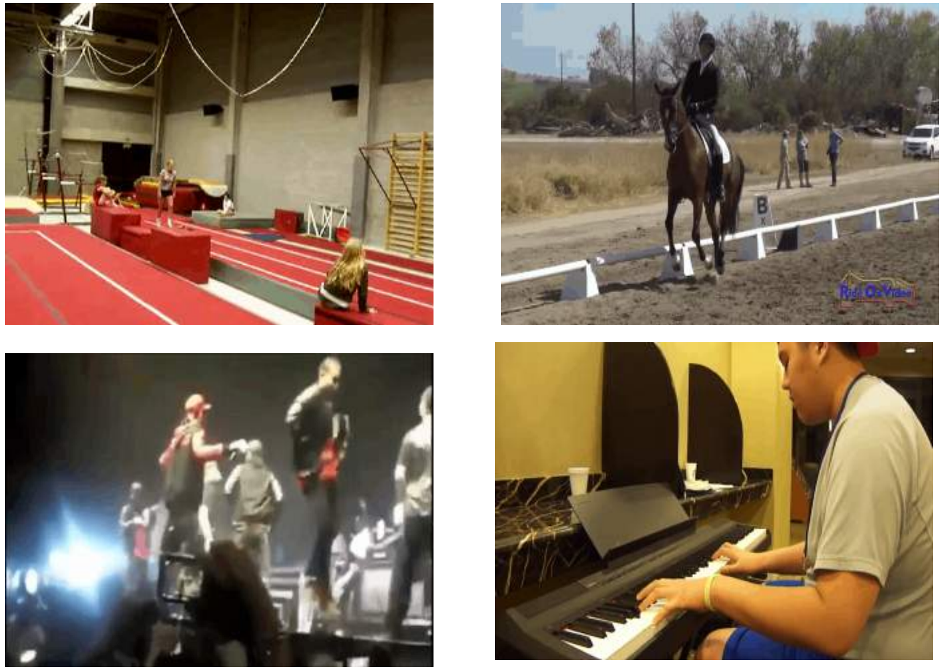
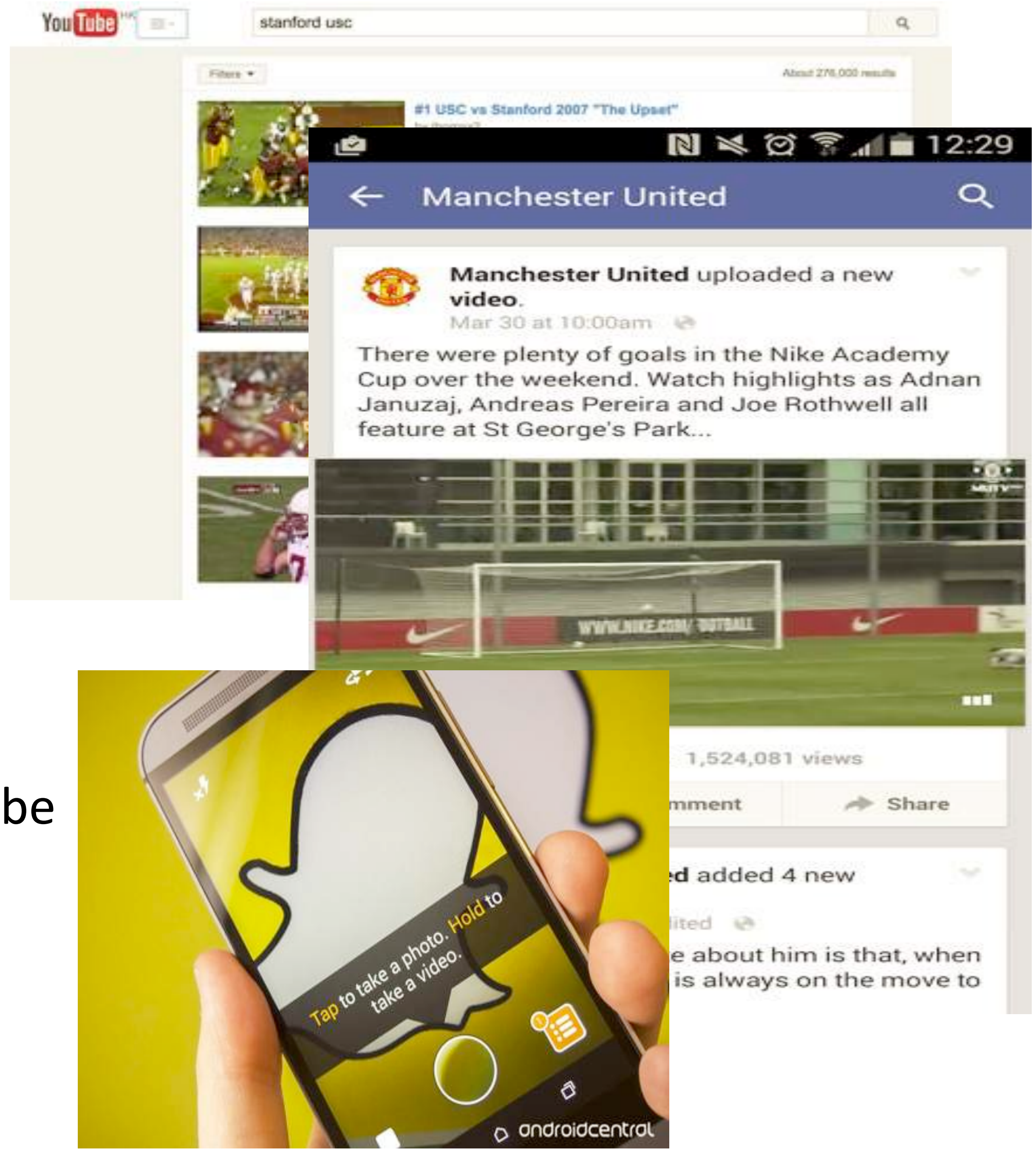
Structured Representations for Video Understanding



Chuang Gan

MIT-IBM
Watson
AI Lab

People also love to share their videos!



Manchester United uploaded a new **video**.
Mar 30 at 10:00am

There were plenty of goals in the Nike Academy Cup over the weekend. Watch highlights as Adnan Januzaj, Andreas Pereira and Joe Rothwell all feature at St George's Park...

ed added 4 new
ited
e about him is that, when
is always on the move to

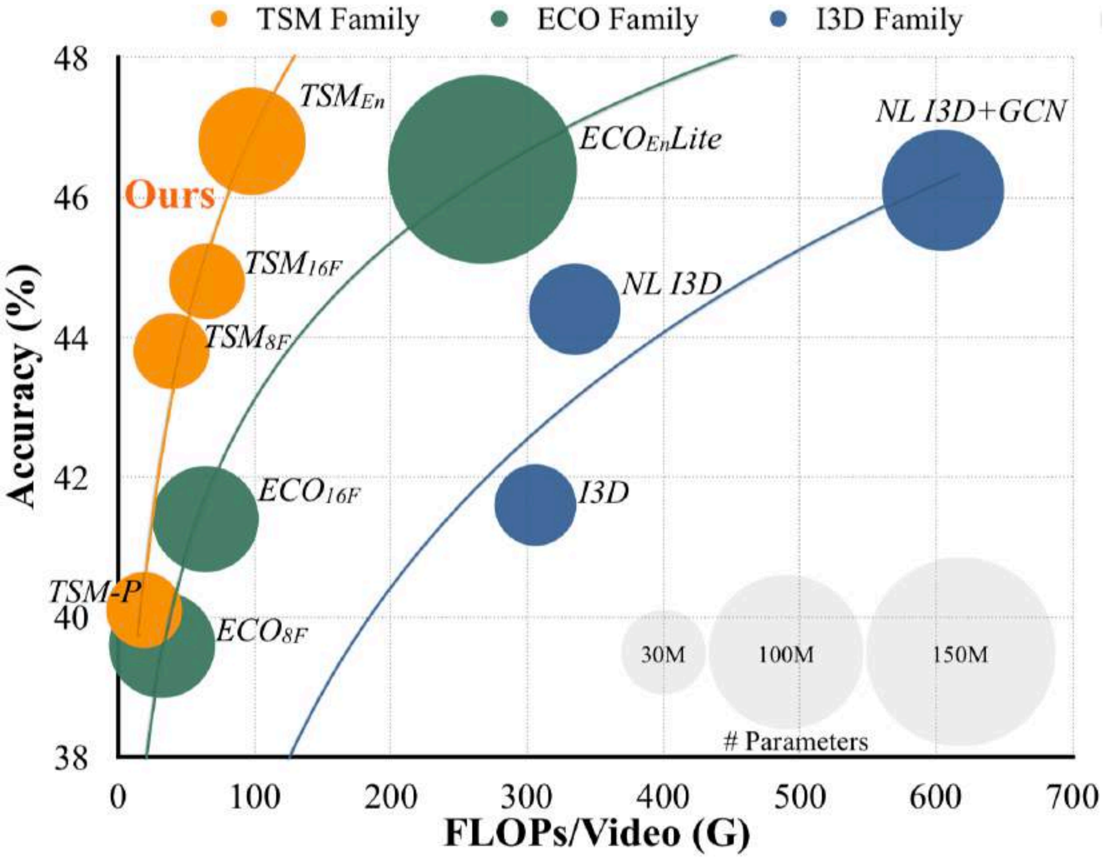
300 hours of new YouTube video every minute.

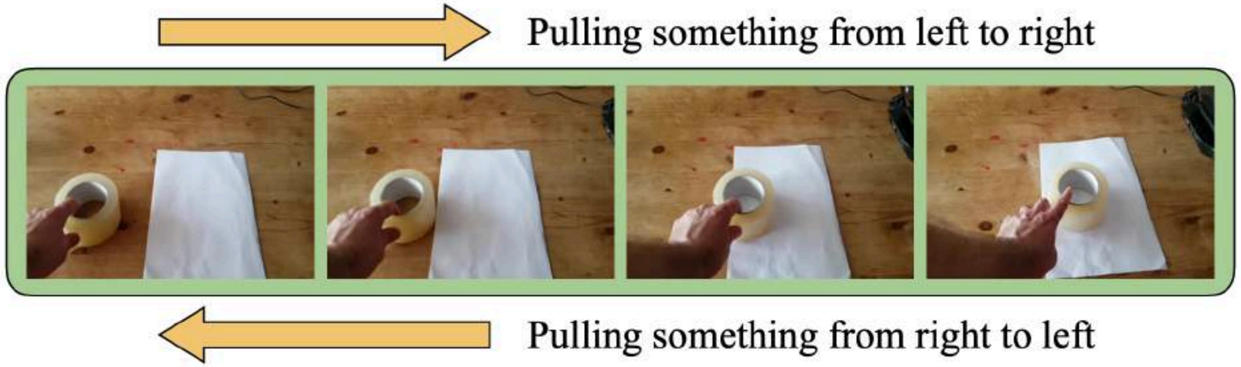
Temporal Shift Module (TSM)

Ji Lin

Chuang Gan

Song Han



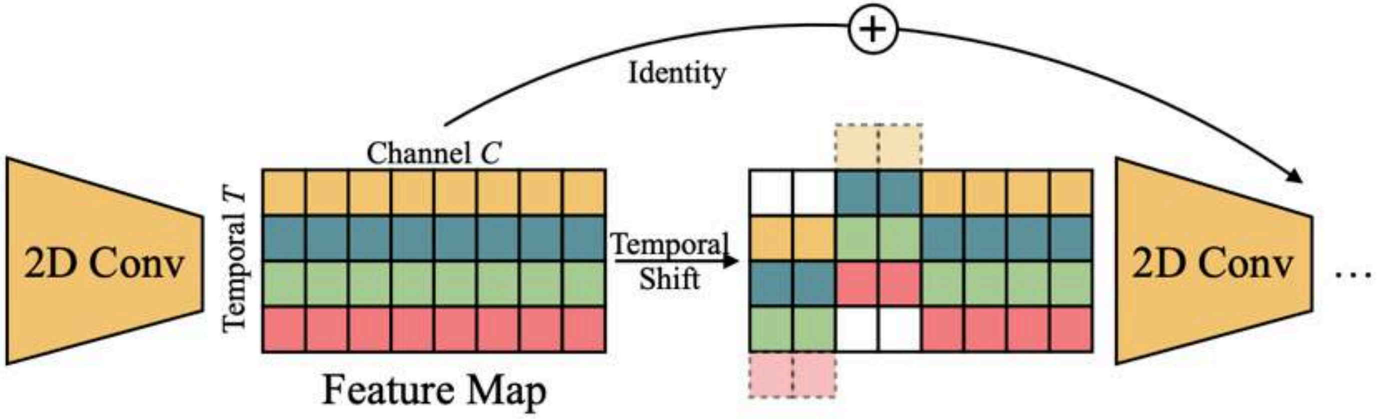


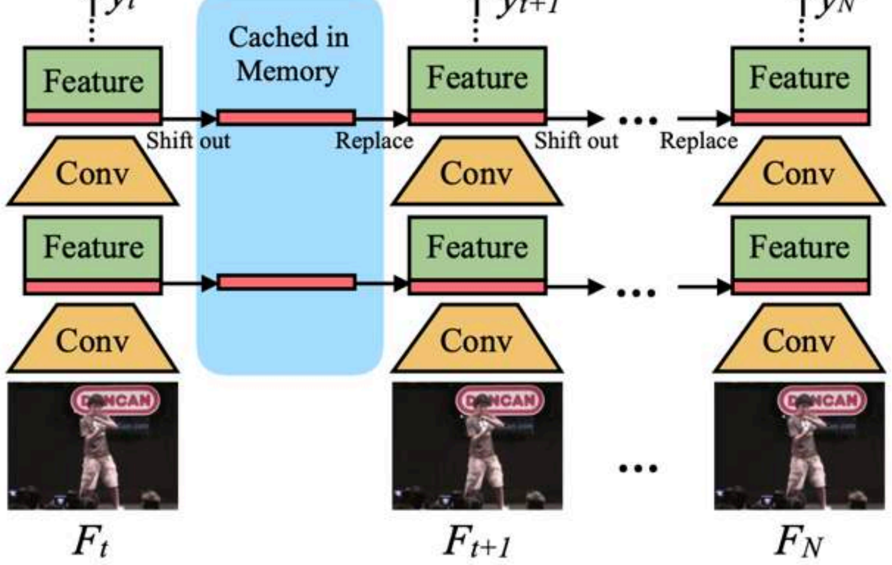
Pulling something from left to right
Pulling something from right to left

TSM: Temporal Shift Module for Efficient Video Understanding. ICCV 2019.

Temporal Shift Module (TSM)

- TSM shifts part of the channels along the temporal dimension to facilitate information exchange
- Support online/offline setting
- It can enable temporal modeling at the cost of *zero FLOPs and zero parameters*





TSM: Accurate and Efficient

Latency Comparison

TSM is **9x** faster than 3D CNN

Measured on NVIDIA Tesla P100. Batch size=1

I3D:

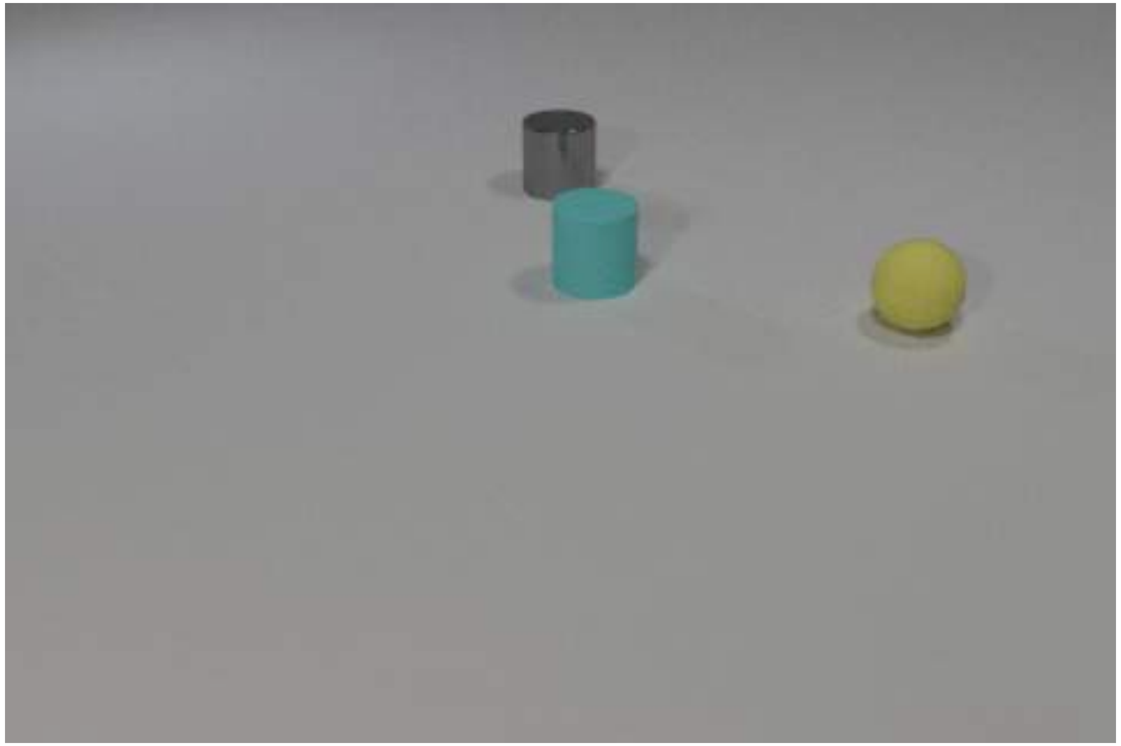
Latency: **164.3** ms/video Acc.: **41.6**%

TSM:

Latency: **17.4** ms/video Acc.: **43.4**%

Have machines achieved the human-level intelligences on video understanding?

What can humans reason about this video?



- Temporal reasoning

“What shape is the second object that collides with the cyan cylinder?”

- Causal reasoning

“What objects are responsible for the second collision?”

“What will happen next?”

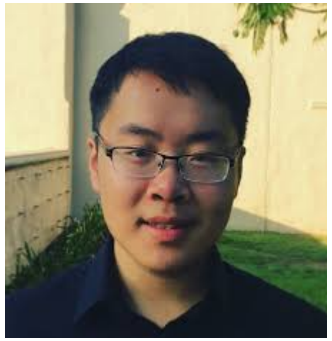
“What would happen without the gray object?”

Can machines understand beyond visual context and reason about causality?

CLEVRER: CoLlision Events for Video REpresentation and Reasoning

http://clevrer.csail.mit.edu

ICLR 2020



Kexin Yi*

Chuang Gan*

Yunzhu Li

Pushmeet Kohli

Jiajun Wu

Antonio Torralba

Josh Tenenbaum

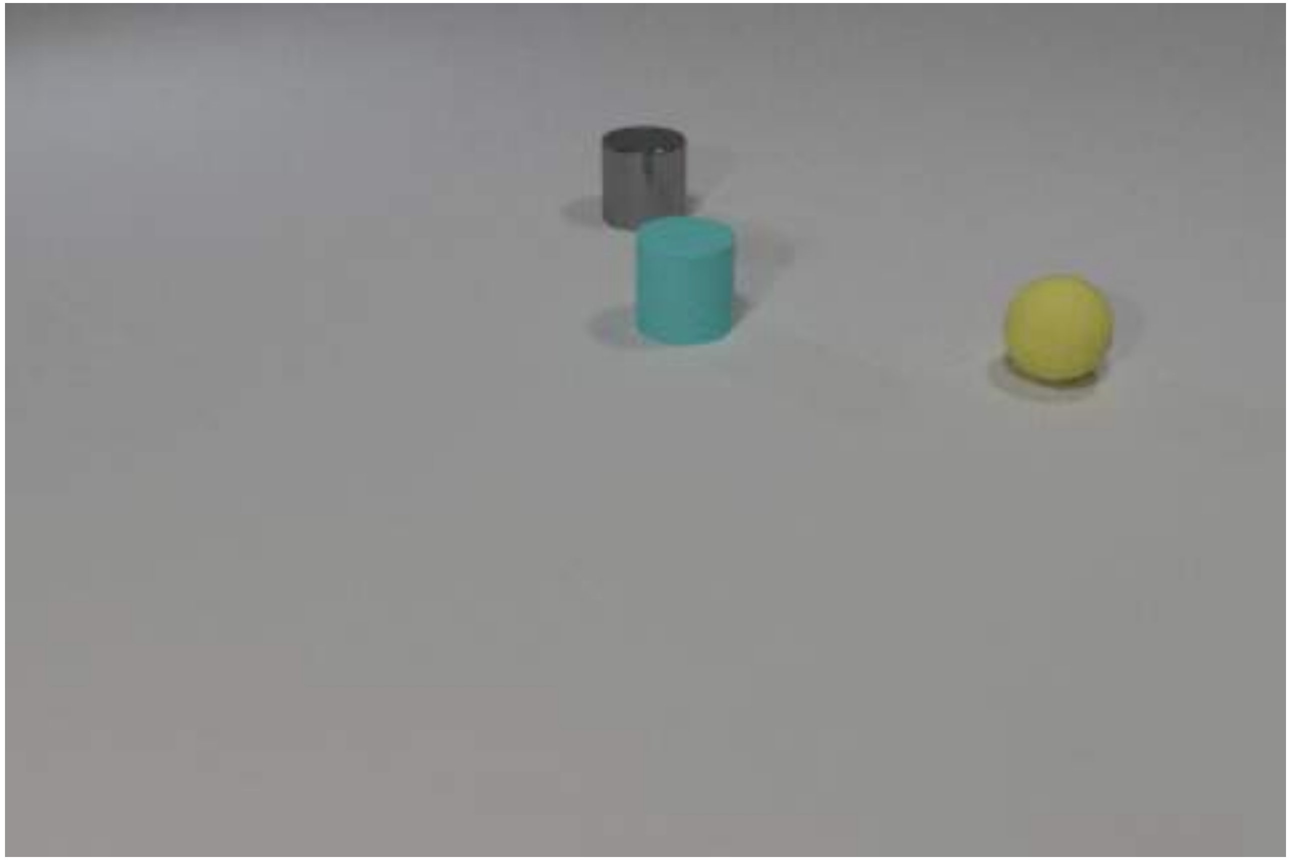
MIT-IBM
Watson
AI Lab

DeepMind

(* equal contributions)

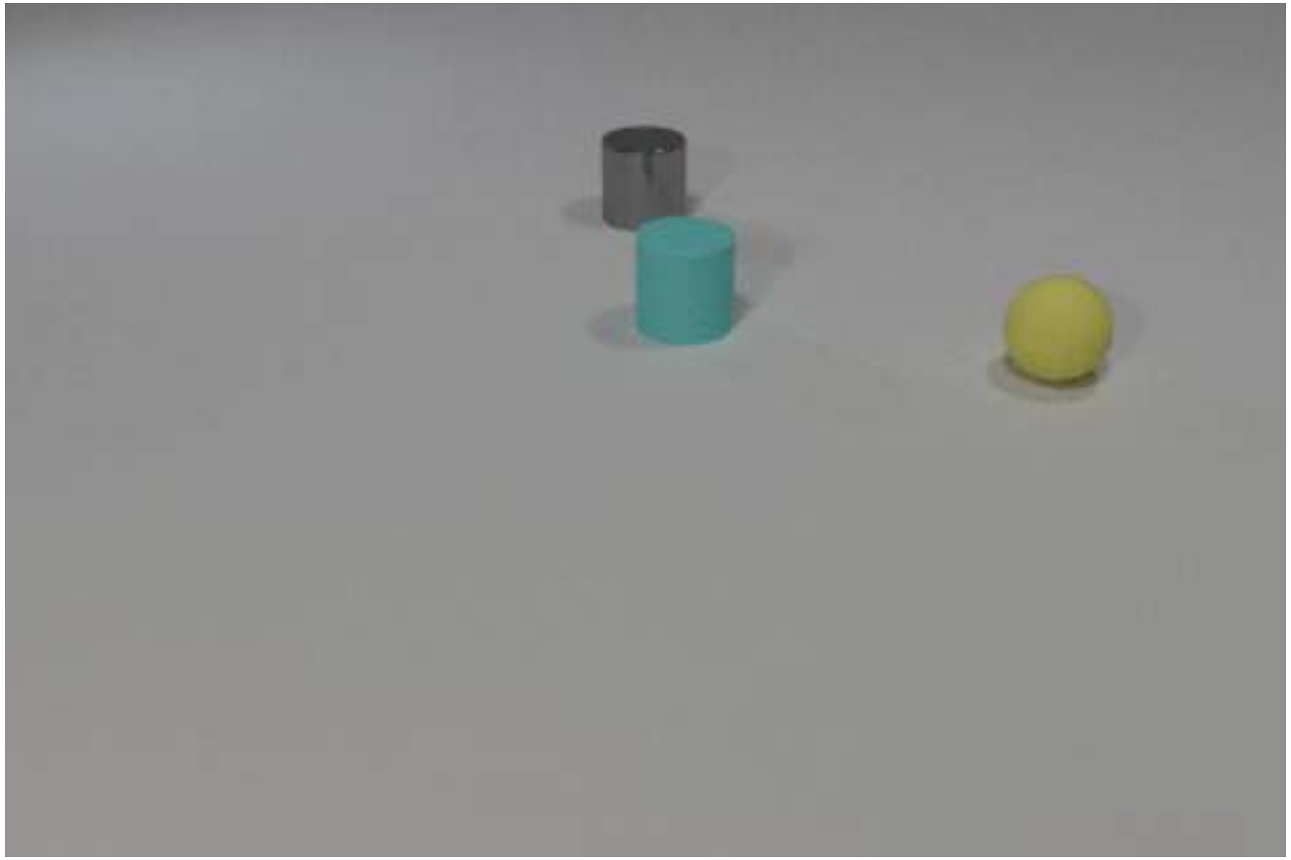
CLEVRER Dataset

- 20,000 Synthetic videos and 300,000+ questions
- Controlled biases
- Diagnostic annotations
- Causal *reasoning*
  - Descriptive
  - Explanatory
  - Predictive
  - Counterfactual



Descriptive Reasoning

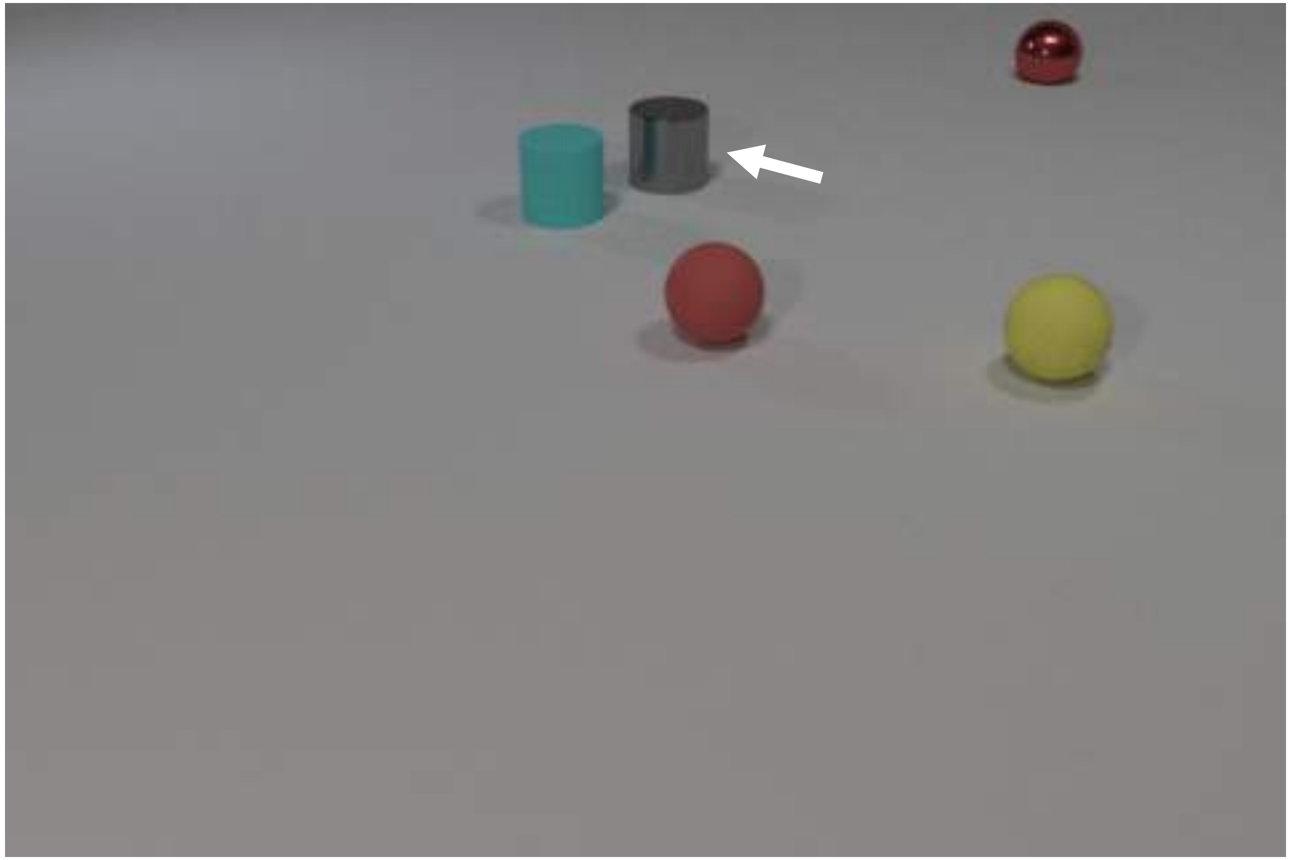
Q: *What is the material of the last object to collide with the cyan cylinder?*



Descriptive Reasoning

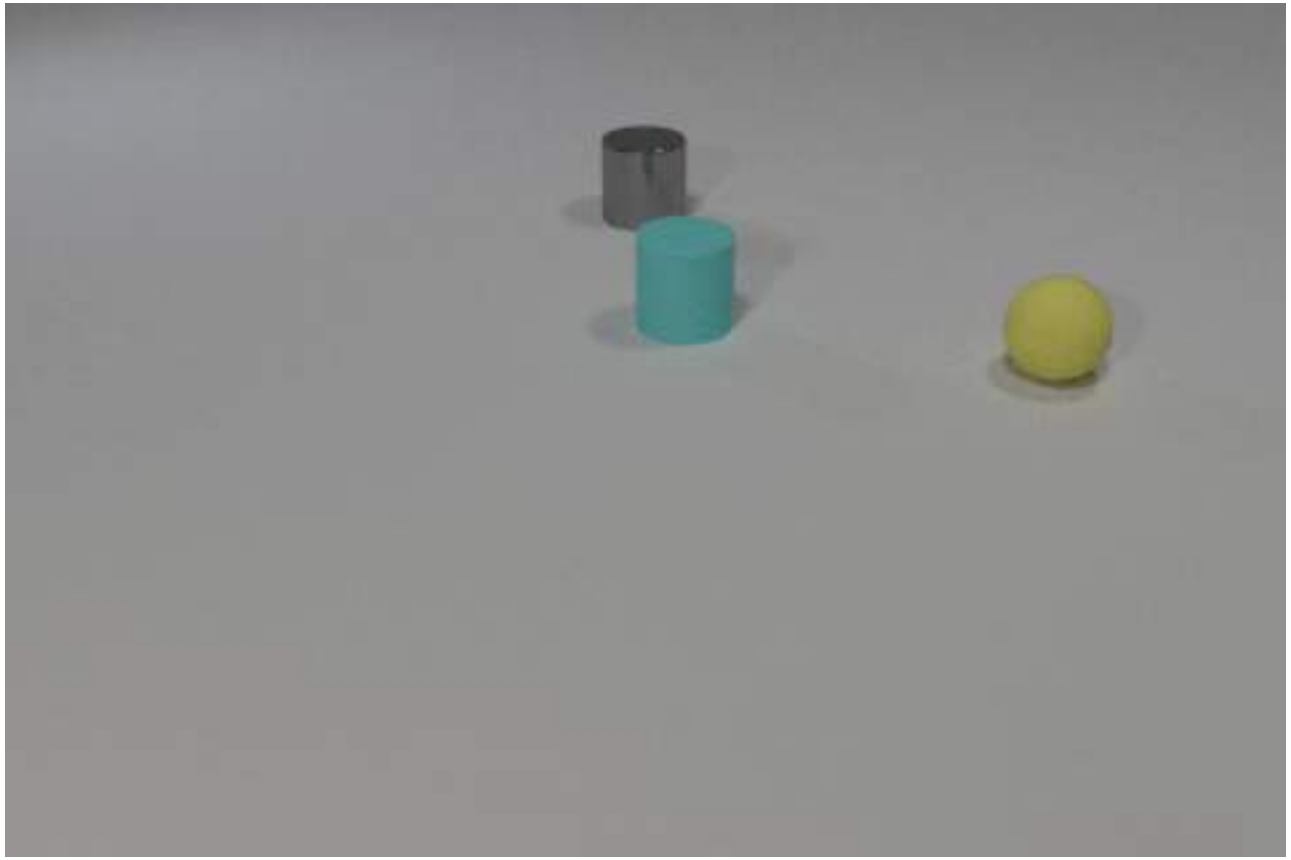
Q: *What is the material of the last object to collide with the cyan cylinder?*

A: *Metal*



Explanatory Reasoning

Q: *What is responsible for the collision between the rubber and metal cylinder?*



✗ *A. The presence of the yellow sphere*

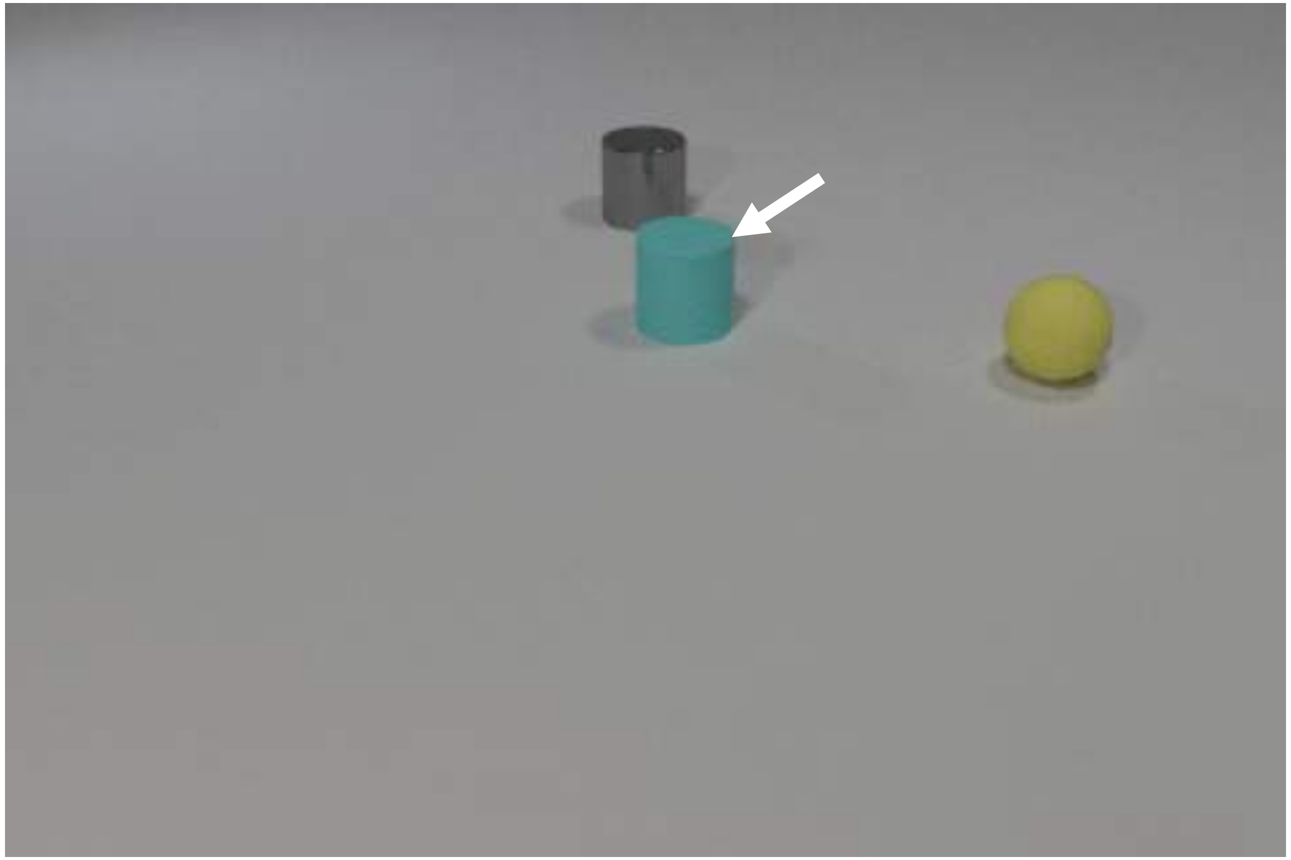
✓ *B. The collision between the rubber cylinder and the red rubber sphere*

Predictive Reasoning

Q: *What will happen next?*

✗ *A. The metal sphere and the gray cylinder collide*

✓ *B. The red rubber sphere collides with the metal sphere*

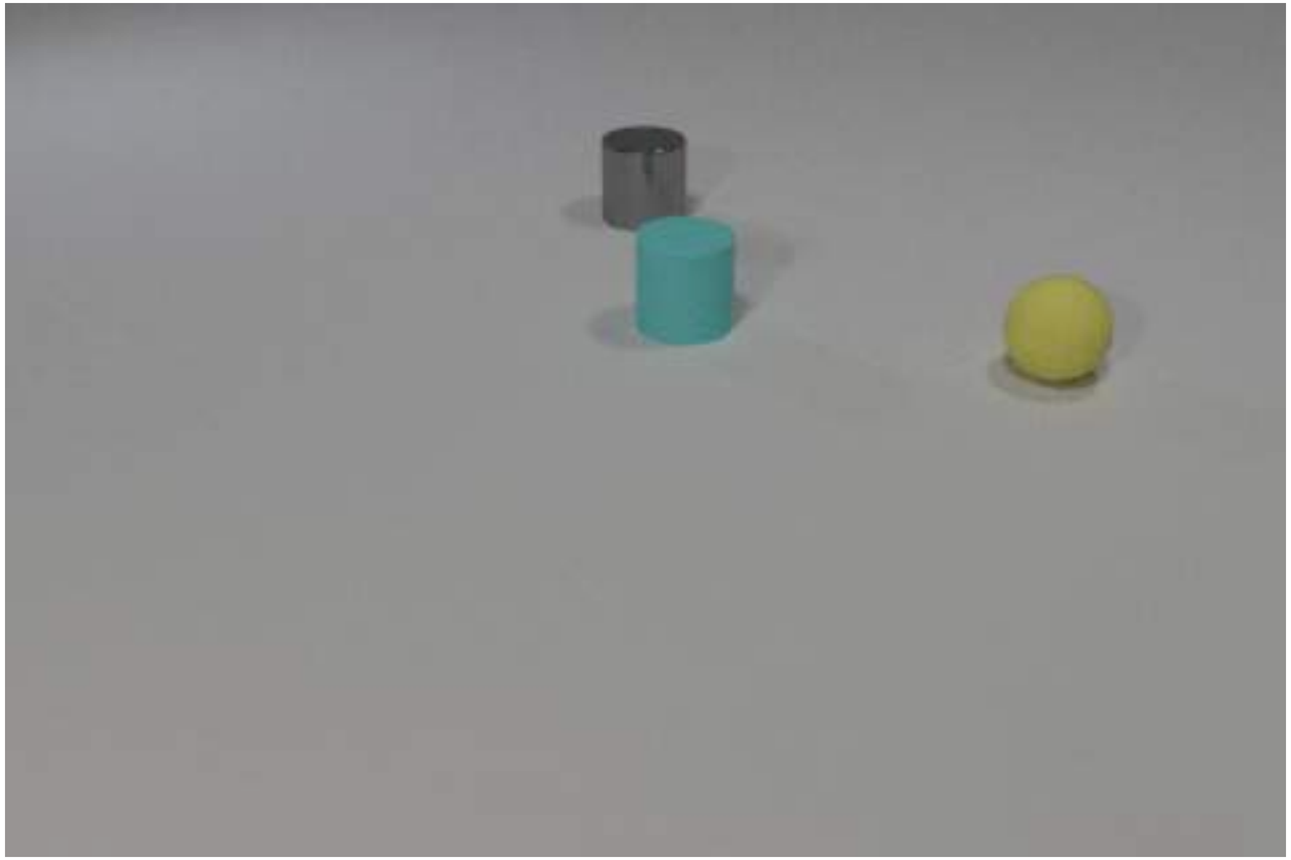


Counterfactual Reasoning

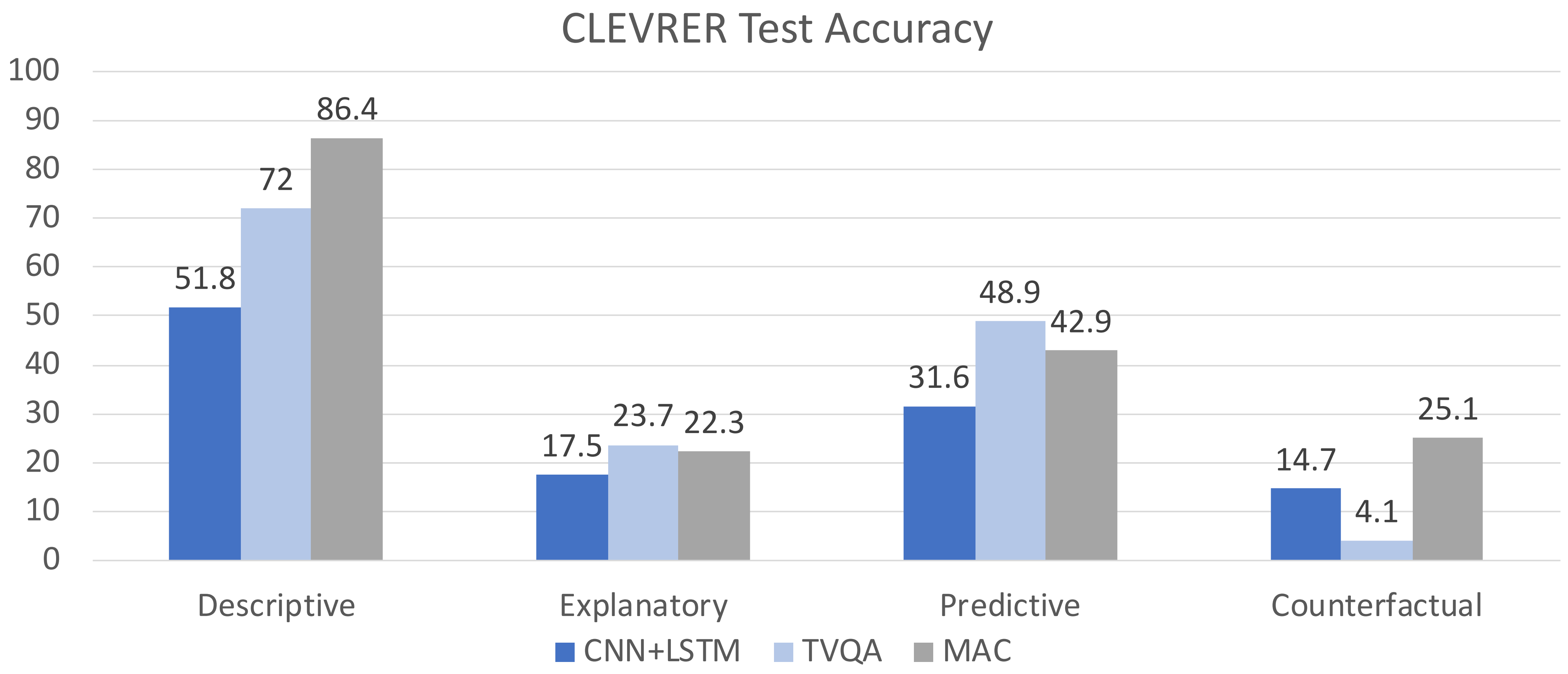
Q: *What will happen without the cyan cylinder?*

✗ *A. The red rubber sphere and the metal sphere collide*

✓ *B. The red rubber sphere and the gray object collide*



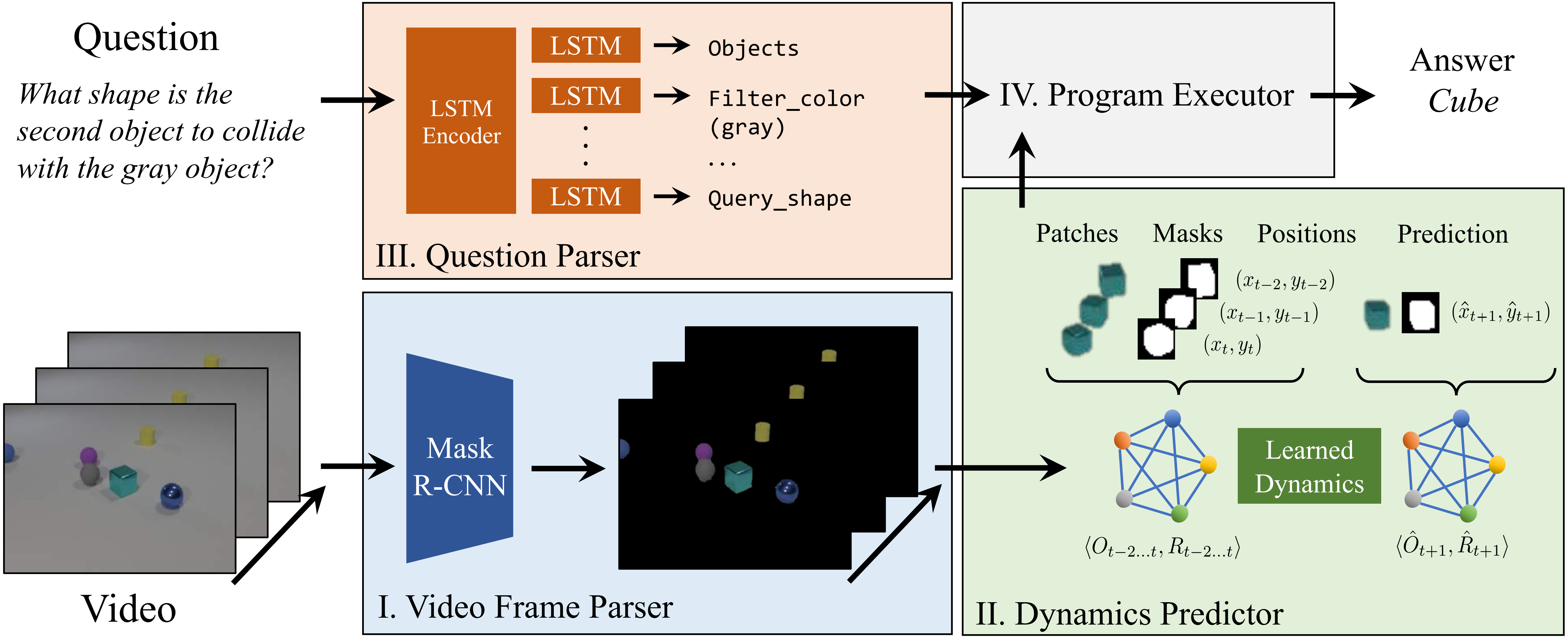
Evaluation



Causal reasoning requires the dynamics of the video's internal state

[Lei et al. EMNLP 2018]
[Hudson et al. ICLR 2019]

Neuro-Symbolic Dynamics Reasoning (NS-DR)



Question

What shape is the second object to collide with the gray object?

LSTM Encoder

LSTM → `Objects`

LSTM → `Filter_color (gray)`

⋮ `...`

LSTM → `Query_shape`

III. Question Parser

IV. Program Executor

Answer

Cube

Video

Mask R-CNN

I. Video Frame Parser

Patches

Masks

Positions

Prediction

(x_{t-2}, y_{t-2})

(x_{t-1}, y_{t-1})

(x_t, y_t)

$(\hat{x}_{t+1}, \hat{y}_{t+1})$

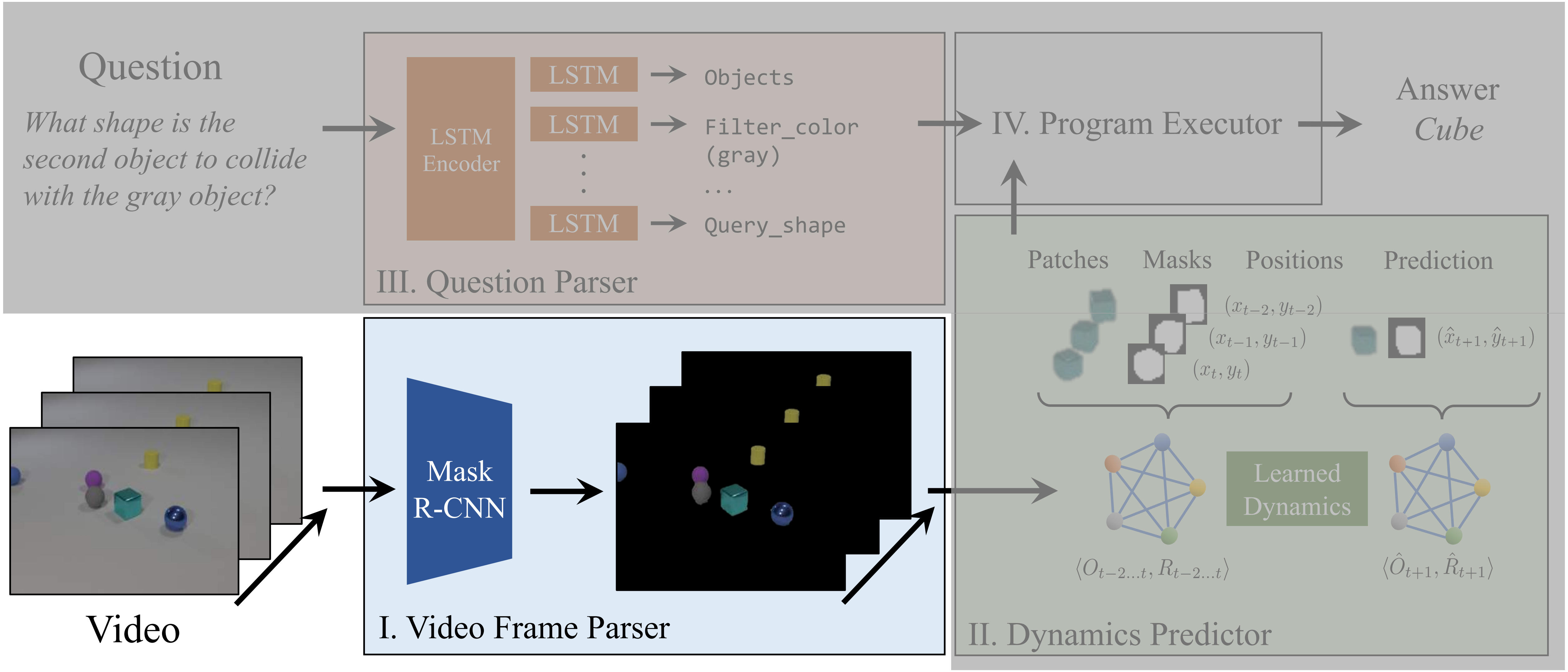
Learned Dynamics

$\langle O_{t-2\ldots t}, R_{t-2\ldots t} \rangle$

$\langle \hat{O}_{t+1}, \hat{R}_{t+1} \rangle$

II. Dynamics Predictor

Neuro-Symbolic Dynamics Reasoning (NS-DR)



Question

What shape is the second object to collide with the gray object?

LSTM Encoder

LSTM → `Objects`

LSTM → `Filter_color (gray)`

`...`

LSTM → `Query_shape`

III. Question Parser

IV. Program Executor

Answer

Cube

Patches

Masks

Positions

Prediction

(x_{t-2}, y_{t-2})

(x_{t-1}, y_{t-1})

(x_t, y_t)

$(\hat{x}_{t+1}, \hat{y}_{t+1})$

Learned Dynamics

$\langle O_{t-2...t}, R_{t-2...t} \rangle$

$\langle \hat{O}_{t+1}, \hat{R}_{t+1} \rangle$

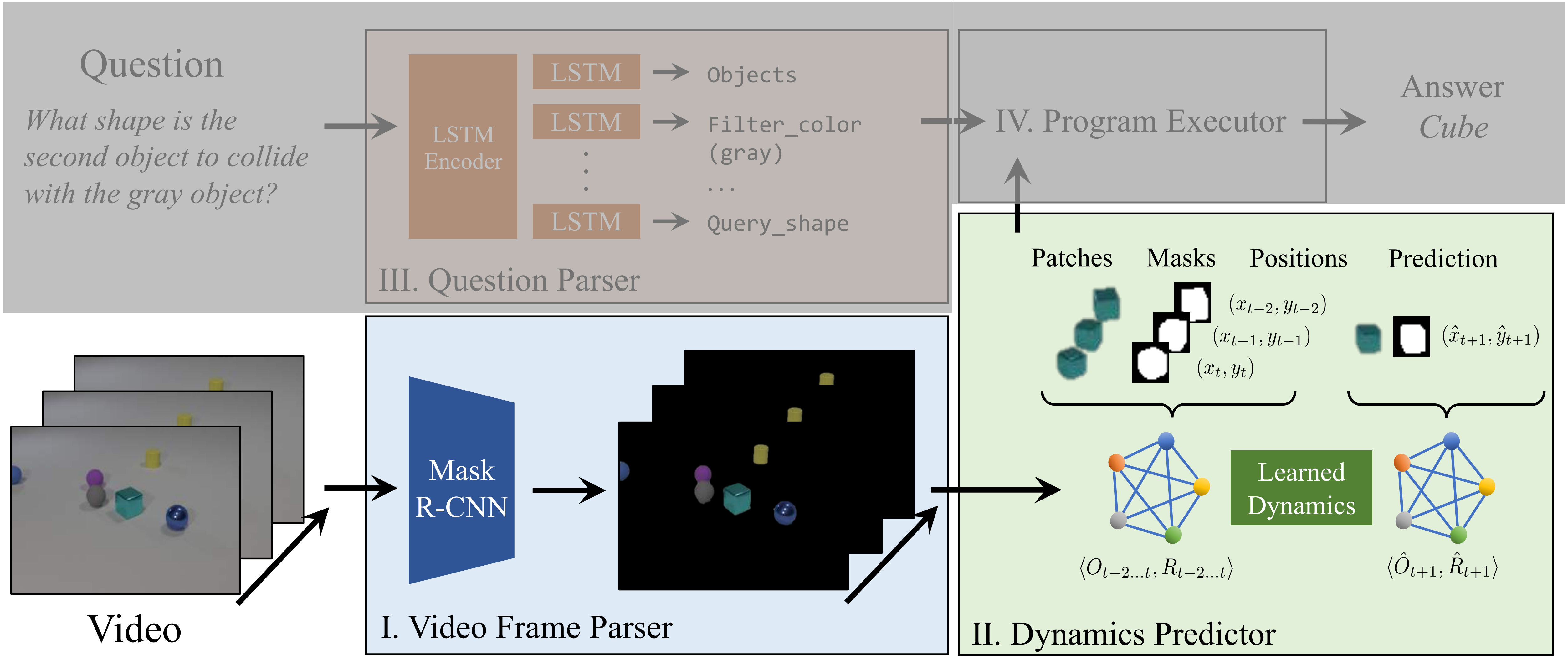
II. Dynamics Predictor

Video

Mask R-CNN

I. Video Frame Parser

Neuro-Symbolic Dynamics Reasoning (NS-DR)



Question

What shape is the second object to collide with the gray object?

LSTM Encoder

LSTM → Objects

LSTM → Filter_color (gray)

...

LSTM → Query_shape

III. Question Parser

IV. Program Executor

Answer

Cube

Patches Masks Positions Prediction

(x_{t-2}, y_{t-2})

(x_{t-1}, y_{t-1})

(x_t, y_t)

$(\hat{x}_{t+1}, \hat{y}_{t+1})$

Learned Dynamics

$\langle O_{t-2\ldots t}, R_{t-2\ldots t}\rangle$

$\langle \hat{O}_{t+1}, \hat{R}_{t+1}\rangle$

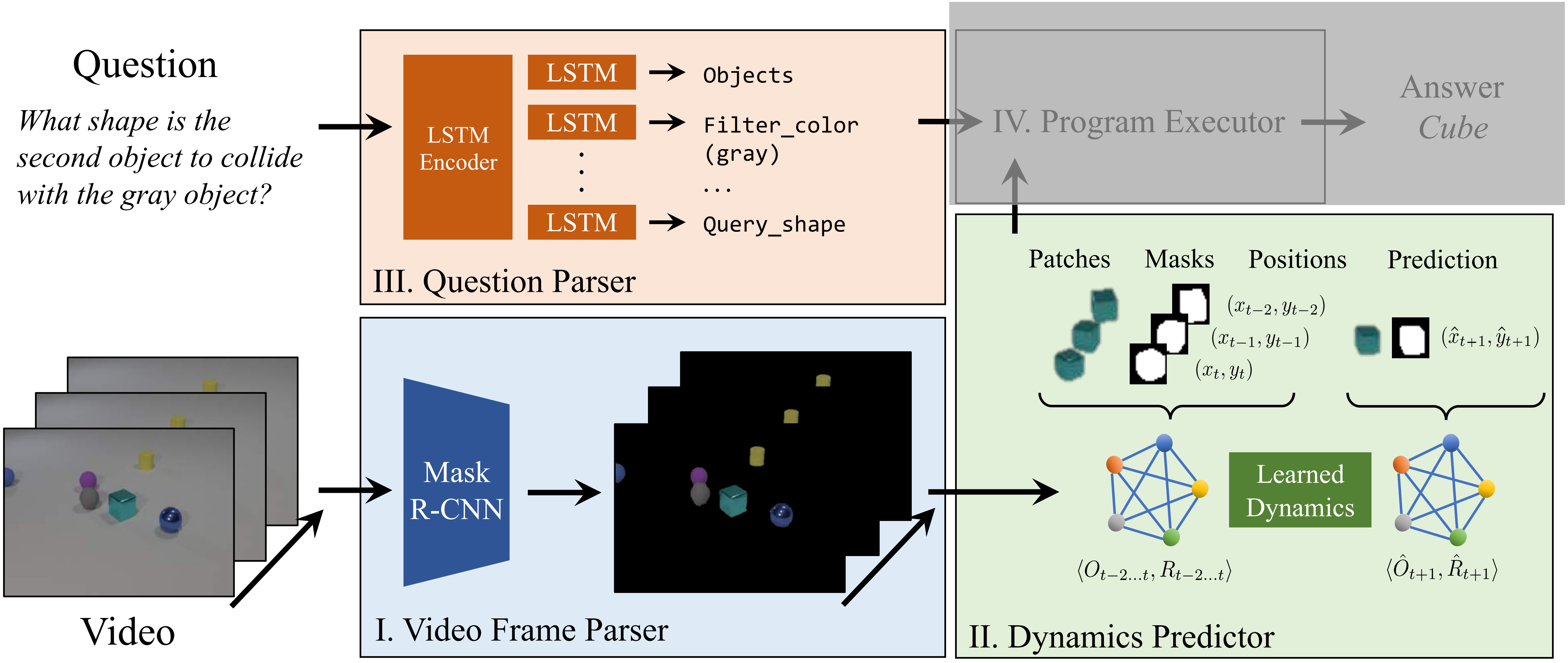
II. Dynamics Predictor

Video

Mask R-CNN

I. Video Frame Parser

Neuro-Symbolic Dynamics Reasoning (NS-DR)



Question

What shape is the second object to collide with the gray object?

LSTM Encoder

LSTM → `Objects`

LSTM → `Filter_color (gray)`

⋮ `...`

LSTM → `Query_shape`

III. Question Parser

IV. Program Executor

Answer

Cube

Video

Mask R-CNN

I. Video Frame Parser

Patches

Masks

Positions

Prediction

(x_{t-2}, y_{t-2})

(x_{t-1}, y_{t-1})

(x_t, y_t)

$(\hat{x}_{t+1}, \hat{y}_{t+1})$

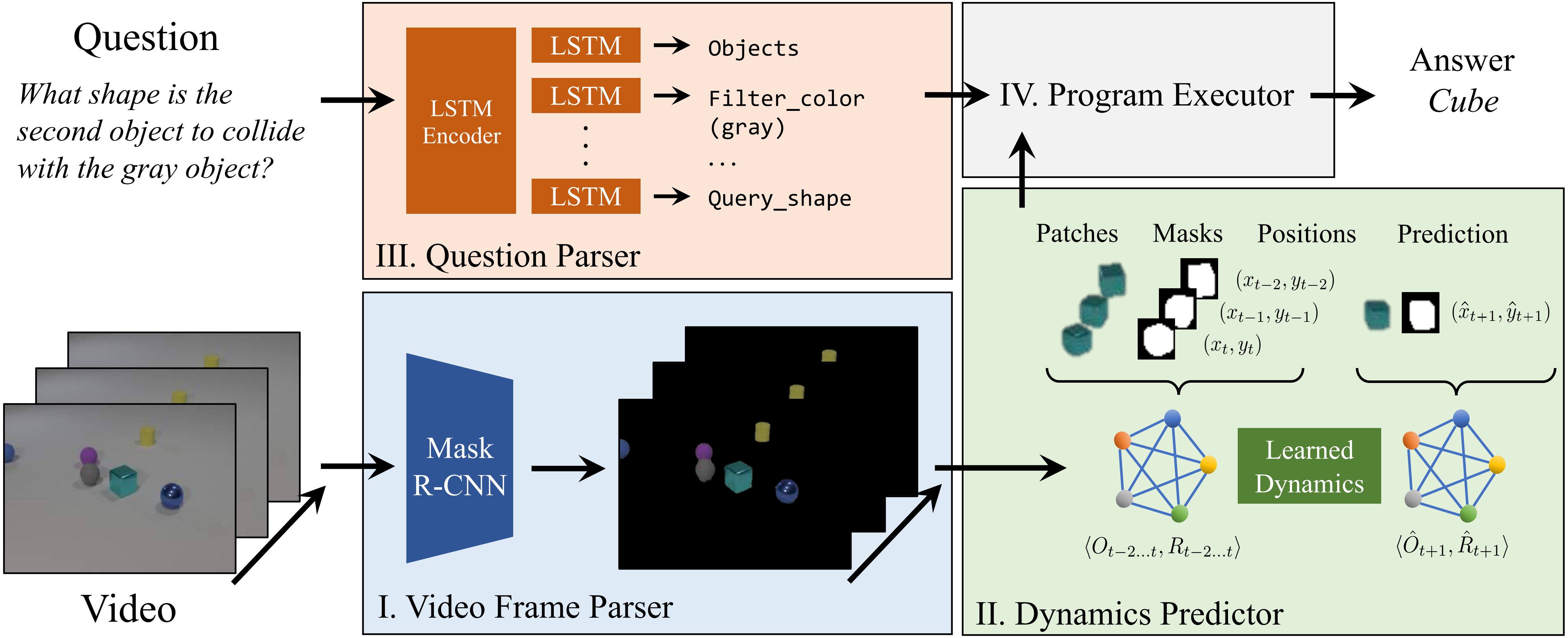
Learned Dynamics

$\langle O_{t-2...t}, R_{t-2...t} \rangle$

$\langle \hat{O}_{t+1}, \hat{R}_{t+1} \rangle$

II. Dynamics Predictor

Neuro-Symbolic Dynamics Reasoning (NS-DR)



Question

What shape is the second object to collide with the gray object?

LSTM Encoder

LSTM → `Objects`

LSTM → `Filter_color (gray)`

LSTM → `Query_shape`

III. Question Parser

IV. Program Executor

Answer

Cube

Patches

Masks

Positions

Prediction

(x_{t-2}, y_{t-2})

(x_{t-1}, y_{t-1})

(x_t, y_t)

$(\hat{x}_{t+1}, \hat{y}_{t+1})$

Learned Dynamics

$\langle O_{t-2\ldots t}, R_{t-2\ldots t} \rangle$

$\langle \hat{O}_{t+1}, \hat{R}_{t+1} \rangle$

II. Dynamics Predictor

Mask R-CNN

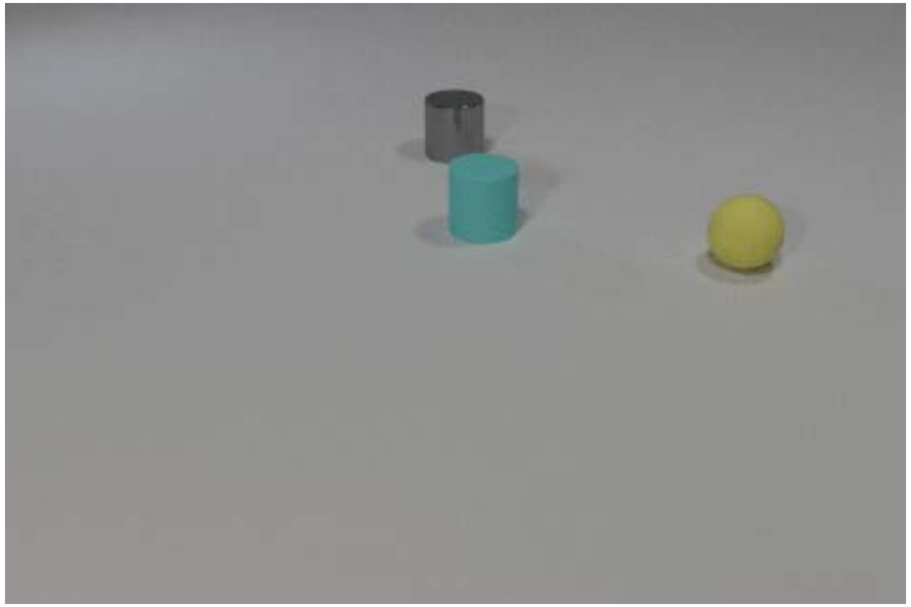
I. Video Frame Parser

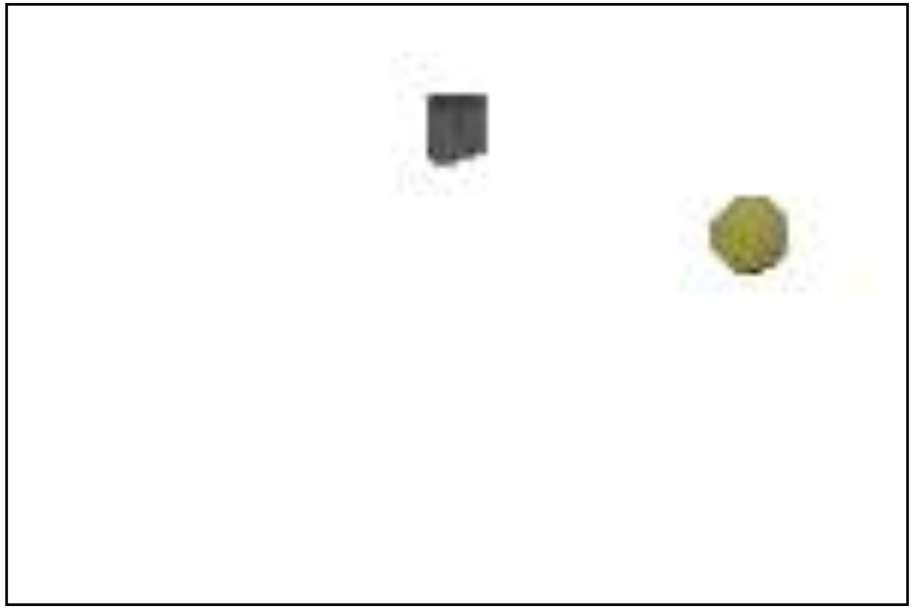
Video

Example: Counterfactual Dynamics Rollout

Example: Counterfactual Dynamics Rollout

- (Remove cyan cylinder)





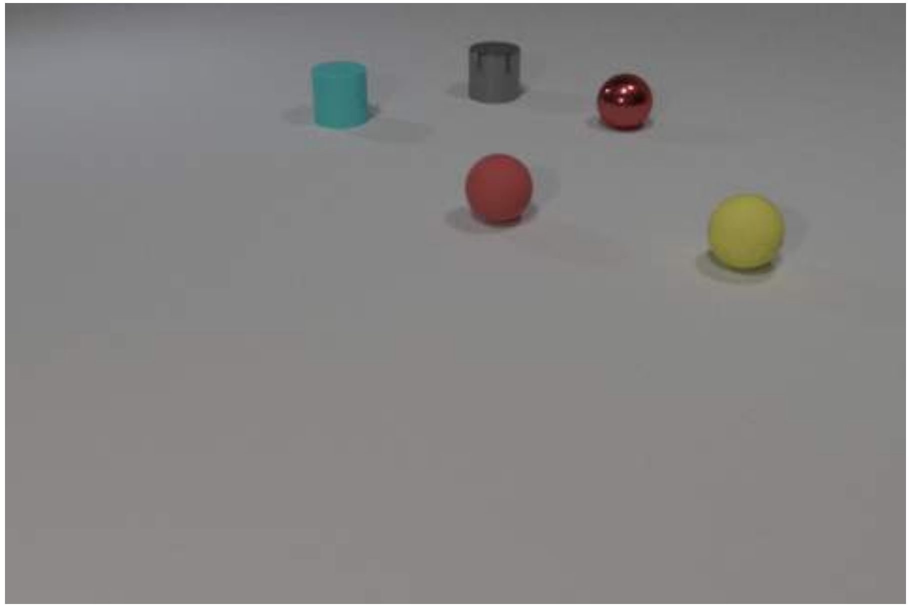
Dynamics predictor output

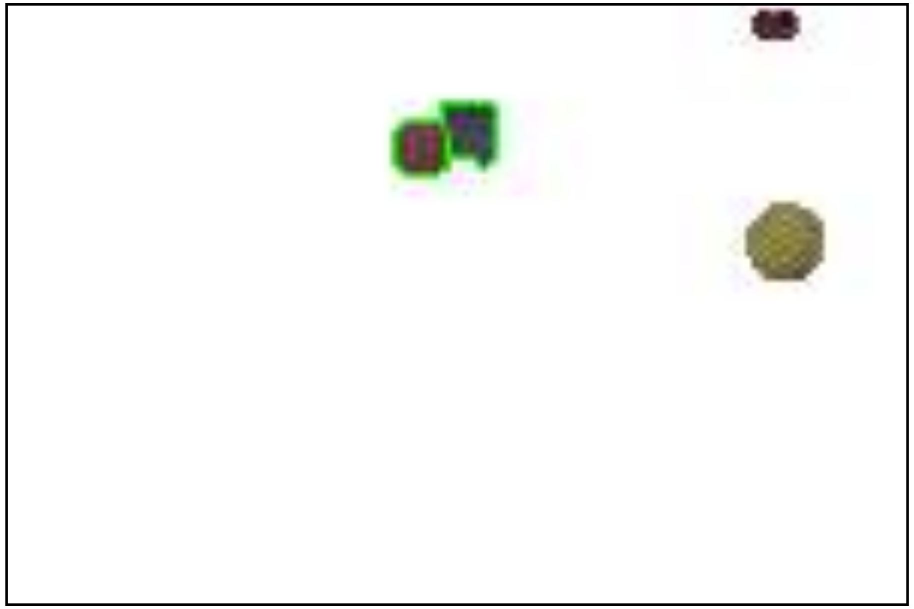
Example: Counterfactual Dynamics Rollout

- (Remove cyan cylinder)

Counterfactual events:

Type: Collision
Frame: 75
Objects: Red rubber sphere vs. Gray metal cylinder





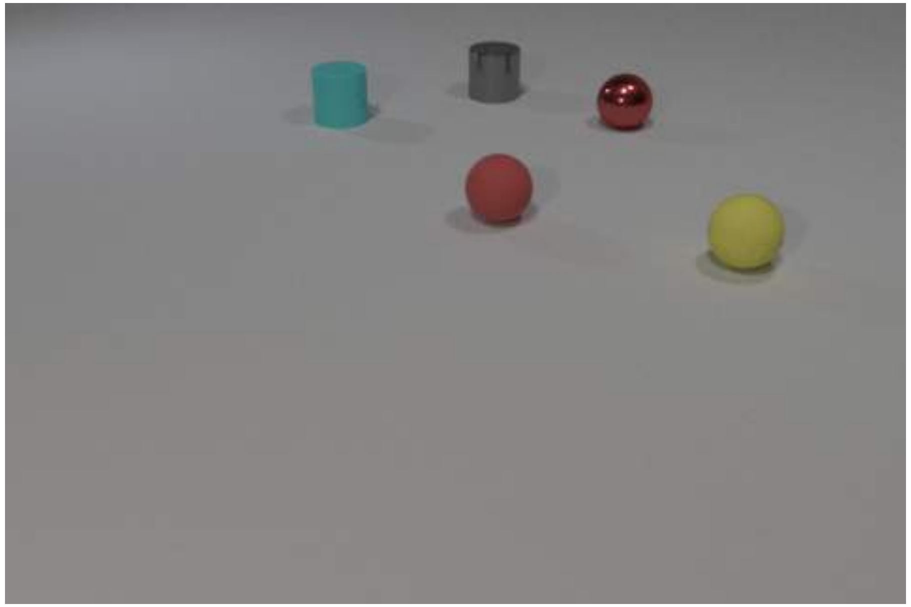
Dynamics predictor output

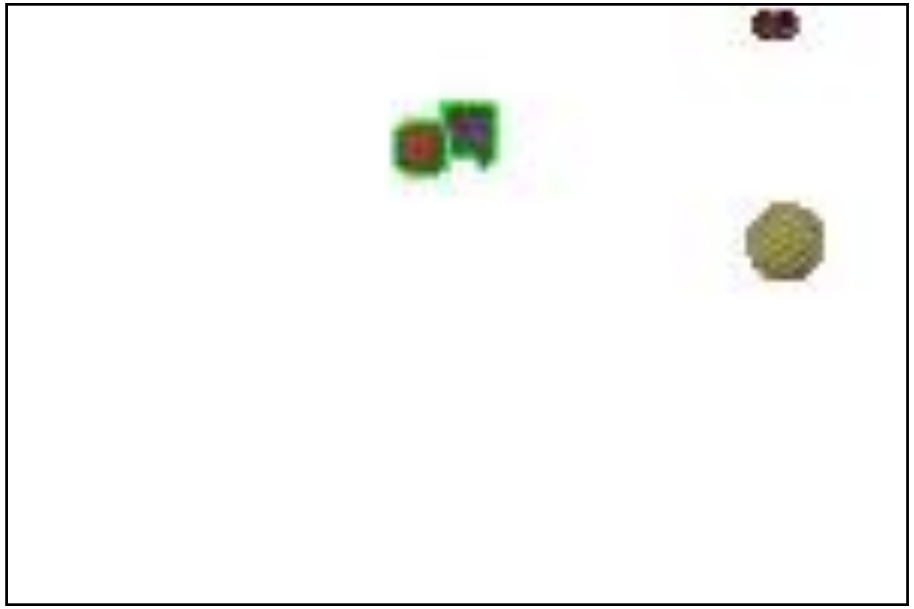
Example: Counterfactual Dynamics Rollout

- (Remove cyan cylinder)

Counterfactual events:

Type: Collision
Frame: 75
Objects: Red rubber sphere vs. Gray metal cylinder

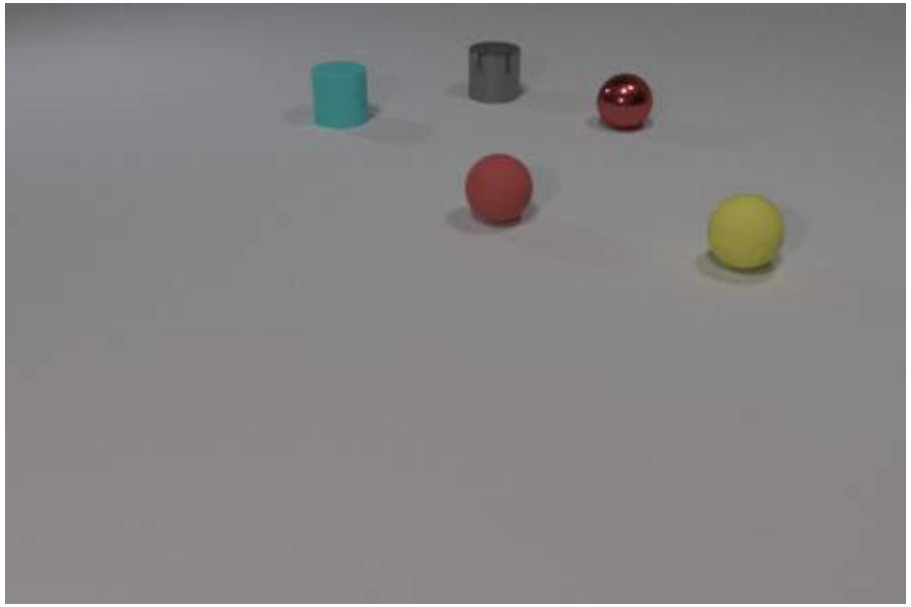




Dynamics predictor output

Example: Counterfactual Dynamics Rollout

- (Remove cyan cylinder)

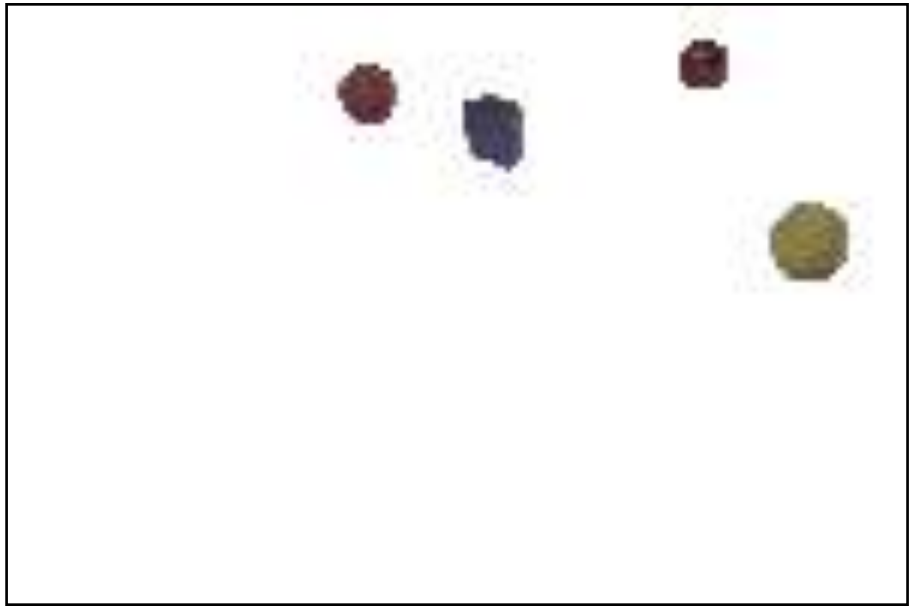


Counterfactual events:

Type: Collision
Frame: 75
Objects: Red rubber sphere vs. Gray metal cylinder

Question: *What will happen without the cyan cylinder?*

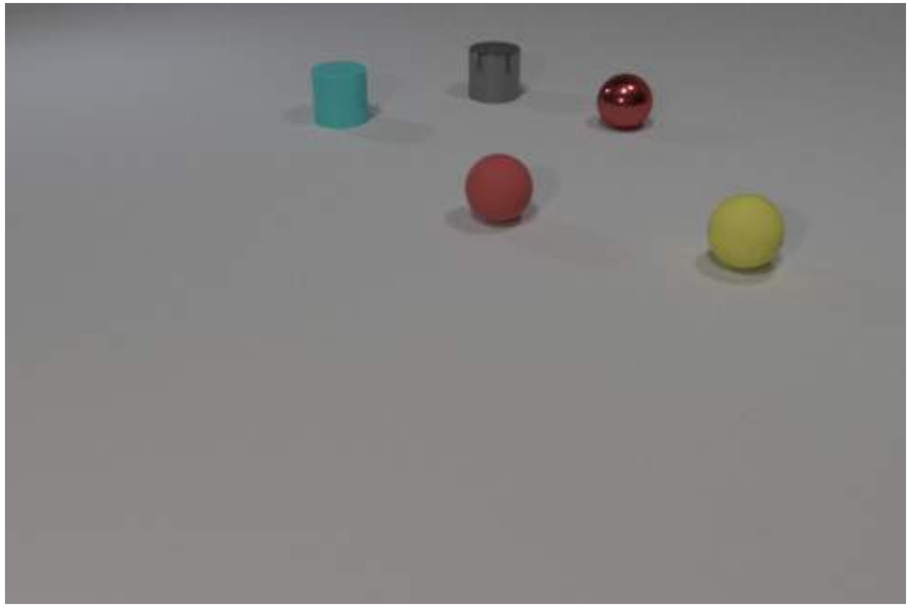
Choice: *The red rubber sphere collides with the metal cylinder*



Dynamics predictor output

Example: Counterfactual Dynamics Rollout

- (Remove cyan cylinder)



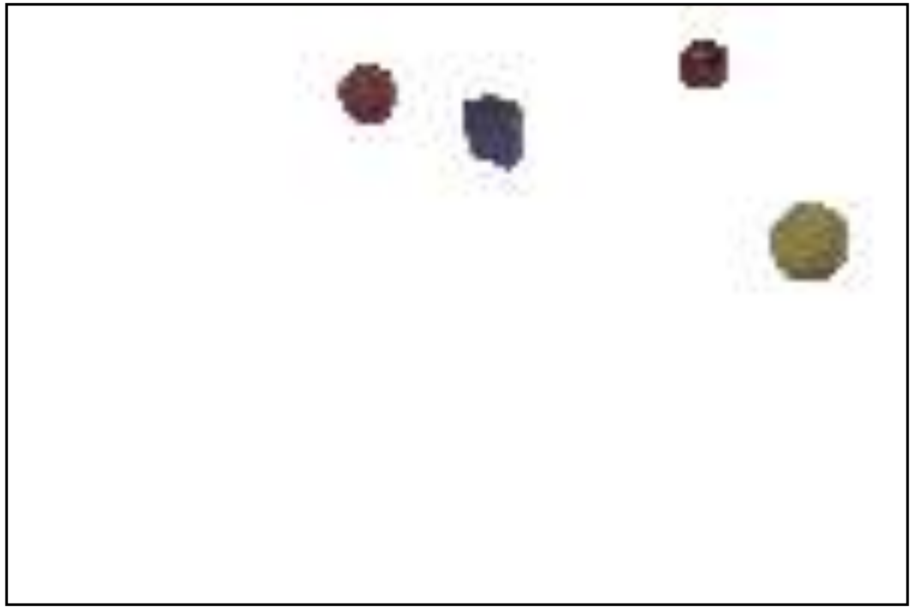
Counterfactual events:

Type: Collision
Frame: 75
Objects: Red rubber sphere vs. Gray metal cylinder

Question: *What will happen without the cyan cylinder?*

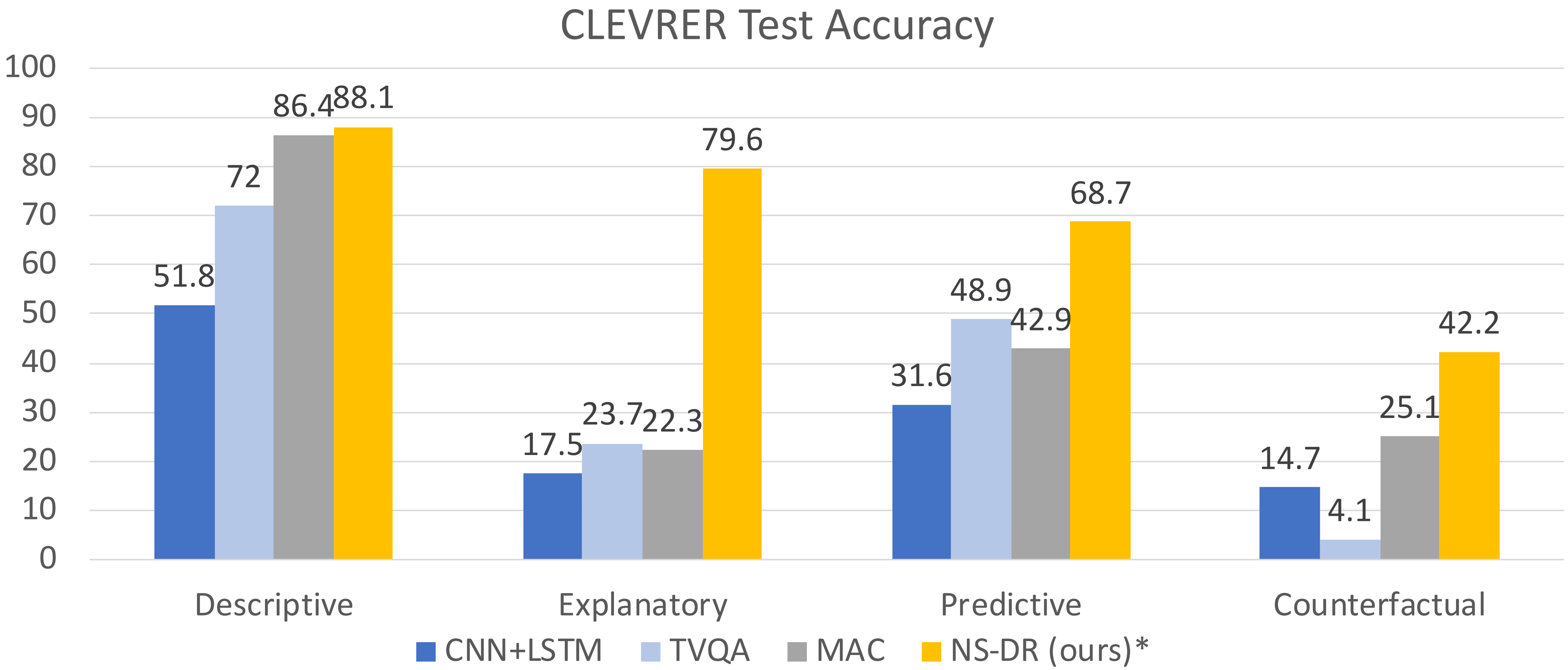
⟶ Answer: *Yes*

Choice: *The red rubber sphere collides with the metal cylinder*



Dynamics predictor output

Evaluation



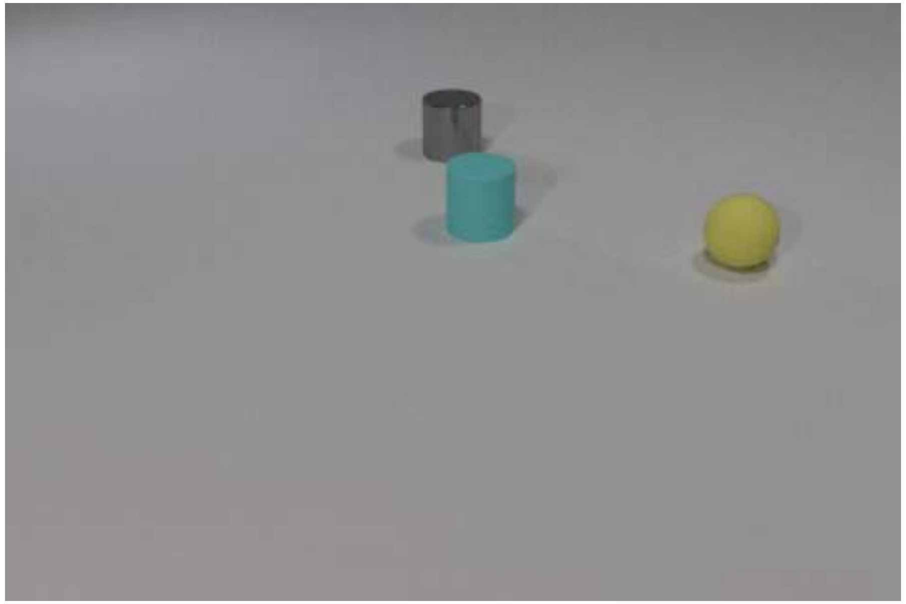
[Lei et al. EMNLP 2018]

[Hudson et al. ICLR 2019]

*: Extra Supervision

Take home message 1

- CLEVRER: video dataset for diagnostic temporal and causal reasoning
  - Descriptive: *"How many…?", "What color…?", "Is there…?"*
  - Explanatory: *"What is responsible for…?"*
  - Predictive: *"What will happen…?"*
  - Counterfactual: *"What if…?"*
- Neuro-Symbolic Dynamics Reasoning (NS-DR)
  - Dynamics modeling on video state
  - Object-centric representation
  - Symbolic program execution
- Download
  - http://clevrer.csail.mit.edu



Music Gesture for Visual Sound Separation

http://music-gesture.csail.mit.edu

CVPR 2020

Chuang Gan

Deng Huang

Hang Zhao

Josh Tenenbaum

Antonio Torralba

MIT-IBM
Watson
AI Lab

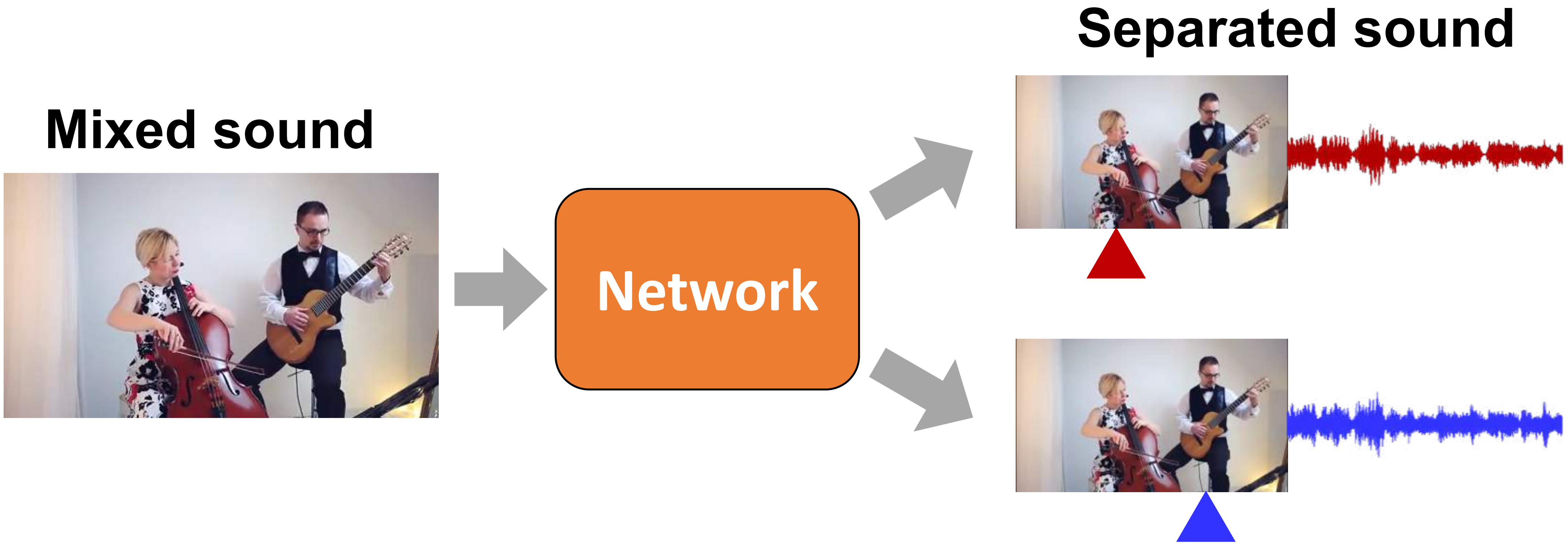
Task: visual sound separation

Given a music performance video…

Mixed sound

Task: visual sound separation

…we aim to separate two sounds played by different instruments.



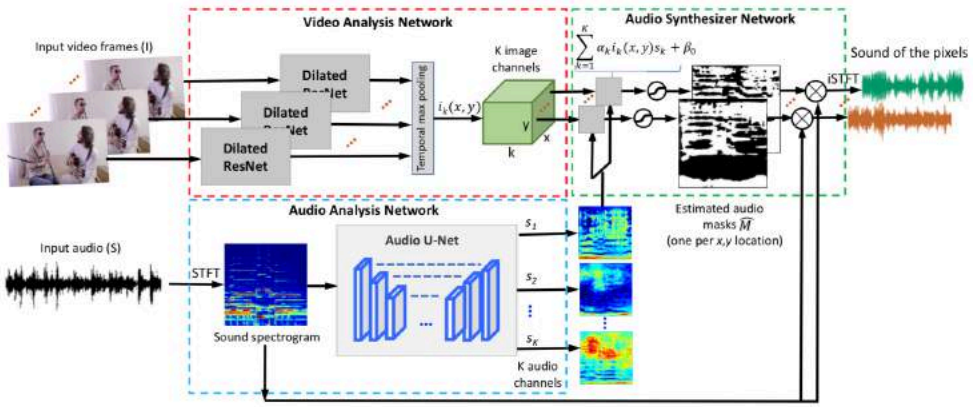
Mixed sound

Network

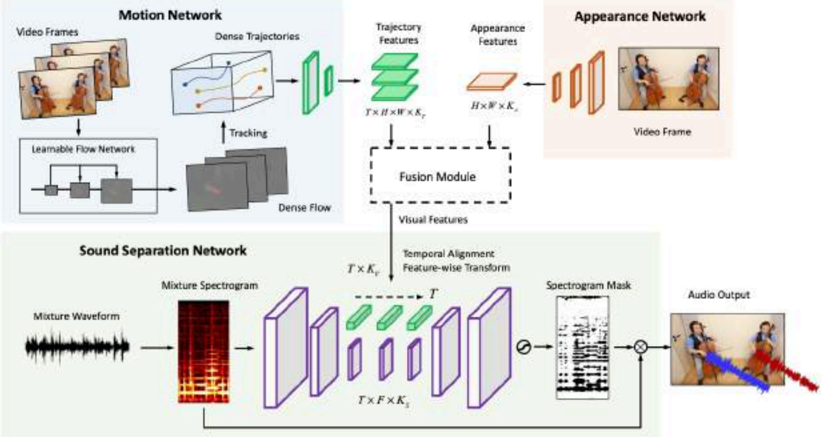
Separated sound

Challenges

- Most existing methods use **raw pixel** or **optical flow** as input.



The Sound of Pixels. ECCV 2018

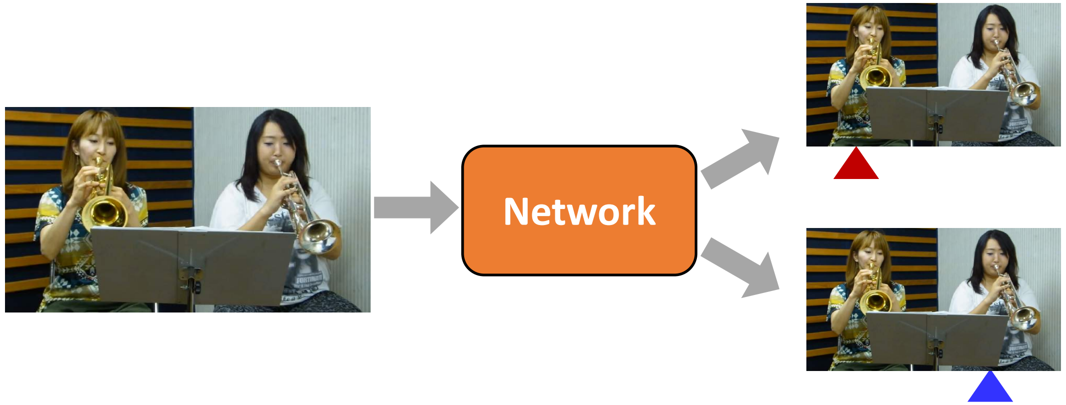


The Sound of Motions. ICCV 2019

Challenges

- Most existing methods use **raw pixel** or **optical flow** as input.

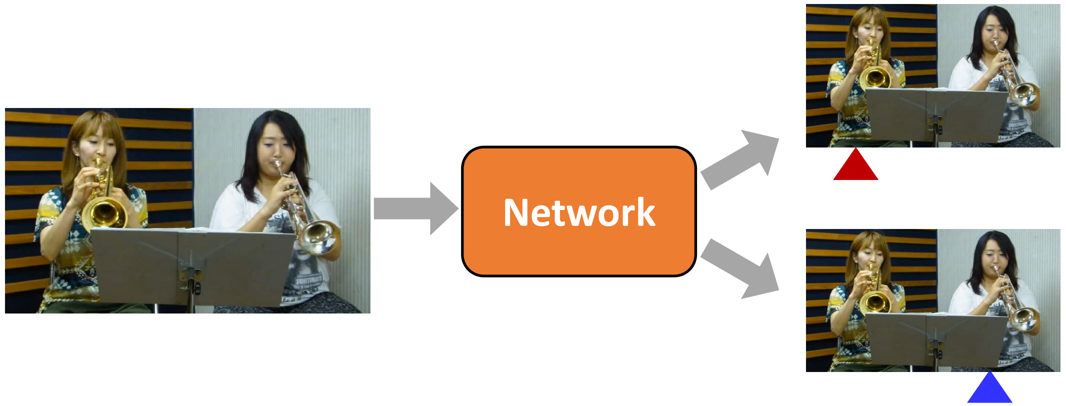
- Problem: limited to **separate multiple instruments of the same types**.



Challenges

- ➢ Most existing methods use **raw pixel** or **optical flow** as input.

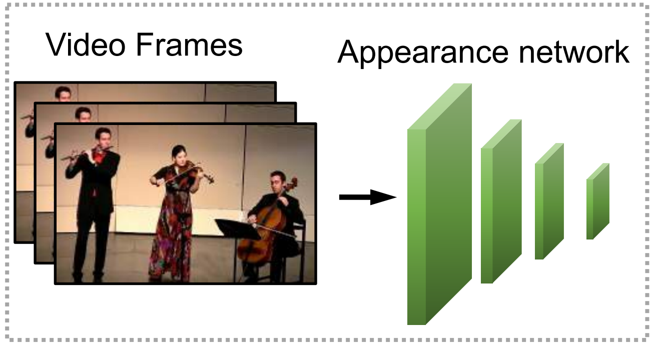
- ➢ Problem: limited to **separate multiple instruments of the same types**.



- ➢ We propose ``**Music Gesture**," a keypoint-based structured representation to explicitly model the body and finger movements of musicians.

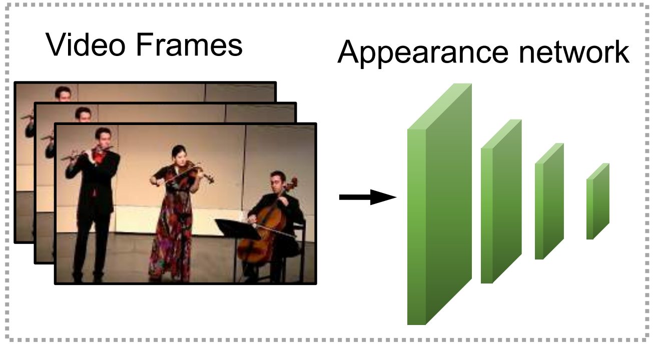
Our method

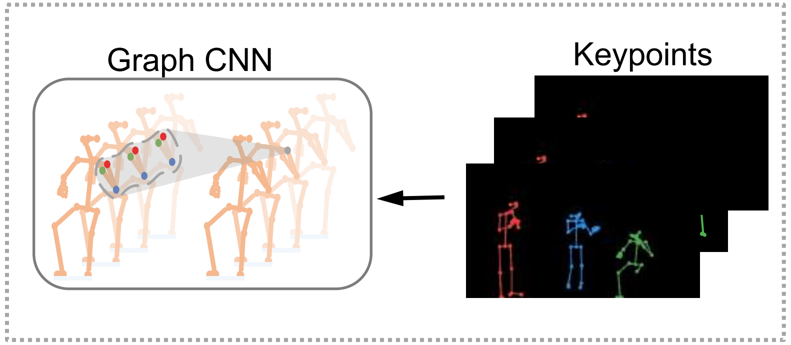
We first encode video frames by using ResNet-50.



Our method

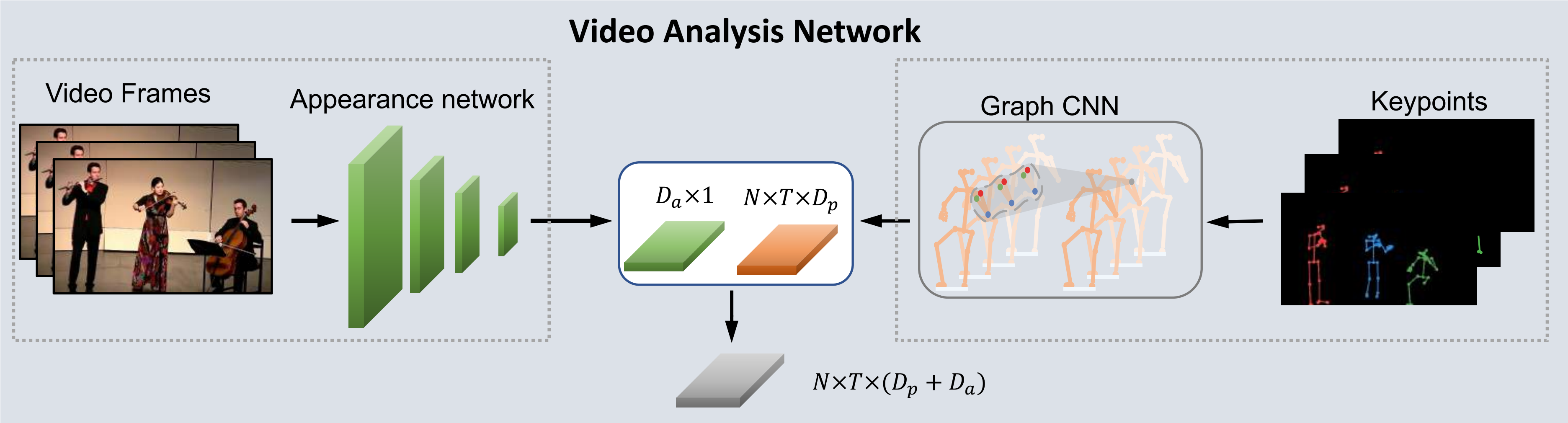
We also encode human keypoints by using graph CNN.





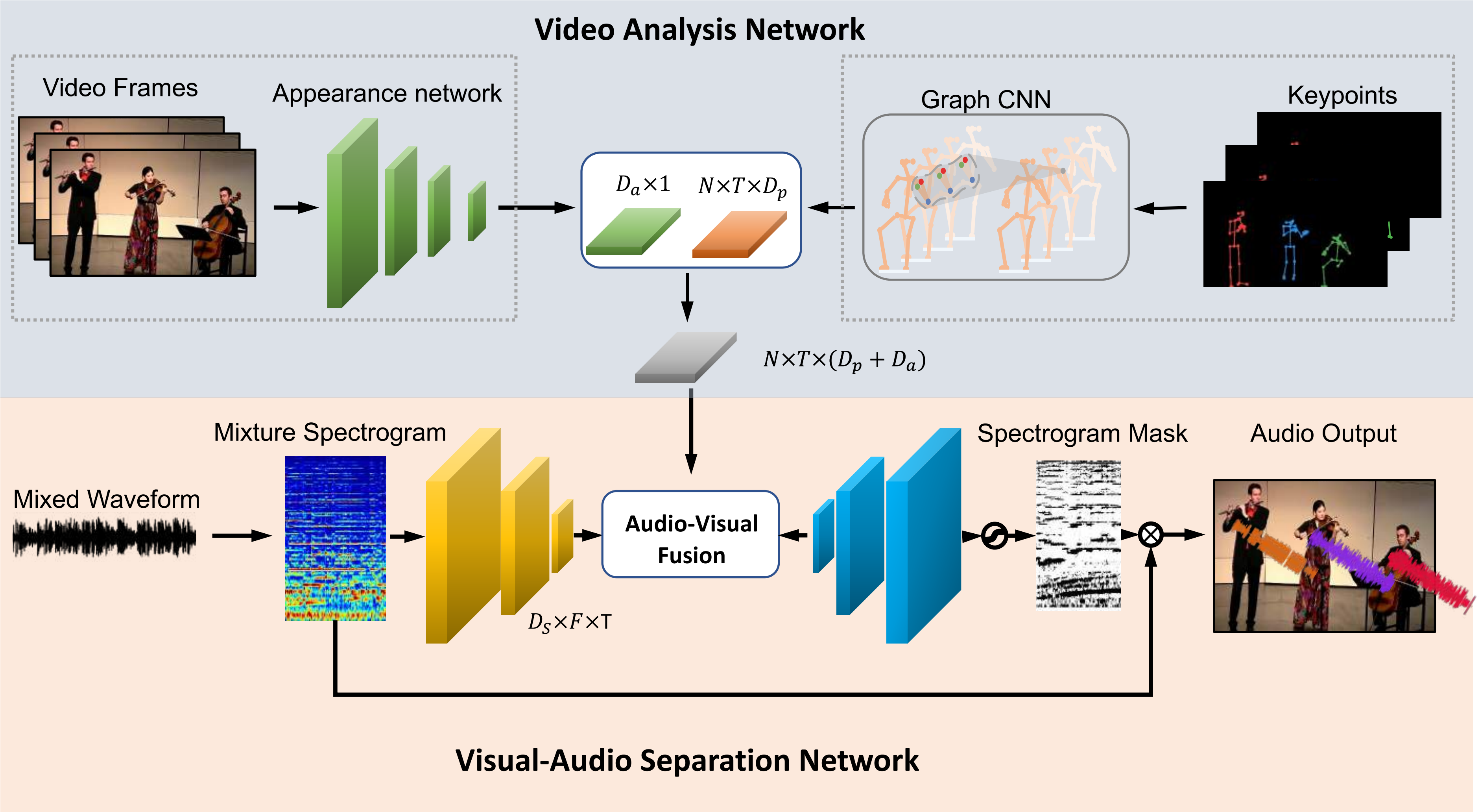
Our method

We combine the appearance and keypoints features as visual information.



Our method

We use the visual information to guide the networks to separate sounds.



Visual sound separation results

Previous method

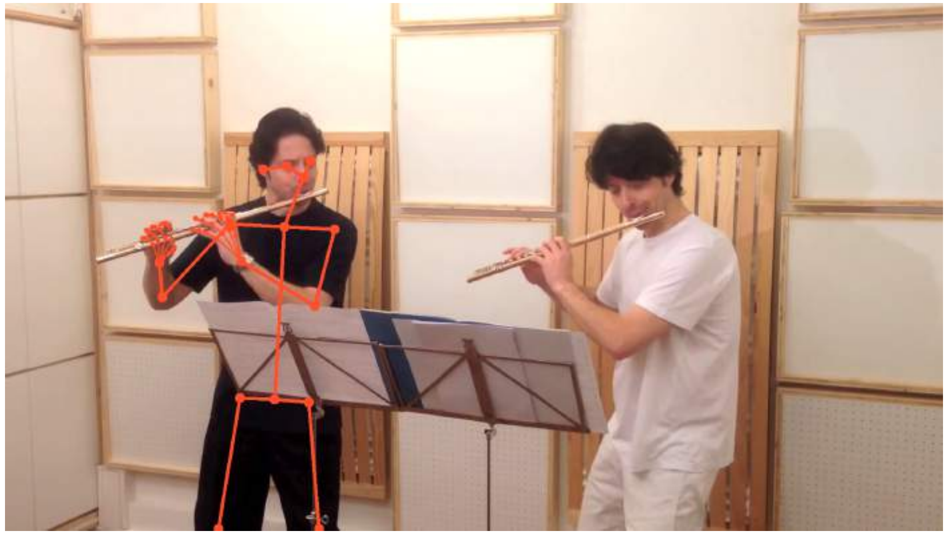
Mixed sound

Separated sound1

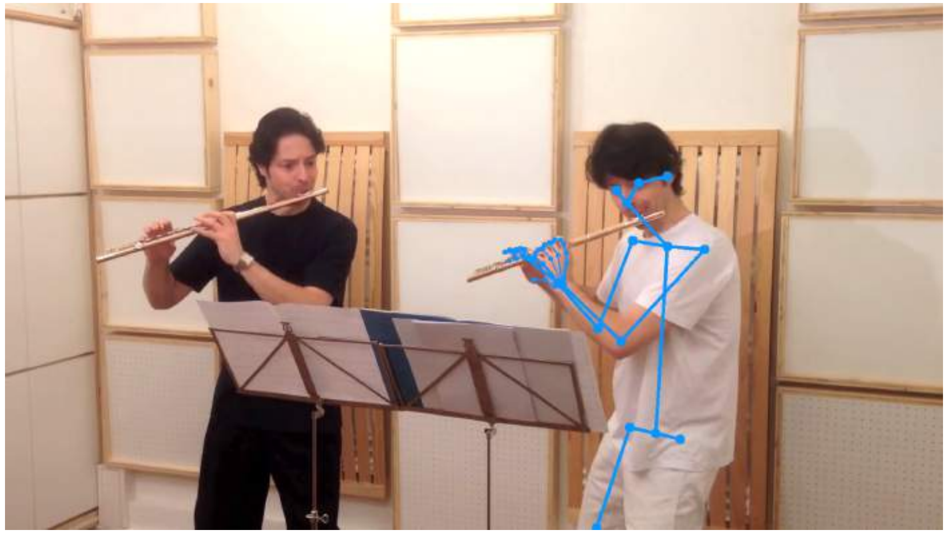
Separated sound2

Our method

Mixed sound



Separated sound1



Separated sound2

Previous method

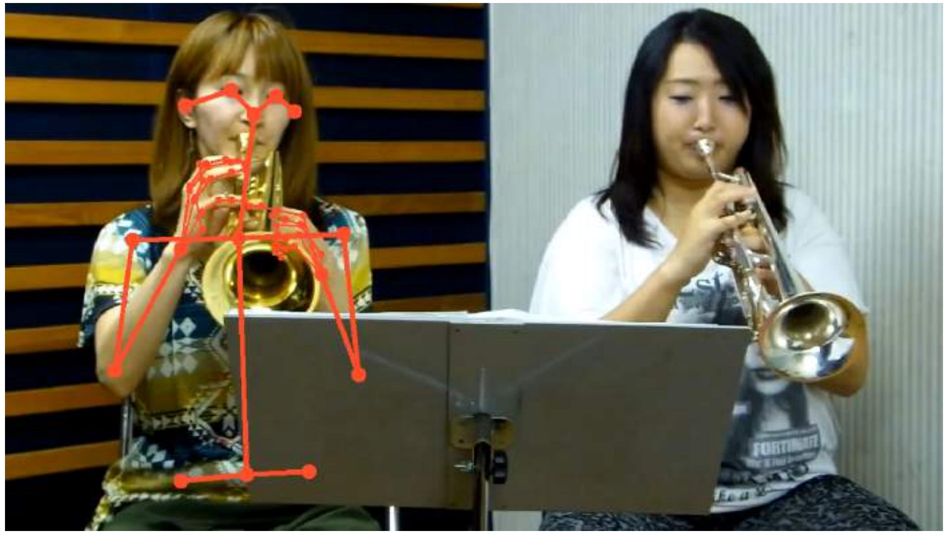
Mixed sound

Separated sound1

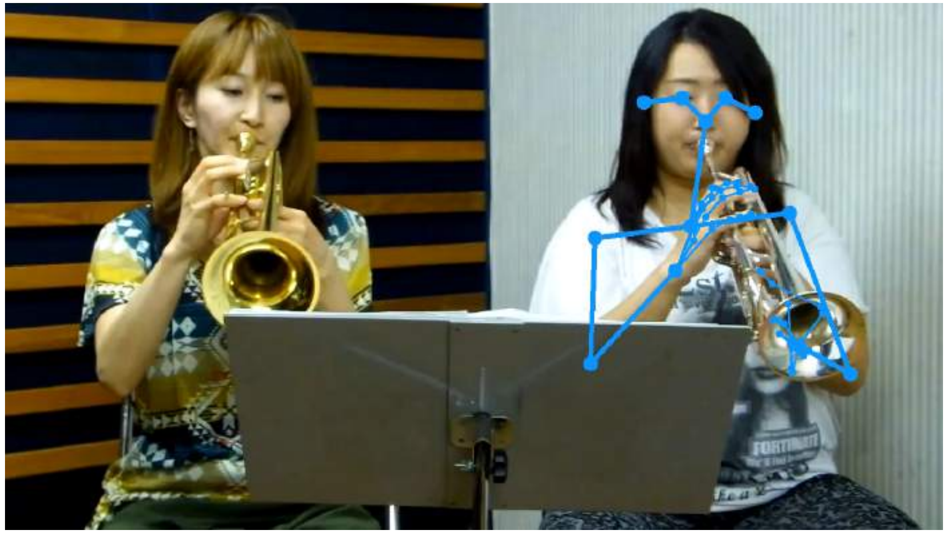
Separated sound2

Our method

Mixed sound



Separated sound1



Separated sound2

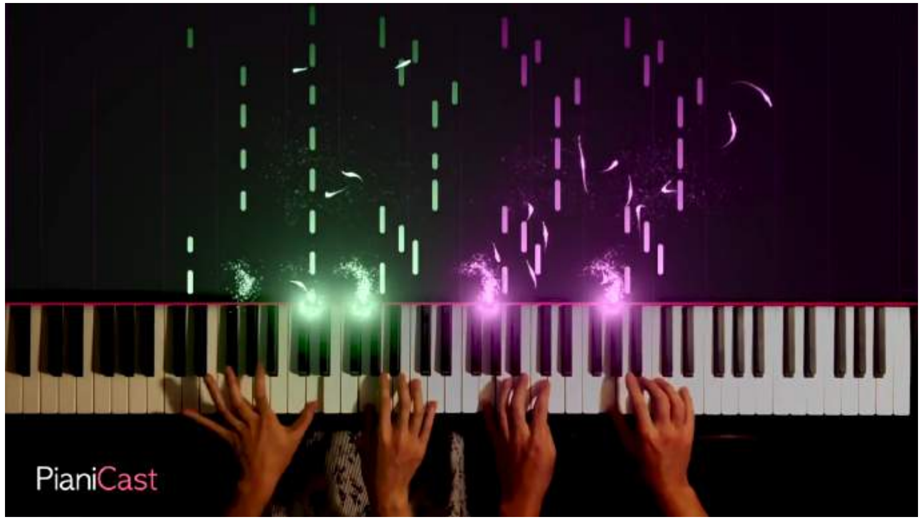
Previous method

Mixed sound

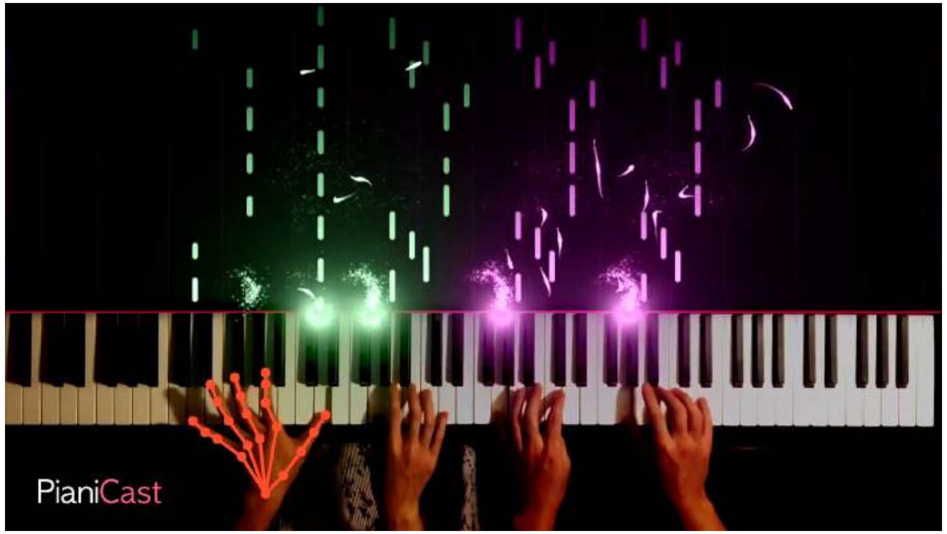
Separated sound1

Separated sound2

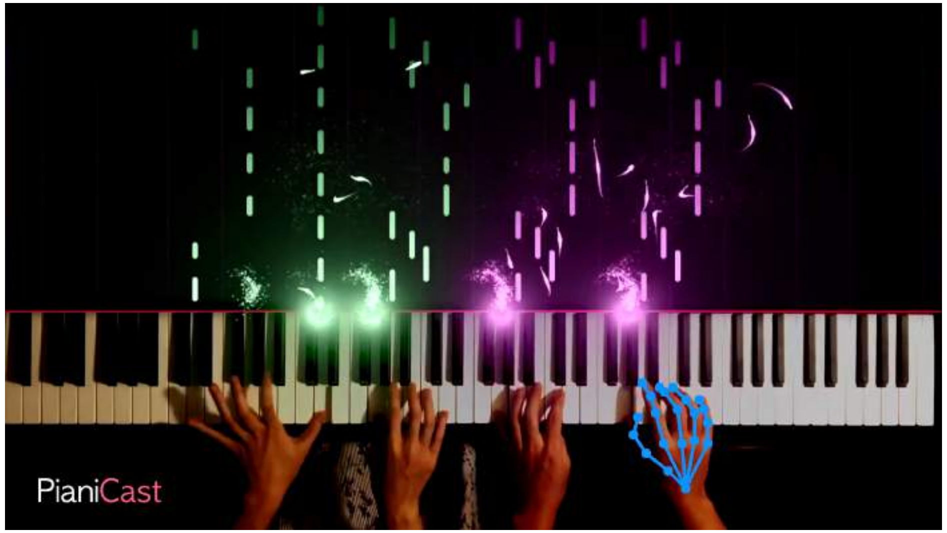
Our method



Mixed sound



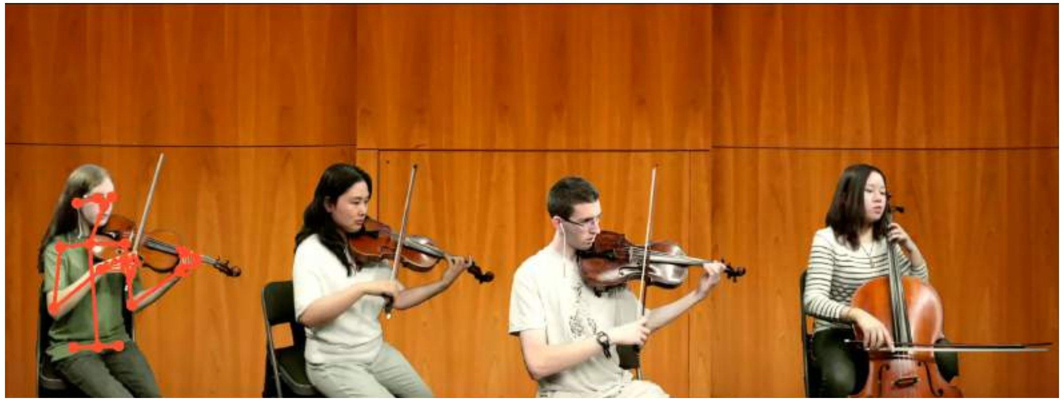
Separated sound1



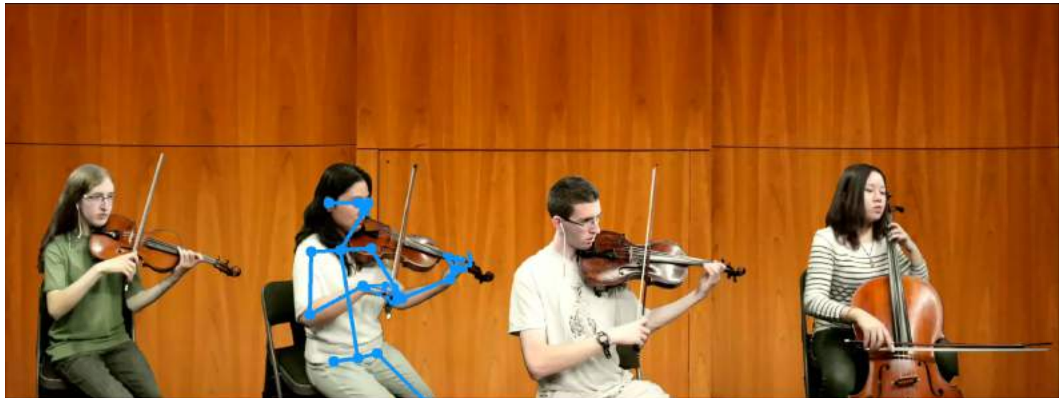
Separated sound2

Multiple instruments

Mixed sound

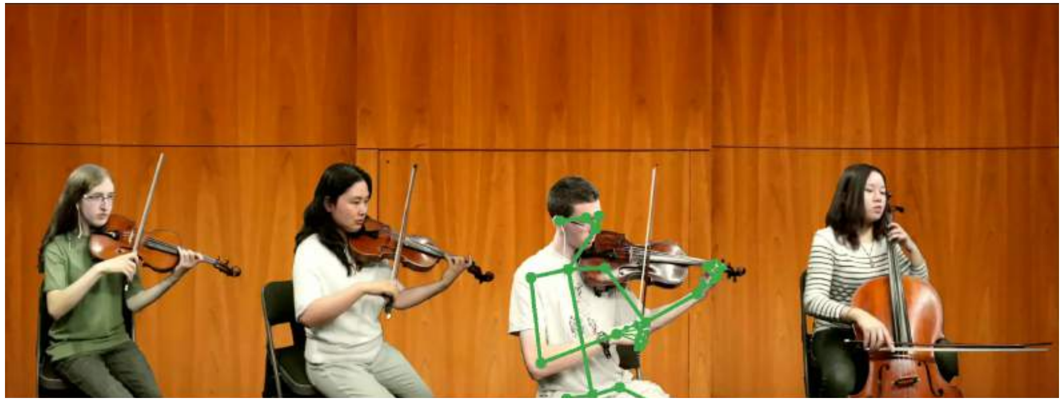


Separated sound1

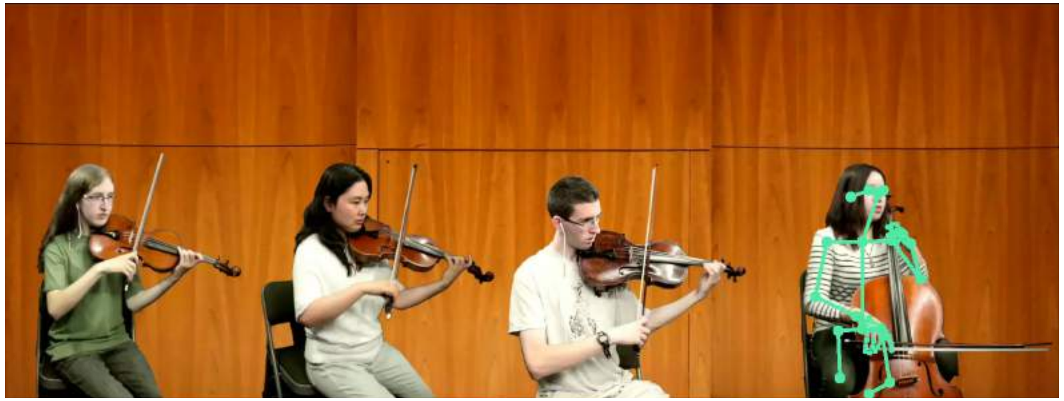


Separated sound2

Mixed sound

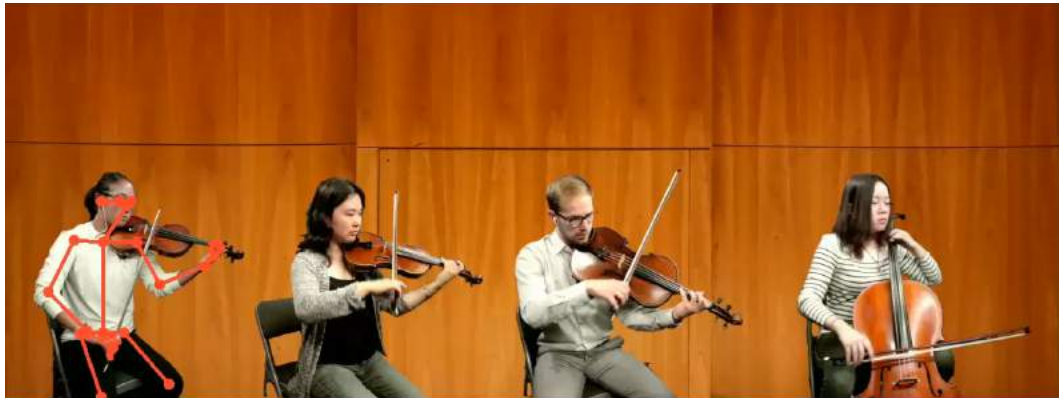


Separated sound3

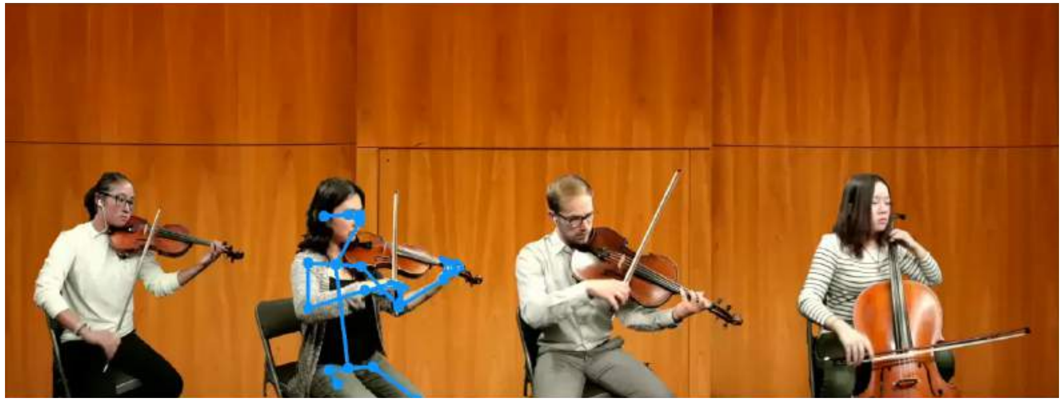


Separated sound4

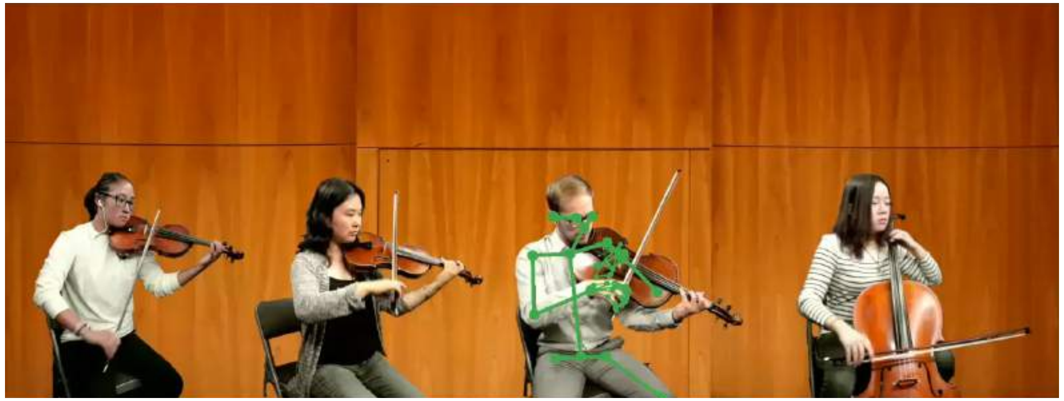
Mixed sound



Separated sound1

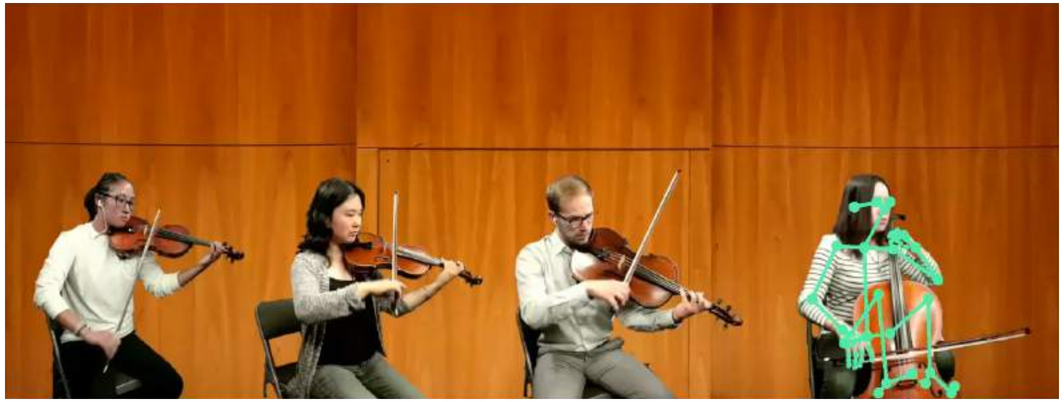


Separated sound2



Separated sound3

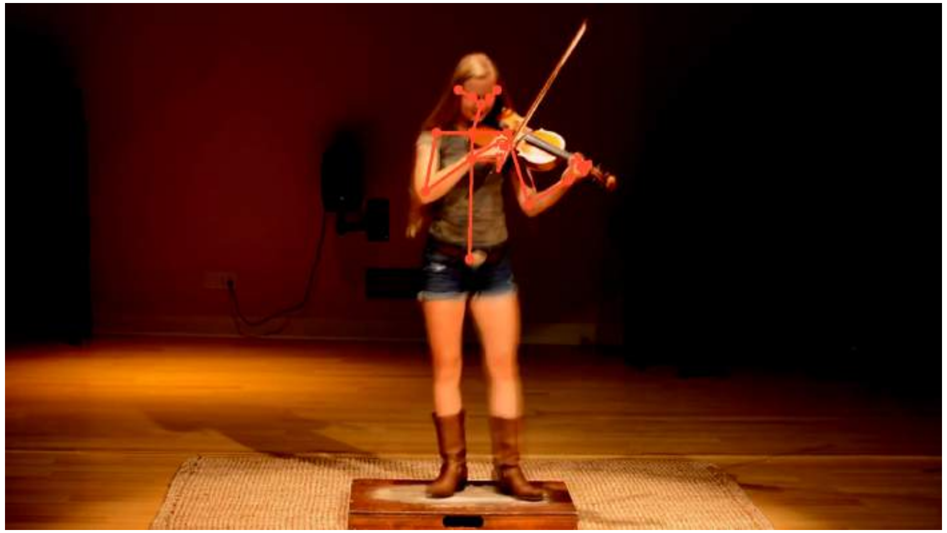
Mixed sound



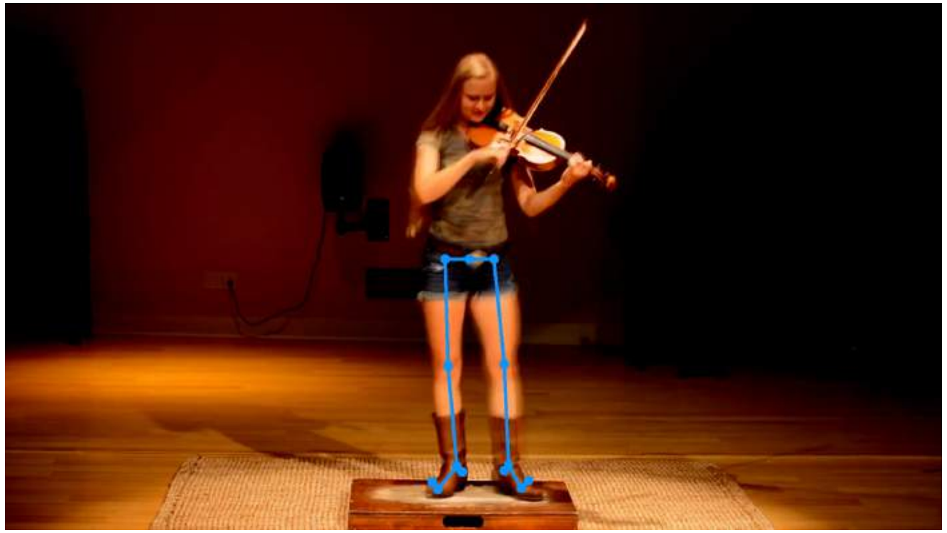
Separated sound4

The sound of body parts

Mixed sound



Separated sound1



Separated sound2

Thank you for your attention!