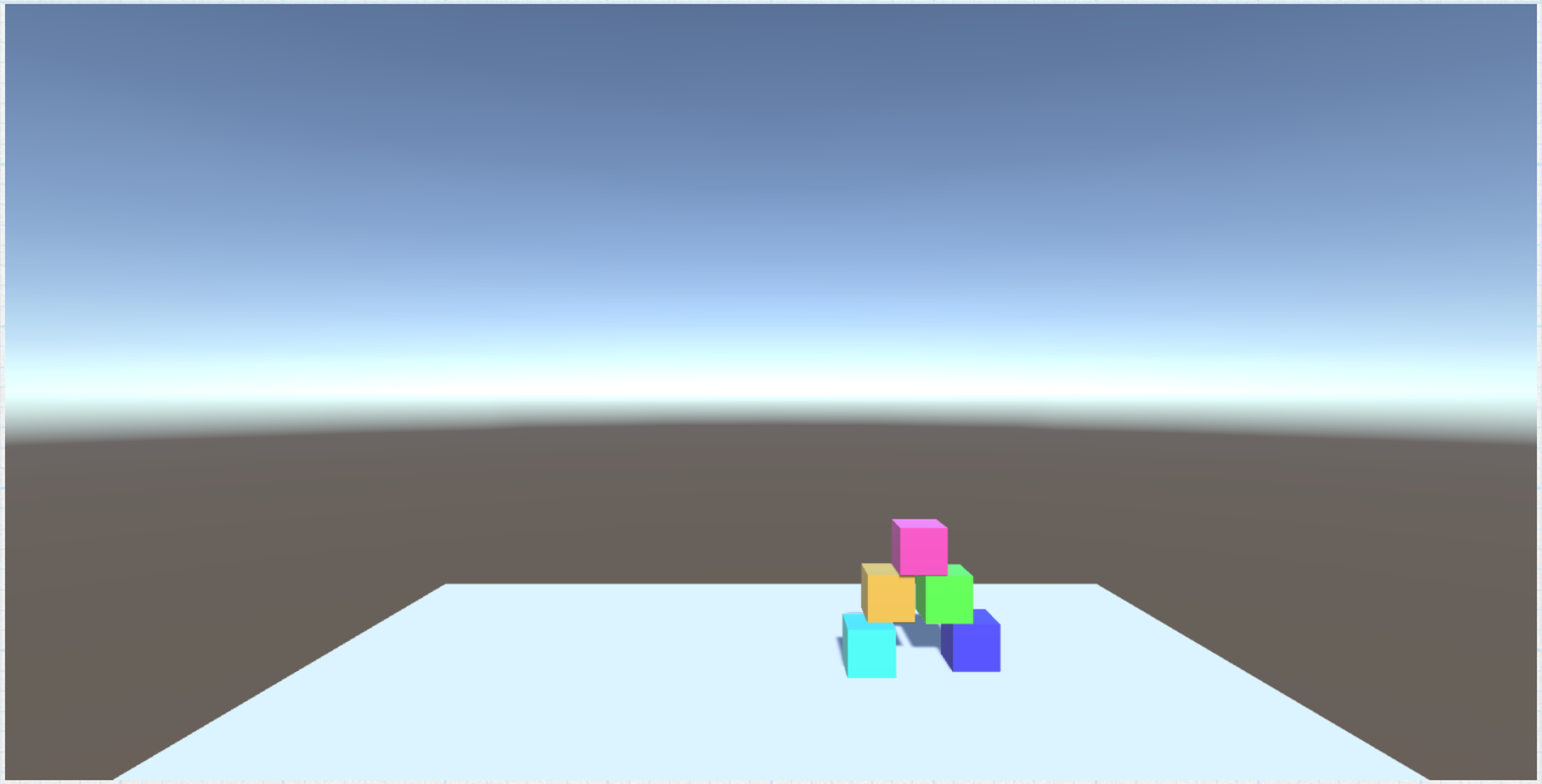
- From VQA point of view, mostly a single state is under consideration and future states are ignored.

- 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.

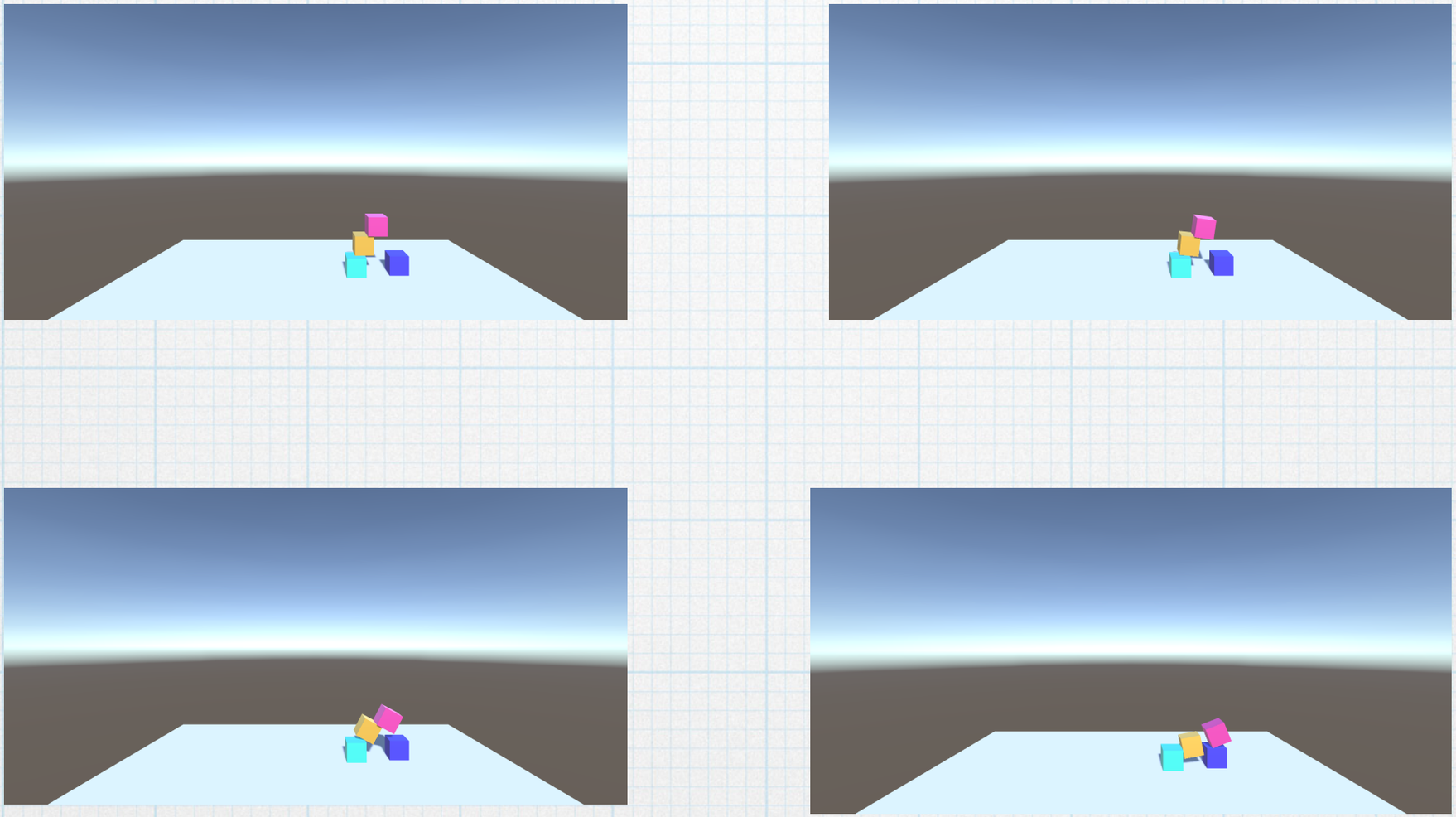
- Besides the model that is capable of simulating such intuitions, we will output a VQA dataset for further improvements.

An Example Question in Intuitive Physics



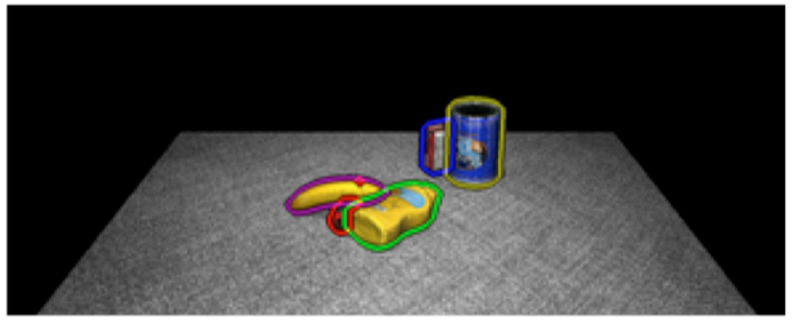
If I remove the object which just on top of blue object, does the object with orange color move?

An Example Question in Intuitive Physics

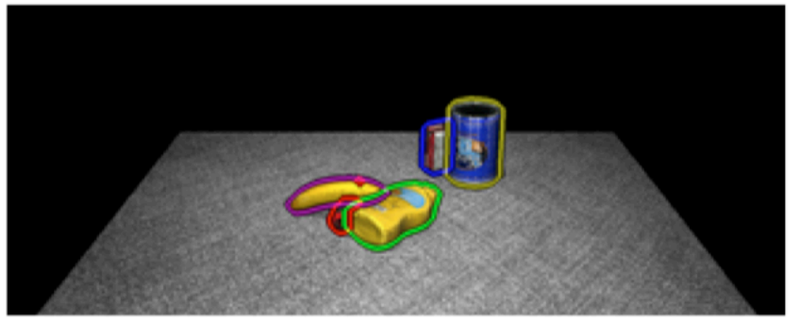


The answer should be: YES.

Answering visual what-if questions: from actions to predicted scene descriptions

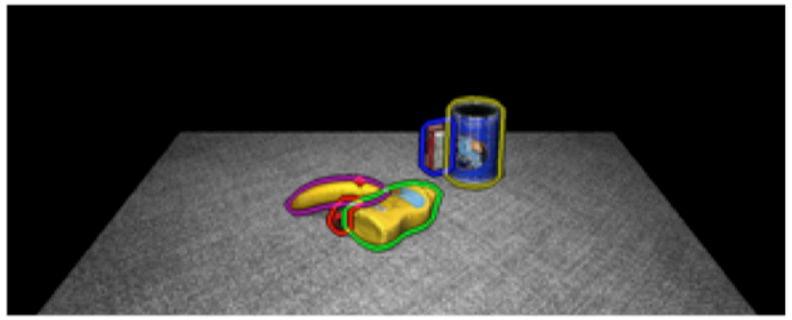
 2) "the robot spins the screw driver in anti-clowise direction" — what happens to the mustard container?	**Ground Truth** "the screw driver pushes the mustard container" **Human Baseline** "the mustard conatainer moves a little due to the impact of spinning screw driver" **Hybrid Model Prediction** "the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

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"

✳Set of actions: 1. Push an object in a specific direction. 2. Rotate an object clockwise or anti-clockwise. 3. Remove an object from the scene. 4. Drop an object on another object.

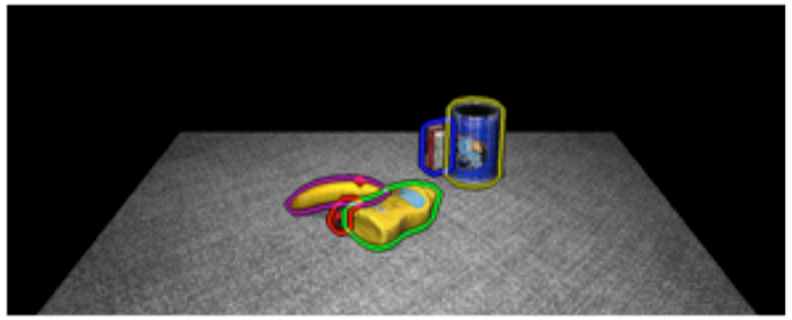
Answering visual what-if questions: from actions to predicted scene descriptions

 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"

- Set of actions: 1. Push an object in a specific direction. 2. Rotate an object clockwise or anti-clockwise. 3. Remove an object from the scene. 4. Drop an object on another object.

- Single type questions: If I do …, what happens to …?

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"

- Set of actions: 1. Push an object in a specific direction. 2. Rotate an object clockwise or anti-clockwise. 3. Remove an object from the scene. 4. Drop an object on another object.

- Single type questions: If I do …, what happens to …?

- A hybrid model using a physics engine