Seminar on Current Works in Computer Vision

CLEVRER: Collision Events for Video Representation and Reasoning

Hakan Sarp Aydemir

Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

Motivation

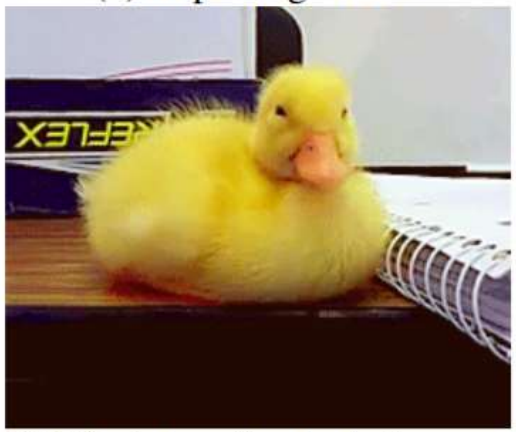
- Human intelligence: Ability to reason about temporal and causal events
- Most video reasoning benchmarks focus on pattern recognition

Video QA

(a) Repetition Count

Q) How many times does the man wrap string? **A)** 5 times

(b) Repeating Action



Q) What does the duck do 3 times? **A)** Shake head

(c) State Transition

Q) What does the bear on right do after sitting? **A)** Stand

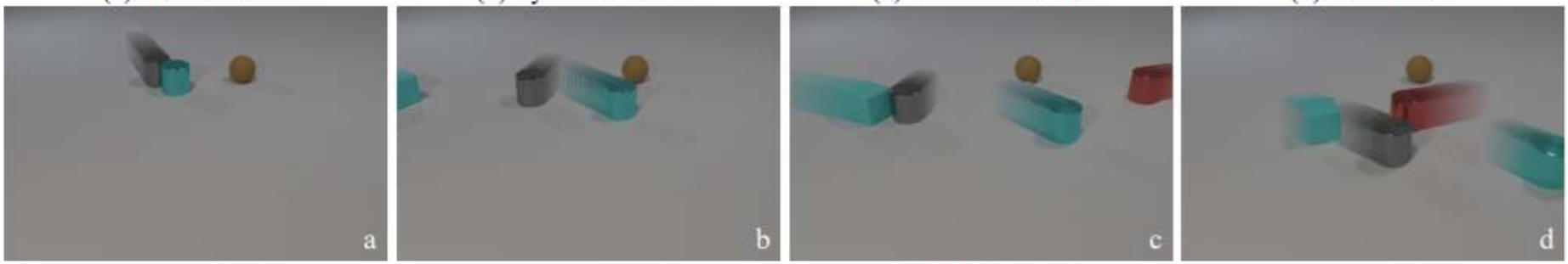
Motivation

- Focus on causal structure instead of pattern recognition
- Explore temporal and causal structures using videos of objects with simple appearance
- CLEVRER is introduced

CLEVRER

- Diagnostic video dataset for temporal and causal reasoning

(a) First collision (b) Cyan cube enters (c) Second collision (d) Video ends



I. Descriptive

Q: *What shape is the object that collides with the cyan cylinder?* **A:** *cylinder*

Q: *How many metal objects are moving when the video ends?* **A:** *3*

II. Explanatory

Q: *Which of the following is responsible for the gray cylinder's colliding with the cube?*

a) The presence of the sphere

b) The collision between the gray cylinder and the cyan cylinder **A:** *b)*

III. Predictive

Q: *Which event will happen next*

a) The cube collides with the red object

b) The cyan cylinder collides with the red object **A:** *a)*

IV. Counterfactual

Q: *Without the gray object, which event will not happen?*

a) The cyan cylinder collides with the sphere

b) The red object and the sphere collide **A:** *a), b)*

Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

Datasets

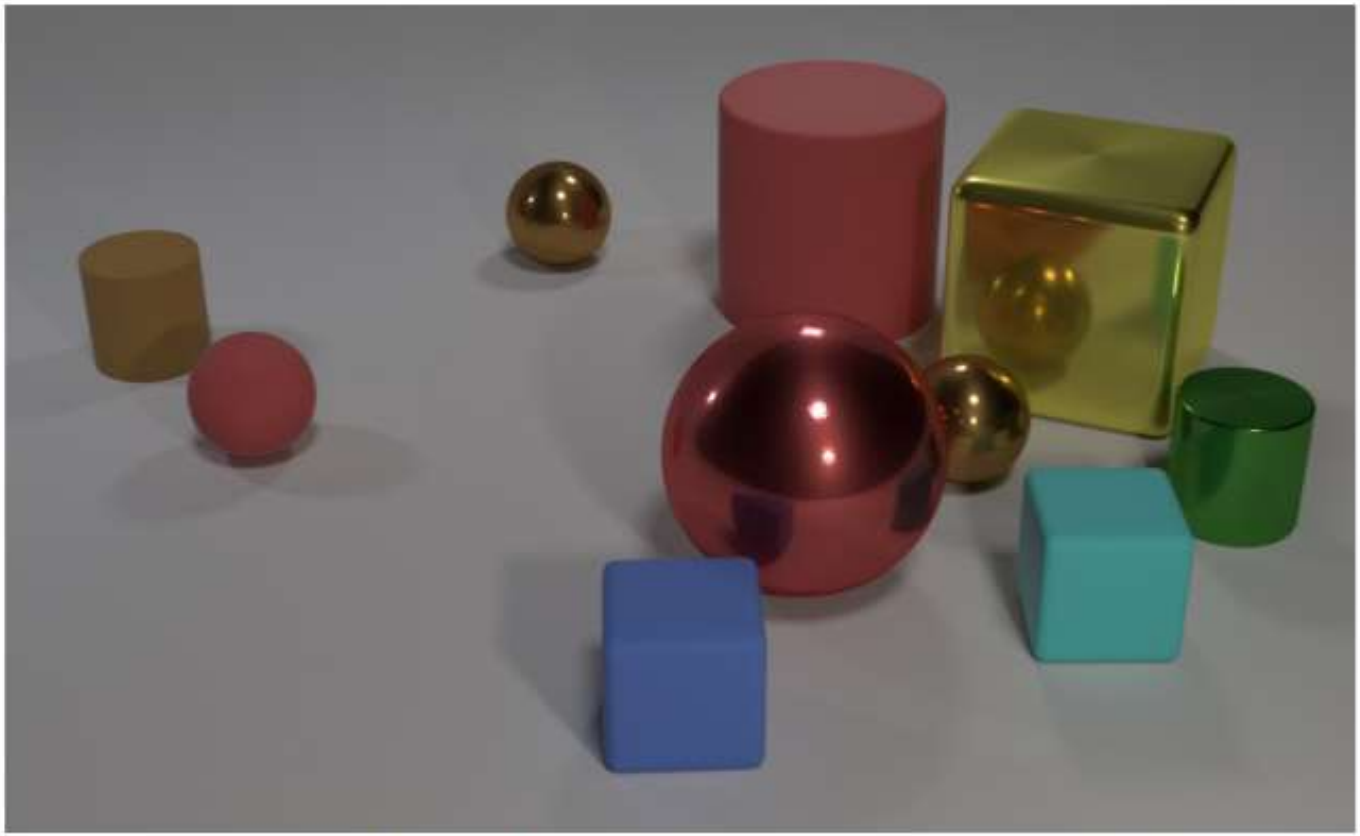
- CLEVRER includes all of the components that other datasets offer

Dataset	Video	Diagnostic Annotations	Temporal Relation	Explanation	Prediction	Counterfactual
VQA	×	×	×	×	×	×
CLEVR	×	✓	×	×	×	×
COG	×	✓	✓	×	×	×
VCR	×	✓	×	✓	×	✓
GQA	×	✓	×	×	×	×
TGIF-QA	✓	×	×	×	×	×
MovieQA	✓	×	✓	✓	×	×
MarioQA	✓	×	✓	✓	×	×
TVQA	✓	×	×	✓	×	×
Social-IQ	✓	×	×	✓	×	×
CLEVRER (ours)	✓	✓	✓	✓	✓	✓

CLEVR

- Diagnostic dataset
  - Visual reasoning abilities
- Requires complex reasoning to solve
  - Synthetic images
  - Automatically generated questions

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?

CLEVR vs CLEVRER

- CLEVR operates on images, CLEVRER on videos
- Objects have similar compositional intrinsic attributes
- Same software and parameters used for rendering
  - Image generation
  - Video generation
- Different focus of questions
  - CLEVR: Static properties of objects
  - CLEVRER: Temporal and causal aspects of objects and events in videos

Research Directions

- Can be positioned in the context of 3 research directions
  - Video understanding
  - Visual question answering
  - Physical and causal reasoning

Video Understanding

- Large scale video datasets => joint video and language understanding
  - Video captioning
  - Localizing video segments from natural language queries
  - Video question answering
- MovieQA, TGIF-QA, TVQA
  - Real-world videos and human-generated questions
- Social-IQ
  - Causal relations in human socia interactions
- COG, MarioQA
  - Simulated environments used to generate data and controllable reasoning tasks

Video Understanding

- CLEVRER
  - Causal relations in object dynamics and physical interactions
  - Introduces tasks: Description, explanation, prediction, counterfactuals
  - Emphasize compositionality in visual and logic context

Visual Question Answering

- VQA
  - Based on cloud-sourced real images and human-generated questions
- CLEVR

Visual Question Answering

- VCR
  - Explanations and hypothesis judgements based on common sense
- GQA
  - Synthetic compositional questions to real images

Physical and Causal Reasoning

- Related to research on learning scene dynamics for physical and causal reasoning.
  - Directly from images
  - OR from symbolic representation of the environment

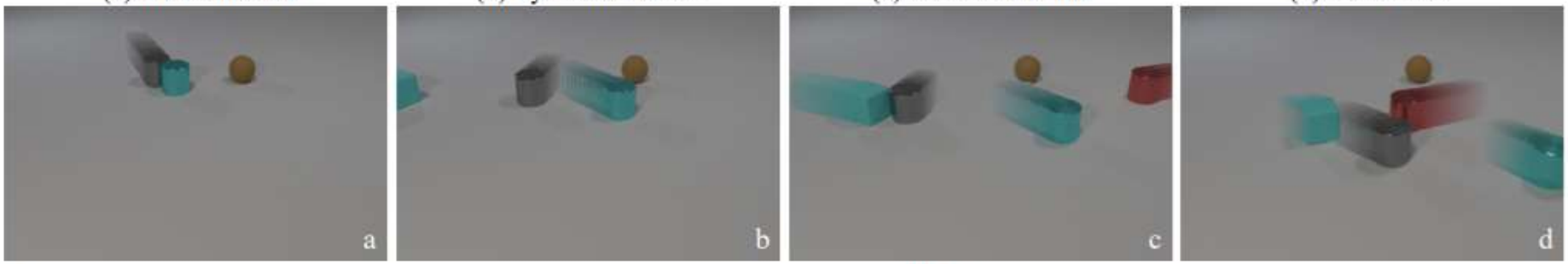
- In CLEVRER: How a learned dynamics model contributes to causal reasoning

Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

The CLEVRER Dataset

(a) First collision (b) Cyan cube enters (c) Second collision (d) Video ends



I. Descriptive

Q: *What shape is the object that collides with the cyan cylinder?* **A:** *cylinder*

Q: *How many metal objects are moving when the video ends?* **A:** *3*

II. Explanatory

Q: *Which of the following is responsible for the gray cylinder's colliding with the cube?*

a) The presence of the sphere

b) The collision between the gray cylinder and the cyan cylinder **A:** *b)*

III. Predictive

Q: *Which event will happen next*

a) The cube collides with the red object

b) The cyan cylinder collides with the red object **A:** *a)*

IV. Counterfactual

Q: *Without the gray object, which event will not happen?*

a) The cyan cylinder collides with the sphere

b) The red object and the sphere collide **A:** *a), b)*

Structure of CLEVRER

- Has 2 parts
  - Videos of object motion and collisions, uses a physics engine
  - Questions that are generated about each video
- 20,000 videos and 300,000 questions

Videos

- 10,000 training, 5,000 validation and 5,000 testing videos
  - 5 second long

- Generated
  - By object motion simulating physics engine
  - By graphs engine that renders the frames
  - From simuated motion traces

- Intrinsic attributes sampled randomy for each object
  - Identical objects are prohibited => unique object with combination of the 3 attributes (Shape, Material, Color)

Videos - Causal Structure

- Randomly initialized objects + new objects with overlapping motion traces => compex causal structure

- Example: A sphere collides with a cube then a cylinder.
  - First collision (Sphere + cube) and the cube cause the second collision

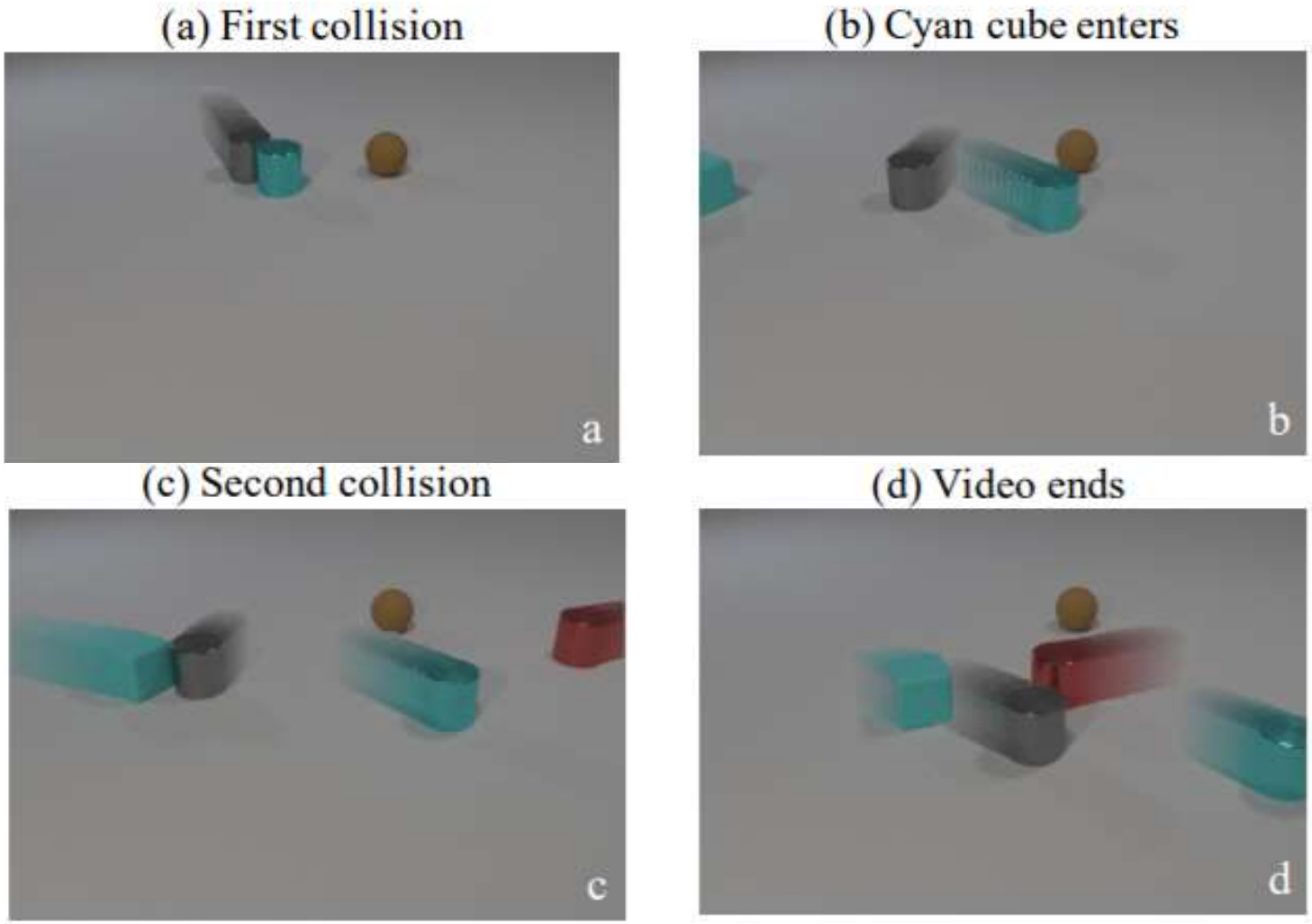
Questions

- Each video paired with machine-generated questions
  - Descriptive
  - Explanatory
  - Predictive
  - Counterfactual

- Each question paired with functional program executable

- No questions about static object properties

Questions - Descriptive

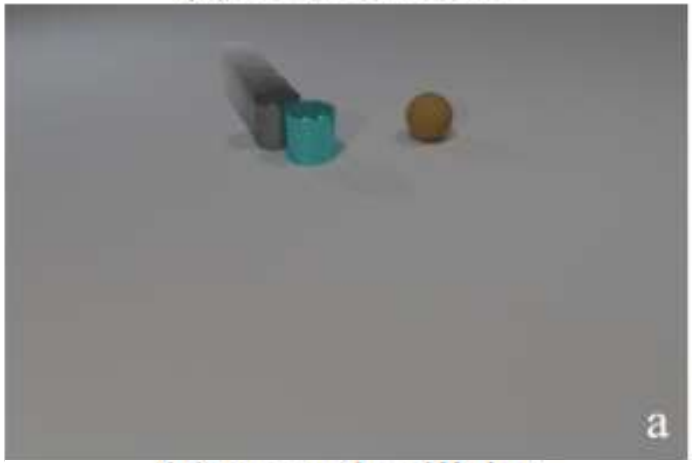


Q: *What shape is the object that collides with the cyan cylinder?* **A:** *cylinder*

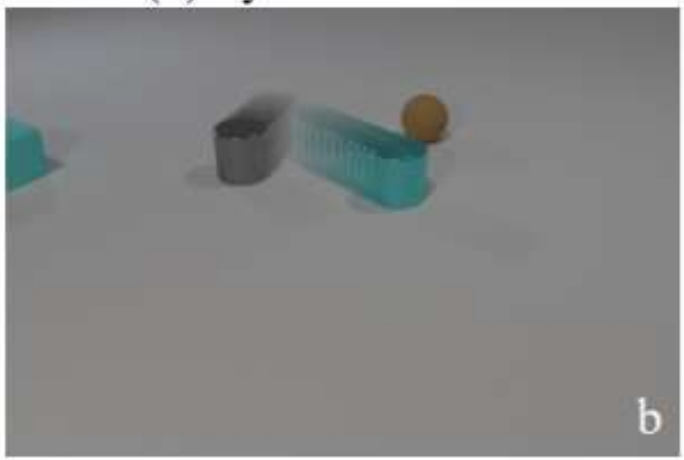
Q: *How many metal objects are moving when the video ends?* **A:** *3*

Questions - Explanatory

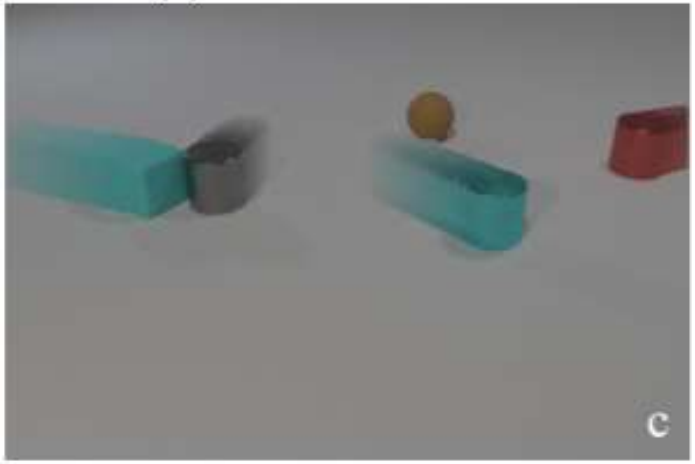
(a) First collision



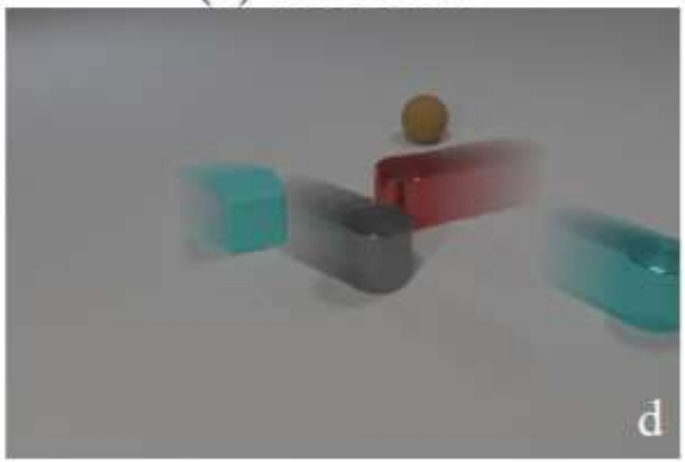
(b) Cyan cube enters



(c) Second collision



(d) Video ends



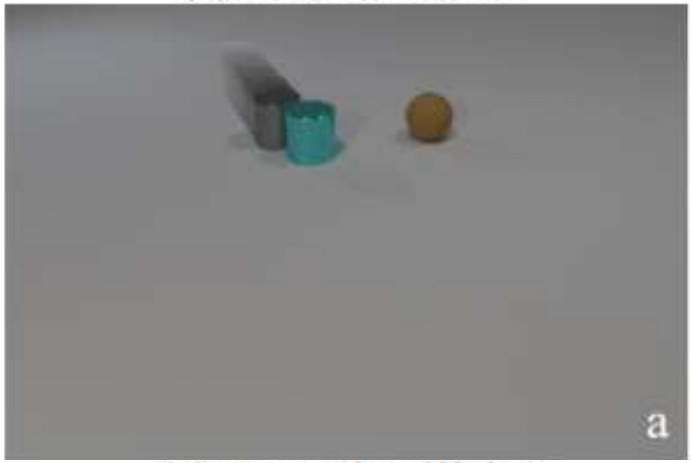
Q: *Which of the following is responsible for the gray cylinder's colliding with the cube?*

a) The presence of the sphere

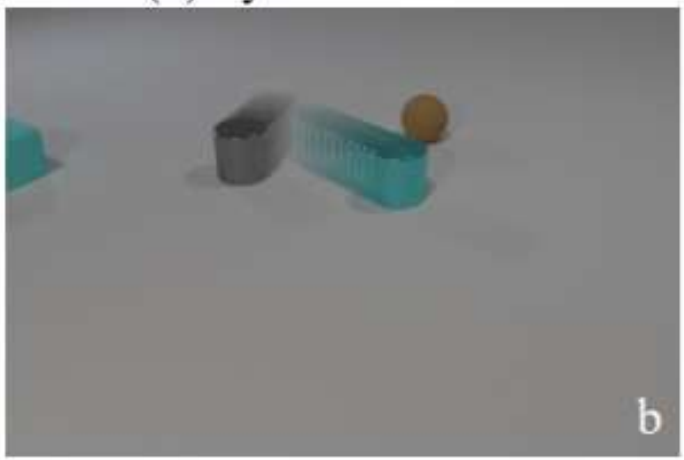
b) The collision between the gray cylinder and the cyan cylinder **A:** *b)*

Questions - Predictive

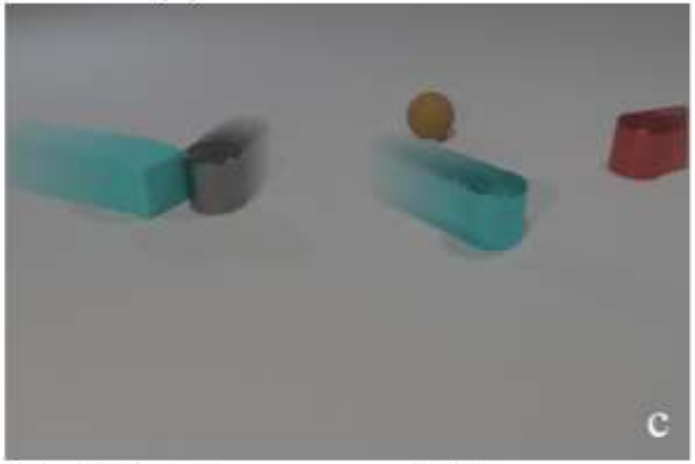
(a) First collision



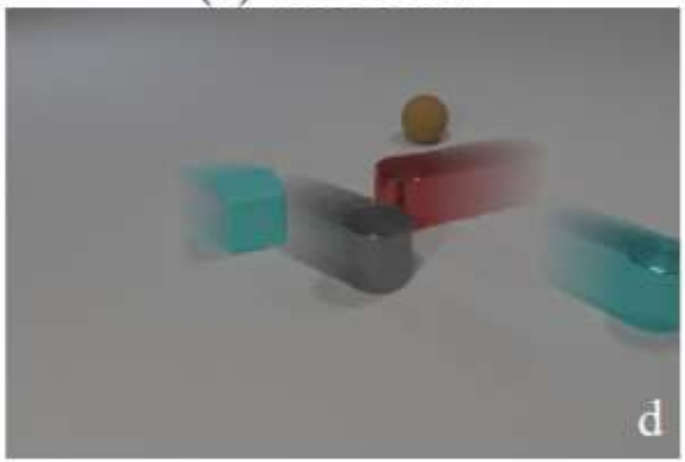
(b) Cyan cube enters



(c) Second collision



(d) Video ends



Q: *Which event will happen next*

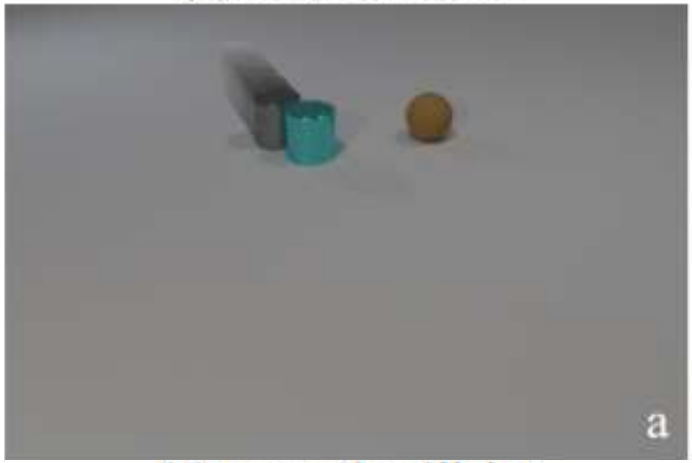
a) The cube collides with the red object

b) The cyan cylinder collides with the red object

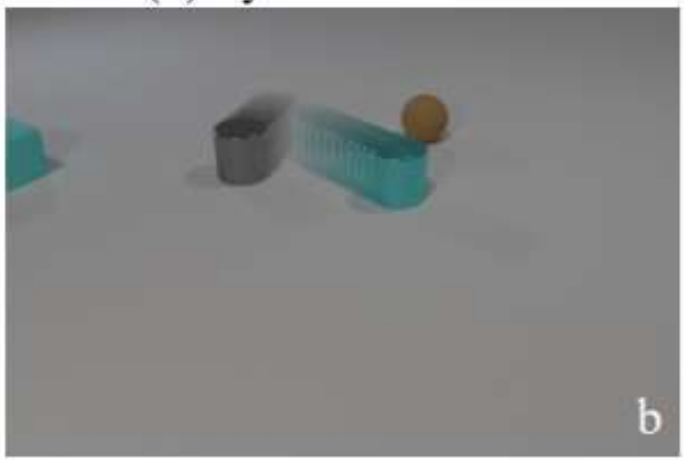
A: *a)*

Questions - Counterfactual

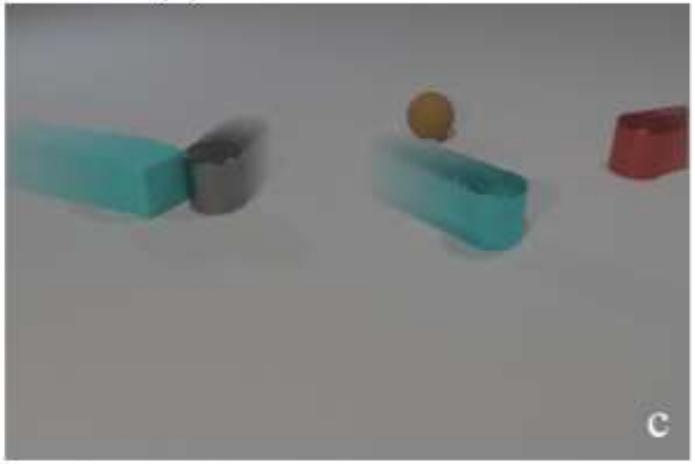
(a) First collision



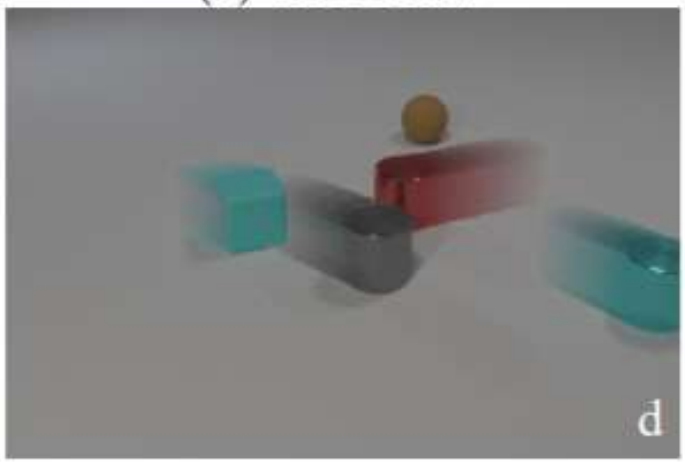
(b) Cyan cube enters



(c) Second collision



(d) Video ends



Q: *Without the gray object, which event will not happen?*

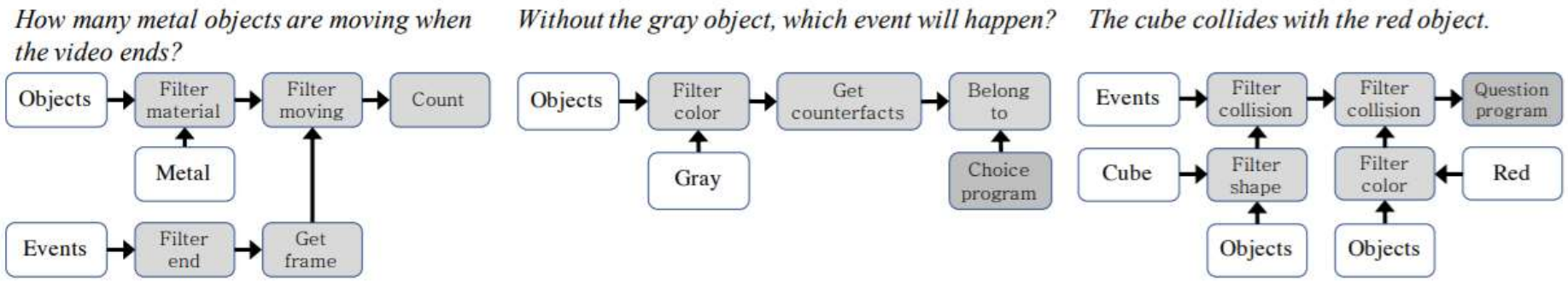
a) The cyan cylinder collides with the sphere

b) The red object and the sphere collide

A: *a), b)*

Program representation

- Tree-structured functional program for each question



Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

Baseline Evaluation on CLEVRER

- Evaluate performances of different baseline models

- 3 Families of baseline models
  - Language-only models
  - Video question answering
  - Compositional visual reasoning

Model Details - Language-only Models

- Weak baselines that depend on question input
  - Assess language biases in CLEVRER

- Q-type (random)
  - Uniformly sample answer from sample space
  - OR randomly select choices for multiple-choice questions

- Q-type (frequent)
  - Choose most frequent answer in training set for each question type

Model Details - Language-only Models

- LSTM
  - Word embedding pretrained on Google News corpus
  - TO encode input question
  - AND process the sequence with LSTM

Model Details - Video Question Answering

- Models that rely on both video and language inputs

- CNN+MLP extracts features via CNN
  - Questions encoded by taking the avg. of pretrained word embeddings
  - Features sent to an MLP for answer prediction

- CNN+LSTM also extracts features via CNN
  - Uses final state of an LSTM for answer prediction

Model Details - Video Question Answering

- TVQA
  - Multi-stream end-to-end neural model
  - State of the art for video question answering
- TVQA+
  - Attribute-aware object-centric features applied to TVQA
- Memory
  - Incorporate heterogeneous memory with multimodal attention

Model Details - Compositional visual reasoning

- CLEVR
  - Emphasizes complexity and compositionality in the logic and visual context
- IEP (V)
  - IEP applies neural program execution for visual reasoning on images
  - Same approach => program based video reasoning task
- Tbd-net (V)
  - Spatial-temporal attention over the video feature space

Model Details - Compositional visual reasoning

- MAC (V)
  - MAC modified to apply temporal attention unit accross video frames
  - Generate latent encoding for video

- MAC (V+)
  - MAC (V) augmented
  - Add segmentation masks of objects in frames + label them by their intrinsic attributes => object aware video features

Results - Evaluation metric

- Descriptive questions
  - Accuracy => compare predicted answer token with ground truth

- Multiple choice questions
  - Per-option accuracy => Overall correctness on single options (Across all questions)
  - Per-question accuracy => Correctness of the question as a whole (All choices selected correctly)

Results

Methods	Descriptive	Explanatory		Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.	per opt.	per ques.
Q-type (random)	29.2	50.1	8.1	50.7	25.5	50.1	10.3
Q-type (frequent)	33.0	50.2	16.5	50.0	0.0	50.2	1.0
LSTM	34.7	59.7	13.6	50.6	23.2	53.8	3.1
CNN+MLP	48.4	54.9	18.3	50.5	13.2	55.2	9.0
CNN+LSTM	51.8	62.0	17.5	57.9	31.6	61.2	14.7
TVQA+	72.0	63.3	23.7	70.3	48.9	53.9	4.1
Memory	54.7	53.7	13.9	50.0	33.1	54.2	7.0
IEP (V)	52.8	52.6	14.5	50.0	9.7	53.4	3.8
TbD-net (V)	79.5	61.6	3.8	50.3	6.5	56.1	4.4
MAC (V)	85.6	59.5	12.5	51.0	16.5	54.6	13.7
MAC (V+)	86.4	70.5	22.3	59.7	42.9	63.5	25.1

Results - Descriptive reasoning

- To do well, the model needs to
  - Accurately recognize objects and events in the video
  - Understand the compositional logic behind questions

- Stong perception and logic operations on both visual and language input needed

Results - Causal Reasoning

- Results on descriptive questions
  - Power of models that combine visuals and language

- Causal reasoning tasks (explanatory, predictive, counterfactual) need more than perception

Results - Findings

- Object-centric representations essential for causal reasoning tasks
  - Improvement of performance of MAC (V+) over MAC (V)
  - Performance of TVQA+ on predictive questions
- Baseline models don’t have needed components
  - Explicitly model dynamics of objects
  - Explicitly model causal relations between events
  - Struggle on tasks with unobserved scenes and counterfactual tasks
- Promising direction for causal tasks is combining
  - Object-centric representations
  - Dynamics modeling

Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

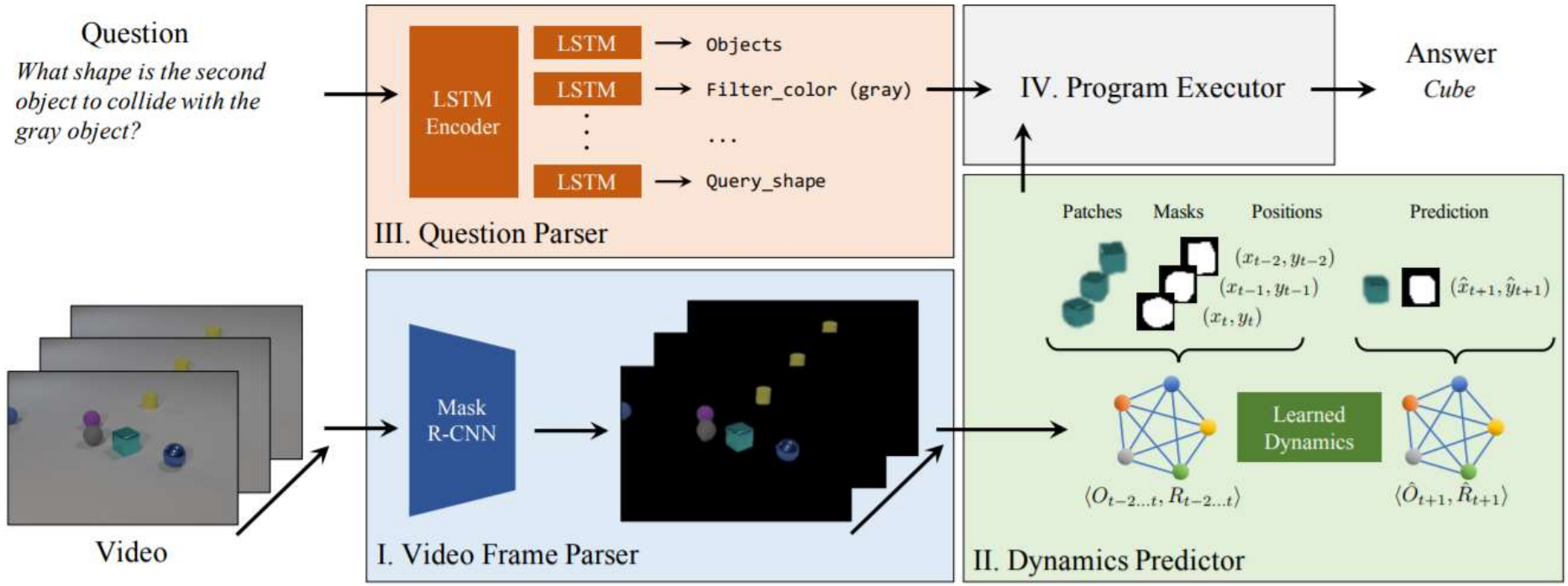
NS-DR

- Unify
  - Object-centric representations
  - Dynamics modeling

- Challenges
  - Disjoint model components should operate on common set of representations of Video, question, dynamics, causal relations
  - Representation aware of compositional relations between objects and events

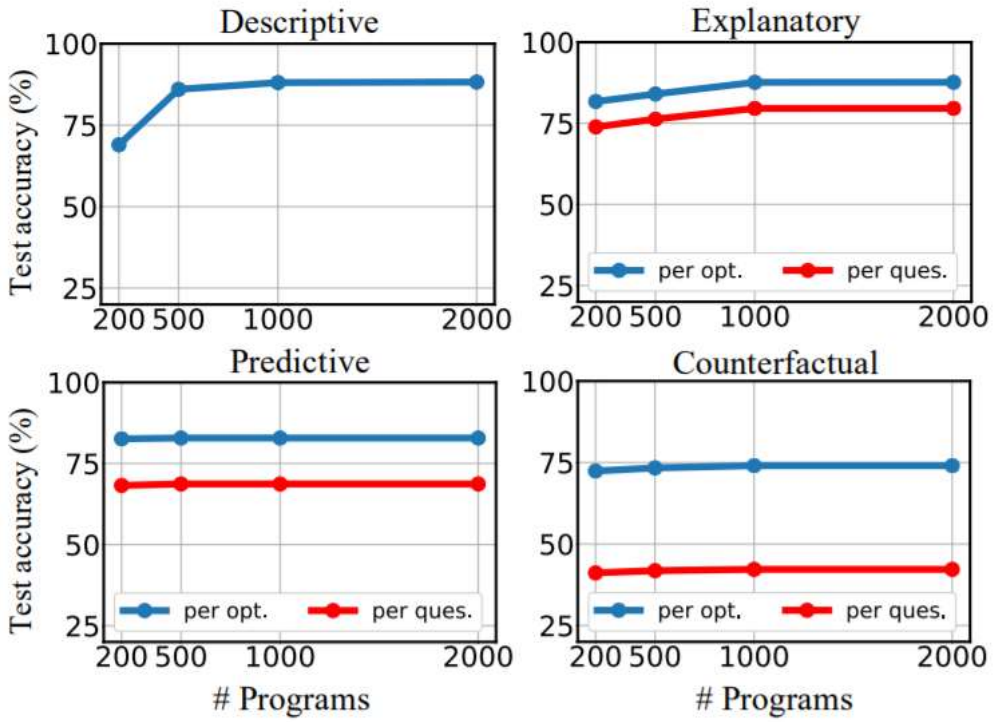
- Framework that joins these components via symbolic representation: NS-DR

Model


Question
What shape is the second object to collide with the gray object?
LSTM Encoder
LSTM
LSTM
LSTM
Objects
Filter_color (gray)
...
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

Results

Methods	Descriptive	Explanatory		Predictive		Counterfactual	
		per opt.	per ques.	per opt.	per ques.	per opt.	per ques.
NS-DR	88.1	87.6	79.6	82.9	68.7	74.1	42.2
NS-DR (NE)	85.8	85.9	74.3	75.4	54.1	76.1	42.0



Overview

- Introduction
- Related Work
- Structure of CLEVRER
- Baseline Evaluation on CLEVRER
- Neuro-Symbolic Dynamic Reasoning
- Conclusion

Conclusion

- Systematic study of temporal and causal reasoning in videos
  - Benchmark tasks introduced to better facilitate the research
  - CLEVRER dataset and NS-DR model

- Capture true intelligence beyond pattern recognition

Thank You!

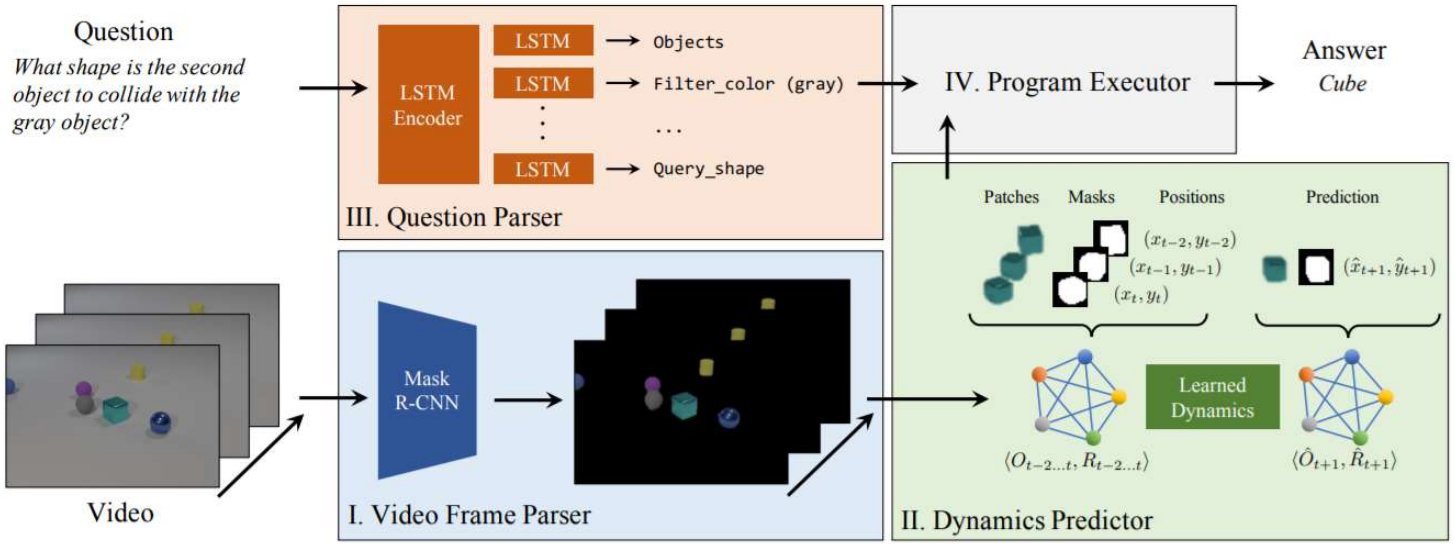
Backup Slides

NS-DR

- Different baseline models evaluated

- NS-DR combines components that are required for causal reasoning tasks

- Symbolic representations



Videos - Objects and Events --Objects

- Objects have 3 intrinsic attributes
  - Shapes: Cube, Sphere, Cylinder
  - Materials: Metal, Rubber
  - Colors: Gray, Red, Blue, Green, Brown, Cyan, Purple, Yellow

Videos - Objects and Events --Events

- 3 Types of events
  - Enter
  - Exit
  - Collision

- Fixed number of participants for each event
  - 2 For collision
  - 1 For enter and exit

Videos - Objects and Events

- Abstract representation of the video formed

- Ground truth annotations + object motion traces => model diagnostics

Videos - Video generation

- Generated from simulated motion traces
  - Object's position and time at each time step

- Each simulation lasts 7 seconds
  - Motion traces first downsampled to fit 25 frames per second
  - First 5 seconds sent to Blender
  - Remaining 2 seconds held-out for predictive tasks

Videos - Causal Structure

- Event can be caused by 2 ways
  - By an object if the event is the first one participated by the object
  - By another event if the cause event happened right before the outcome event on the same object
- Example: A sphere collides with a cube then a cylinder.
  - First collision (Sphere + cube) and the cube cause the second collision
- Randomly initialized objects + new objects with overlapping motion traces => compex causal structure

Questions - Descriptive

- Video's diynamical content and temporal relation
- Open-ended questions, can be answered by a single word
- Sub-types: Count, exist, color, etc.

Questions - Explanatory

- Whether object or event responsible for the event
  - Event A is ancestor of Event B in the causal graph => A responsible for B
  - Object B participates in Event A or any other event responsible for A => O responsible for A
- Multiple choice with at most 4 options
- Model selects all options that match
  - Correct and wrong options sampled equally for balance and bias minimization

Questions - Predictive

- Predict possibe occurences of future events
- Multiple choice
- Sparse post-video events
  - 2 options provided to reduce bias

Questions - Predictive

- Example: Which event will happen next?

a) The cube collides with the red object

b) The cyan cylinder collides with the red object

A: a)

Questions - Counterfactual

- Outcome of the video under hypothetical conditions.
- At most 4 options
- Correct and inorrect option numbers balanced

Questions - Counterfactual

- Example: Without the gray object, which event will not happen?

a) The cyan cylinder collides with the sphere

b) The red object and the sphere collide

A: a), b)

Baseline Evaluation on CLEVRER

- Evaluate performances of different baseline models
  - Multi-class classification for descriptive questions
  - Binary classification for multiple choice questions

Results

- Different performances => Powerful assesment to models' strength and limitations on various domains

- Models trained until convergence
  - Tuned on validation set
  - Evaluated on test set

Results - Descriptive reasoning

- LSTM relies only on question input => performs poorly
  - Suggests CLEVRER has very small bias on questions
- TVQA+ achieves better performance
  - Video QA model
  - Compositionality in question logic and visual context handling limited => doesn't thrive
- TbD-net and MAC achieve better performance
  - Designed for compositional reasoning
  - Neural program execution
  - Joint attention mechanism

Results - Causal Reasoning

- Poor performance of most baseline models
- Models that do well on descriptive questions (MAC (V), TbD-net (V)) perform poorly
- Performance gain on models with object-aware representations
  - TVQA+ achieves high accuracy on predictive questions
  - MAC (V+) better than MAC (V) for all tasks