Object-Centric Representation Learning for Video Question Answering

Long Hoang Dang, Thao Minh Le, Vuong Le, Truyen Tran

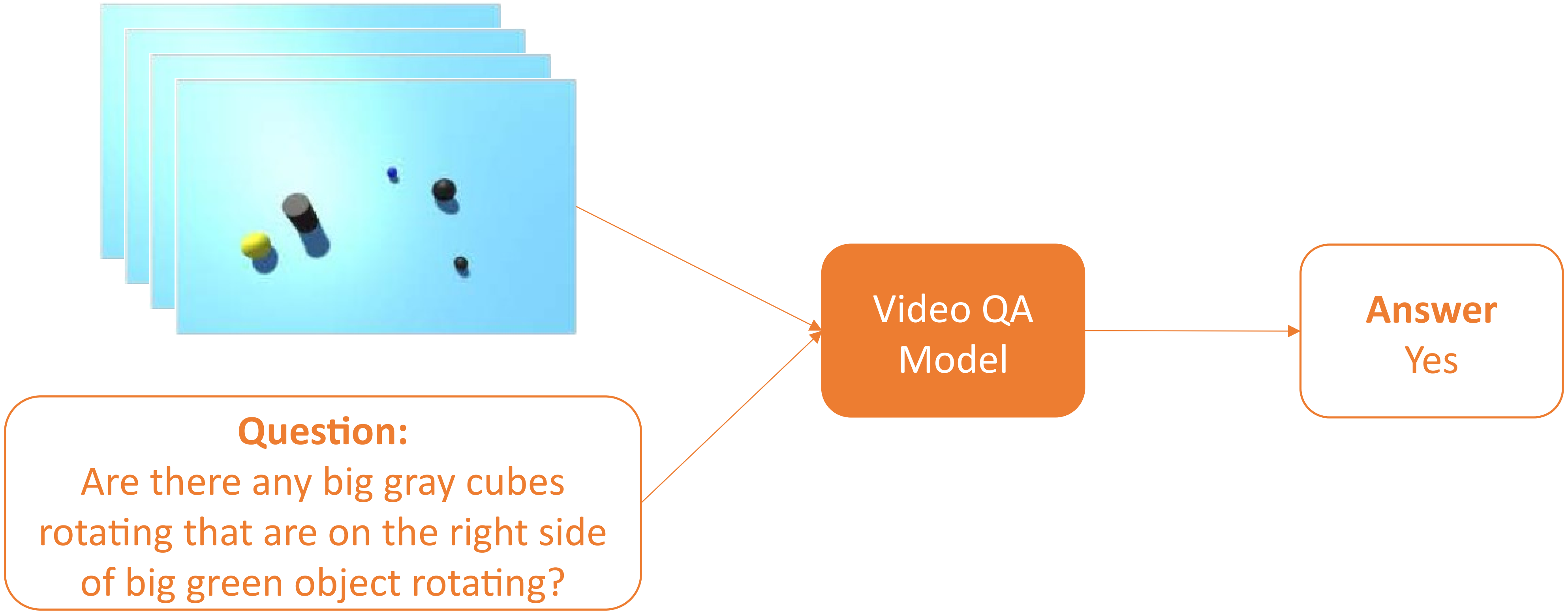
Presented at IJCNN 2021

APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Introduction

General Video QA Framework



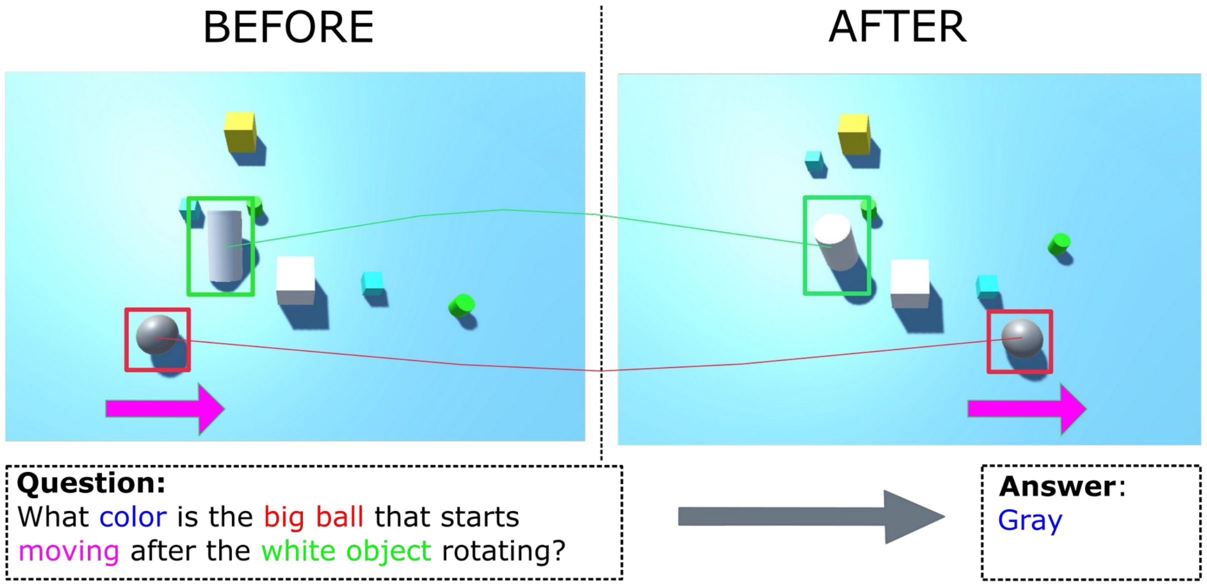
Question:
Are there any big gray cubes rotating that are on the right side of big green object rotating?

Video QA Model

Answer
Yes

Our focus: Object-centric representation

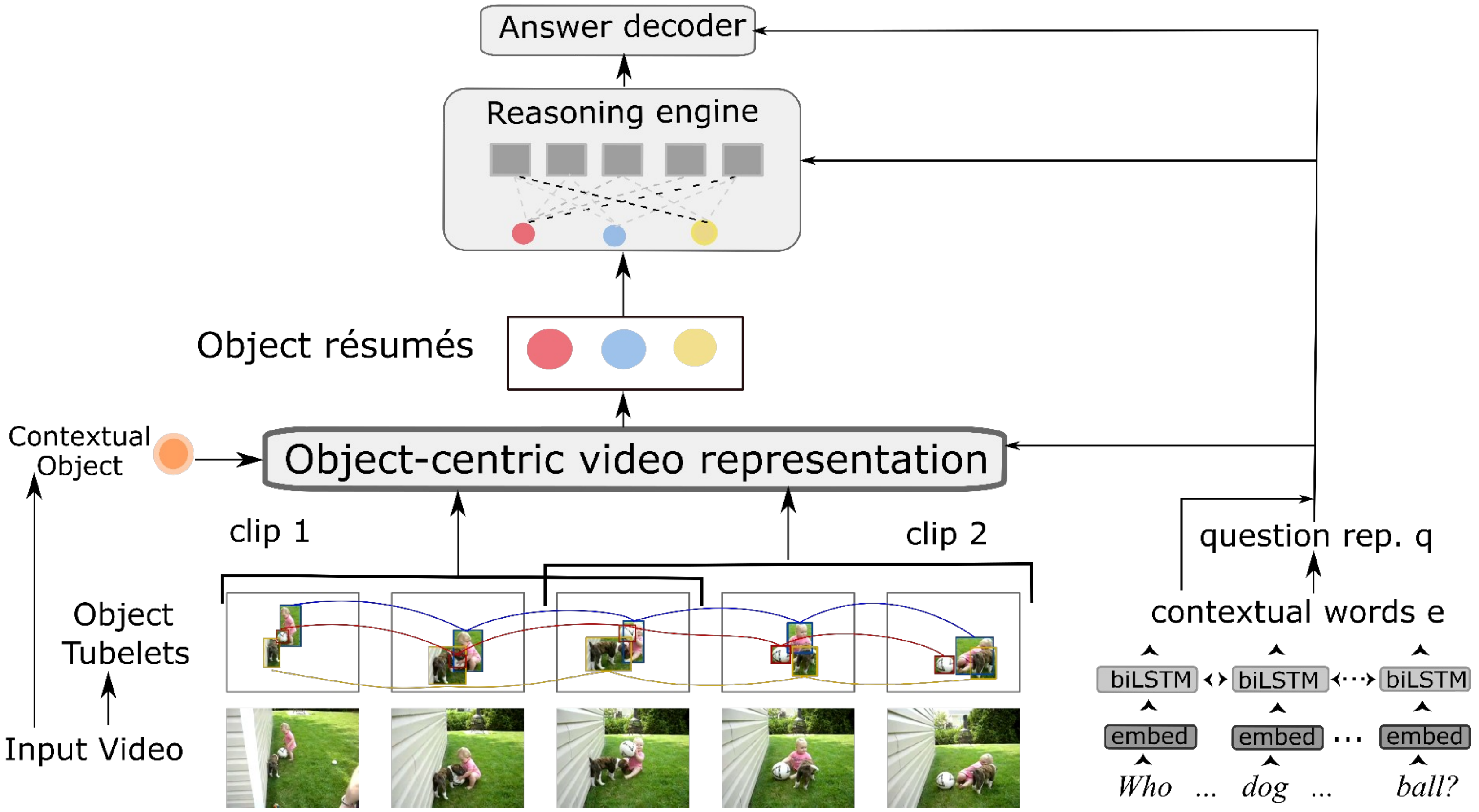
- **Objects** in video are primary constructs that have unique evolving lives throughout space-time
- To predict a correct answer, we need:
  - Understand the **evolution** of the object
  - Capture the contextualized **interaction** with its neighbours



Question:
What color is the big ball that starts moving after the white object rotating?

Answer:
Gray

Object-centric Neural Architecture for Video QA

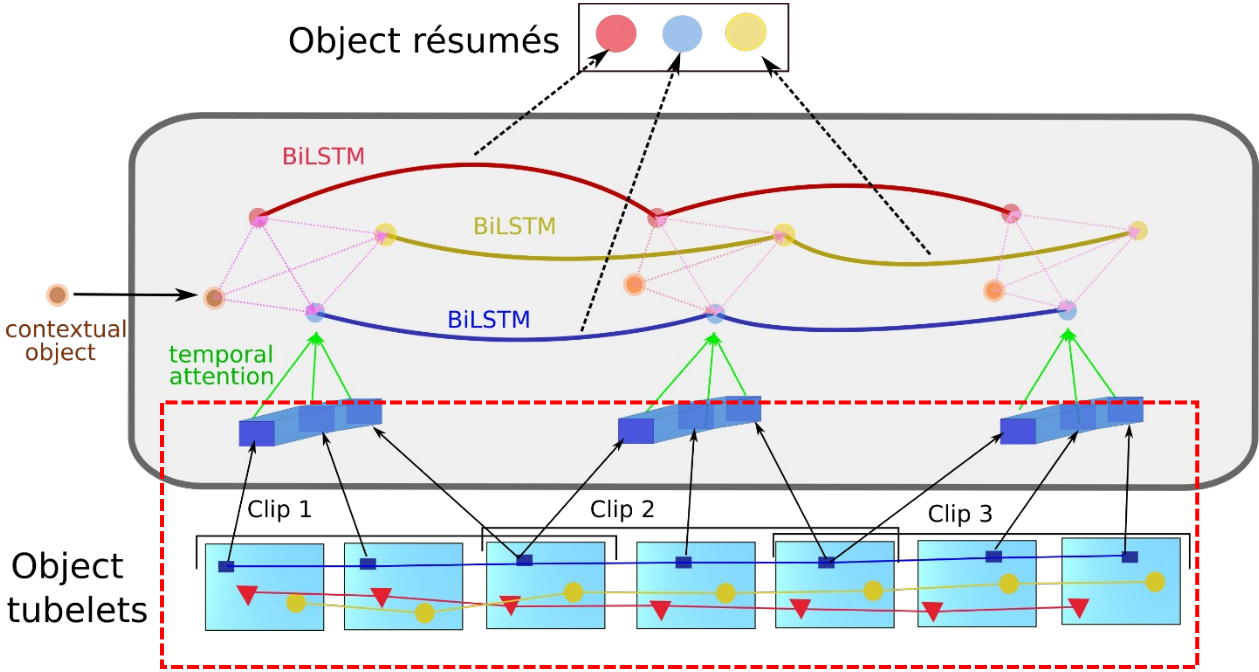


System 1: Object-centric Video Representation

Constructing Object Tubelets

- For each object:
  - The appearance feature
  - The geometrical feature : coordinates of the region box
- The *position-specific appearance* of object:

- Contextual object (ResNet features)



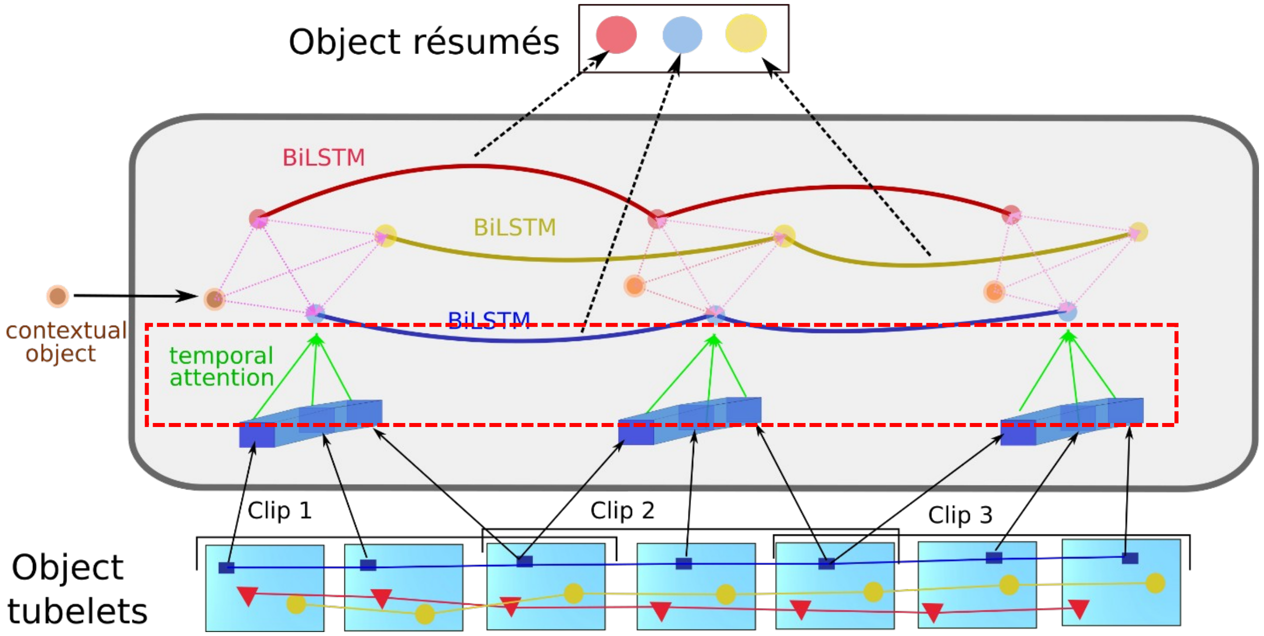
System 1: Object-centric Video Representation

Language-conditioned Representation

- Divide into K equal temporal parts:

  .

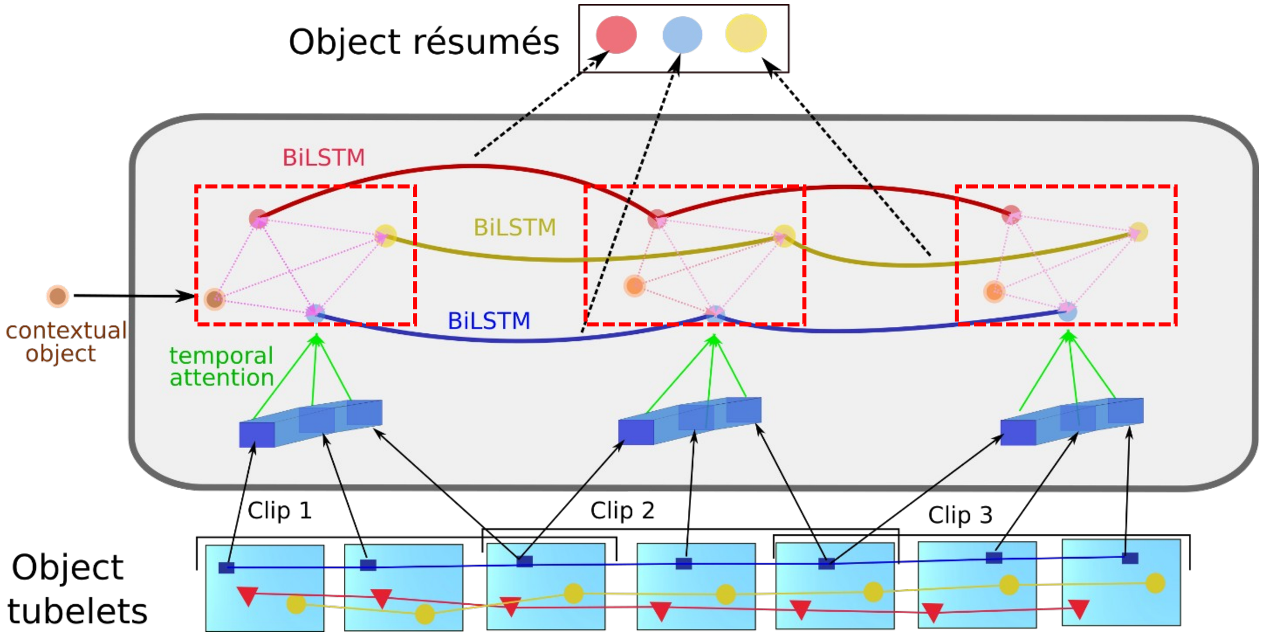
- where is the *position-specific appearance* feature.
- Temporal attention mechanism: reduce irrelevant visual information.
- Binary mask: exclude missed detections of objects.



System 1: Object-centric Video Representation

Query-conditioned Object Graph

- A graph
  - are nodes
  - Adjacent matrix is given by:
- Let , we refine representations of nodes:



System 1: Object-centric Video Representation

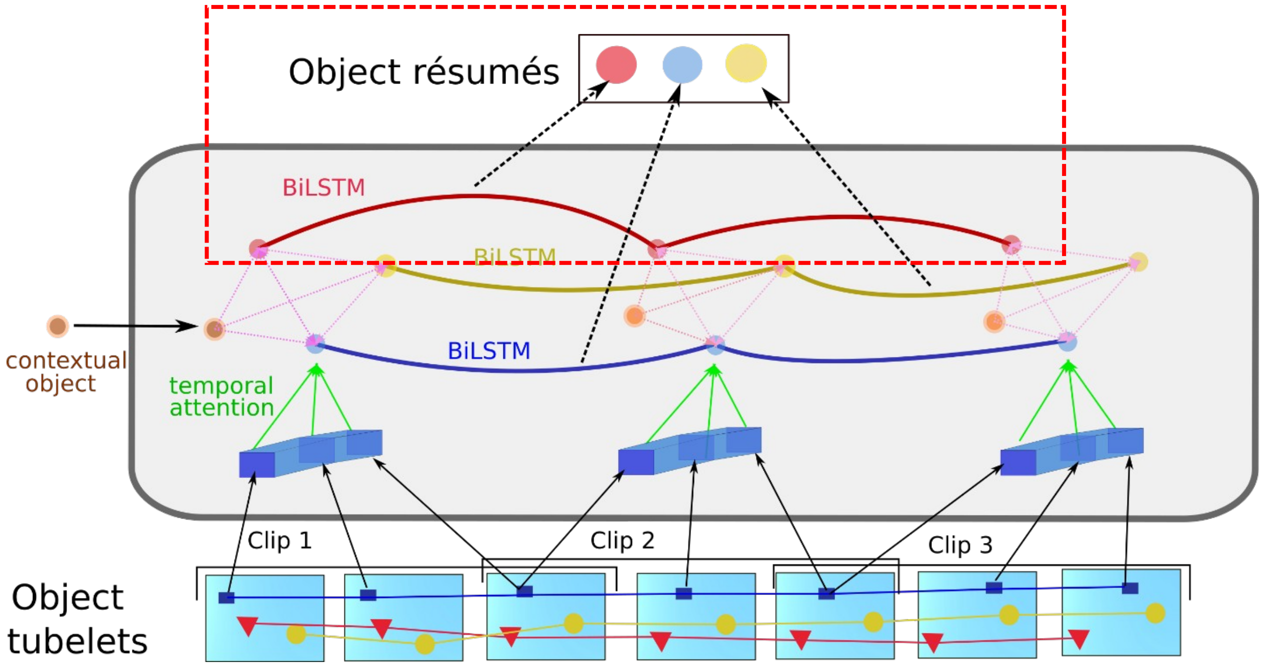
Video as Evolving Object Graph

- Temporal parts are then connected through a BiLSTM:

,

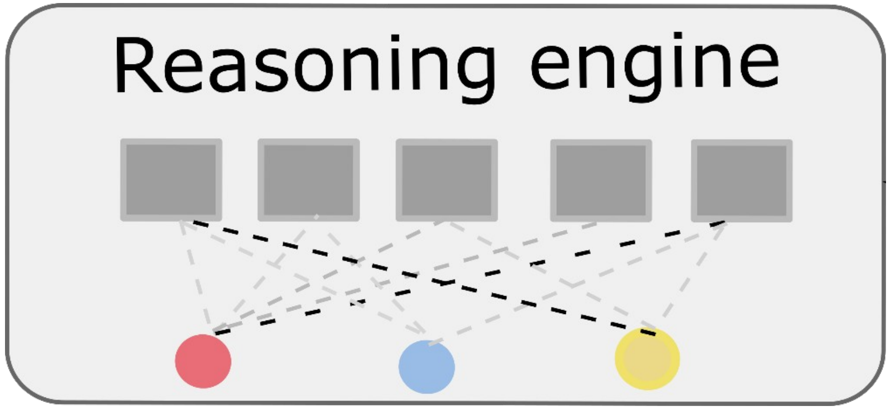
- Compute a résumé for each object by summarizing its lifetime:

and are end states of the BiLSTM.



System 2: General-proposed Reasoning Engine

- Our *object-centric video representation* can combine with a wide range of reasoning models.
  - MACNet (Hudson et al. 2018)
  - LOGNet (Le et al. 2020)



Experiments

Model	Test accuracy (%)	
	MSVD-QA	MSRVTT-QA
ST-VQA	31.3	30.9
Co-Mem	31.7	32.0
AMU	32.0	32.5
HME	33.7	33.0
HRA	34.4	35.1
HCRN	36.1	35.6
OCRL+LOGNet	**38.2**	**36.0**

Comparison with state-of-the-art-methods on three common datasets (MSVD-QA, MSRVTT-QA and SVQA). Our model is referred as OCRL+LOGNet.

Models	Exist	Count	Integer Comparison			Attribute Comparison					Query					All
			More	Equal	Less	Color	Size	Type	Dir	Shape	Colo r	Size	Type	Dir	Shape	
SA+TA	52.0	38.2	74.3	57.7	61.6	56.0	55.9	53.4	57.5	53.0	23.4	63.3	62.9	43.2	41.7	44.9
STRN	54.0	44.7	72.2	57.8	63.0	56.4	55.3	50.7	50.1	50.0	24.3	59.7	59.3	28.2	44.5	47.6
CRN+MAC	72.8	56.7	**84.5**	**71.7**	**75.9**	70.5	76.2	90.7	75.9	57.2	76.1	92.8	91.0	**87.4**	85.4	75.8
OCRL+MAC	77.4	56.7	81.2	64.6	65.0	90.0	93.4	90.1	77.0	93.5	**77.8**	**92.9**	**91.3**	82.5	**89.5**	77.8
OCRL+LOG	**81.7**	**61.5**	83.2	64.9	71.4	**92.7**	**97.2**	**94.6**	**88.8**	**95.7**	75.1	90.9	90.3	82.6	86.8	**79.5**

Conclusion

- Proposed a novel neural architecture for object-centric representation learning in video question answering.
- Introduced the concept of résumé that summarizes the live of an object over the entire video.

Thank you
QA