# Hierarchical Object-oriented Spatio-Temporal Reasoning for Video Question Answering

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


# General Video QA Framework

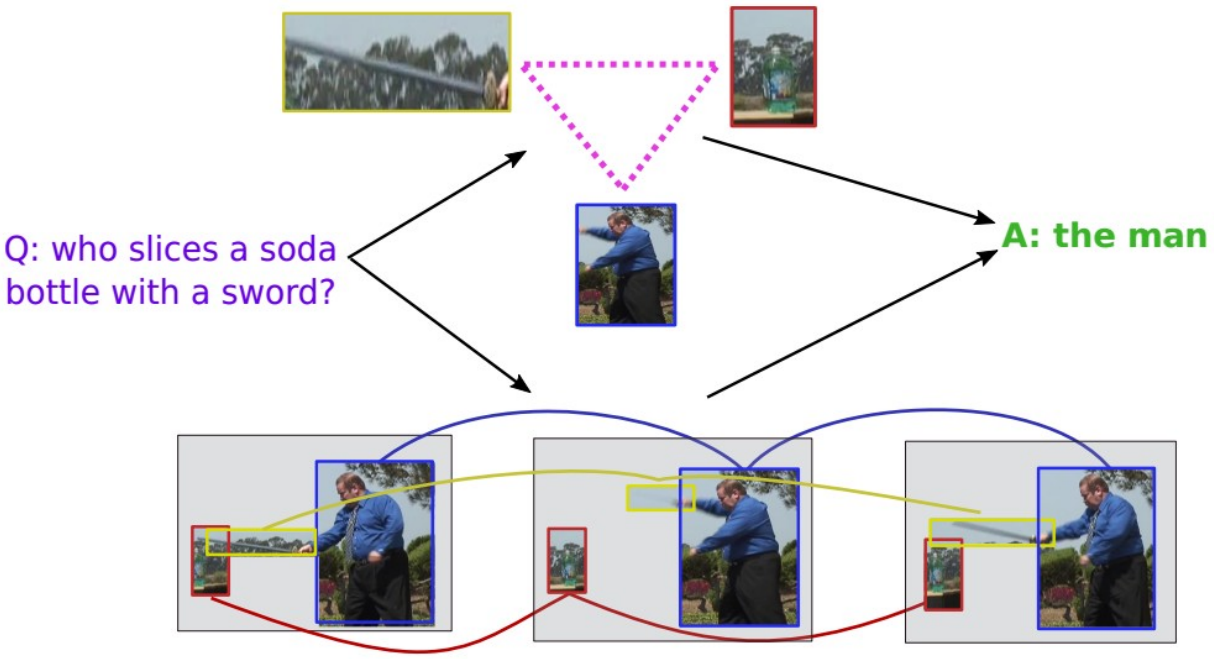
**Question:**
How many times does the cat lick?

Video QA Model

**Answer**
7 times

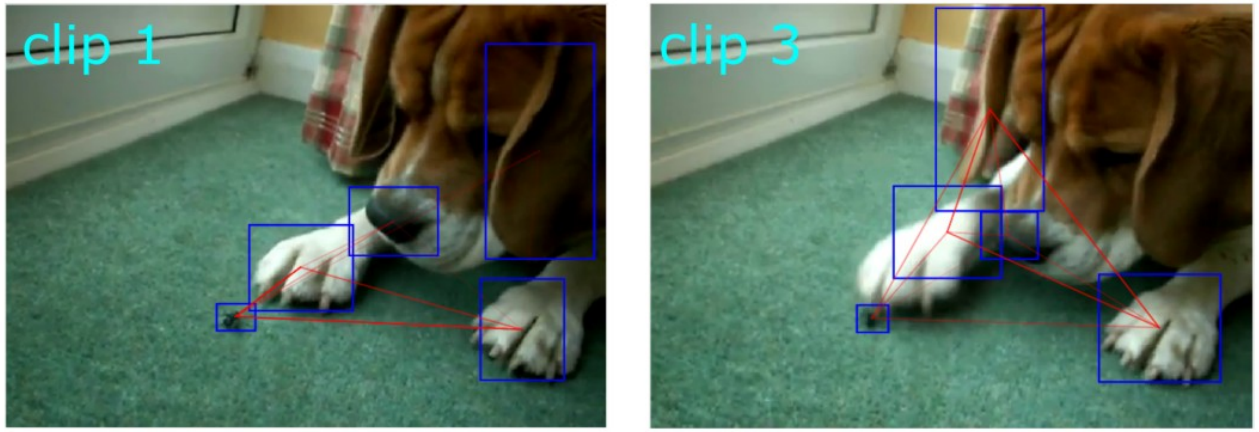
# Challenges in Video QA

- Extracting question-relevant high-level facts from low-level moving pixels over an extended period of time.
- Learning the long-term temporal relation of visual objects conditioning on the linguistic clues.



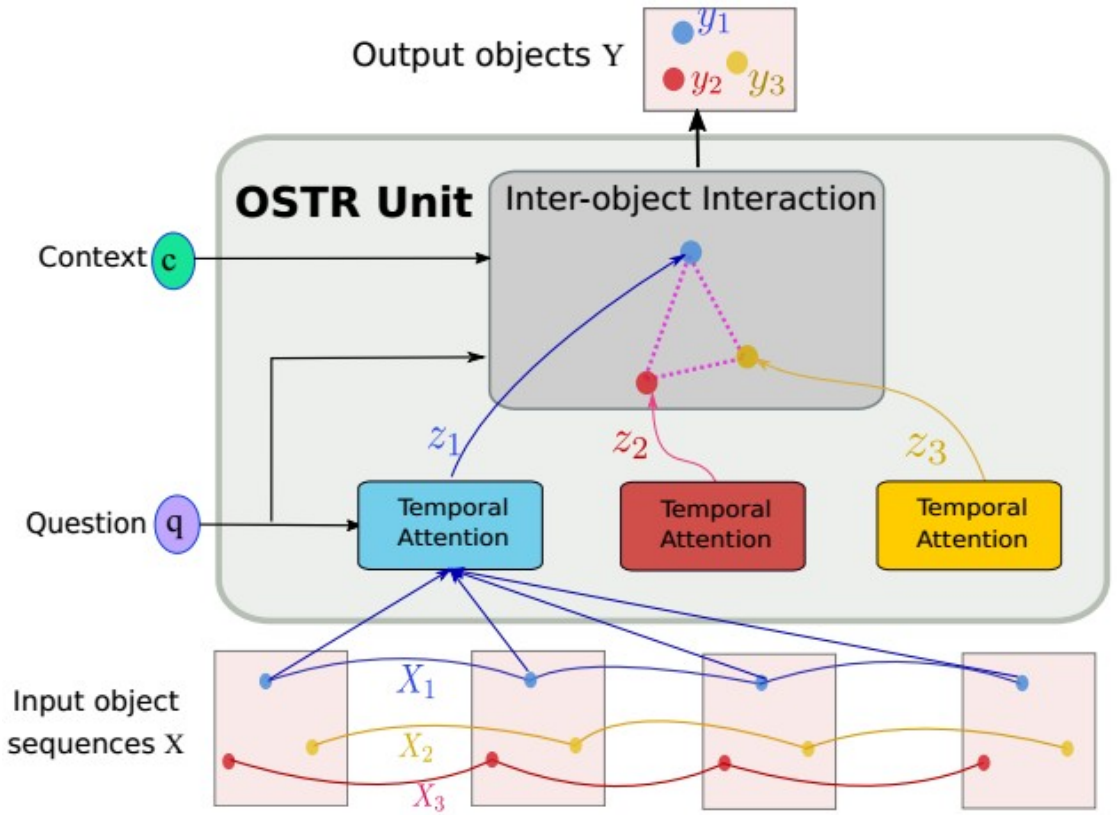
# Object-centric Learning

- Objects pave the way towards more human-like reasoning capability and symbolic computing.
- In video, object changes its appearance and position, and interacts with other objects at arbitrary time.



# Object-oriented Spatio-Temporal Reasoning (OSTR)

- A general purpose neural reasoning unit with dynamic object interactions per context and query.
- The OSTR leads to the efficiency of the reasoning process by containing: Intra-object temporal attention and Inter-object interaction.

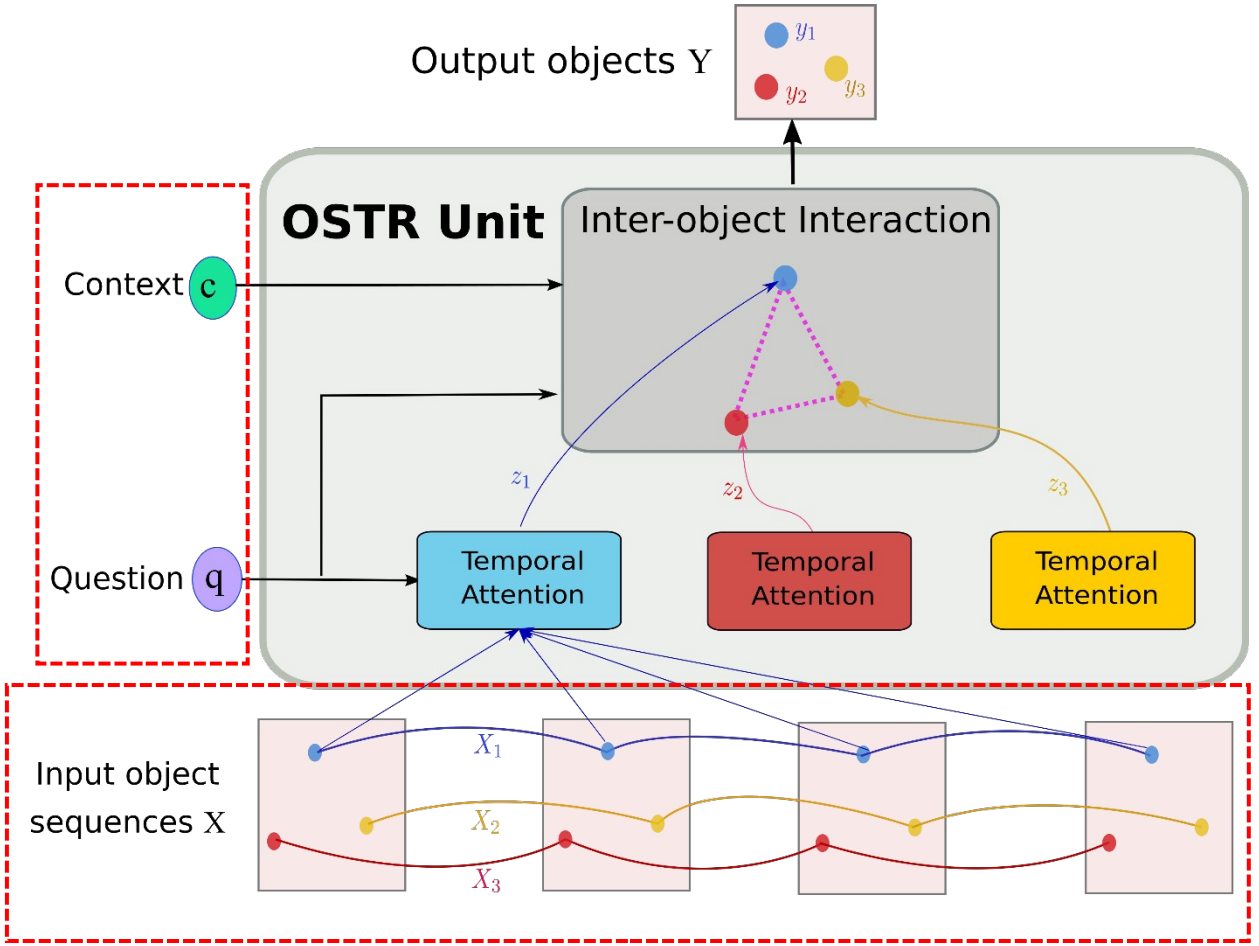


# Object-oriented Spatio-Temporal Reasoning (OSTR)

***Input***

- A set of object sequences

- A context representation :
  - Appearance features (ResNet)
  - Motion feature (ResNeXt)
- A query representation .



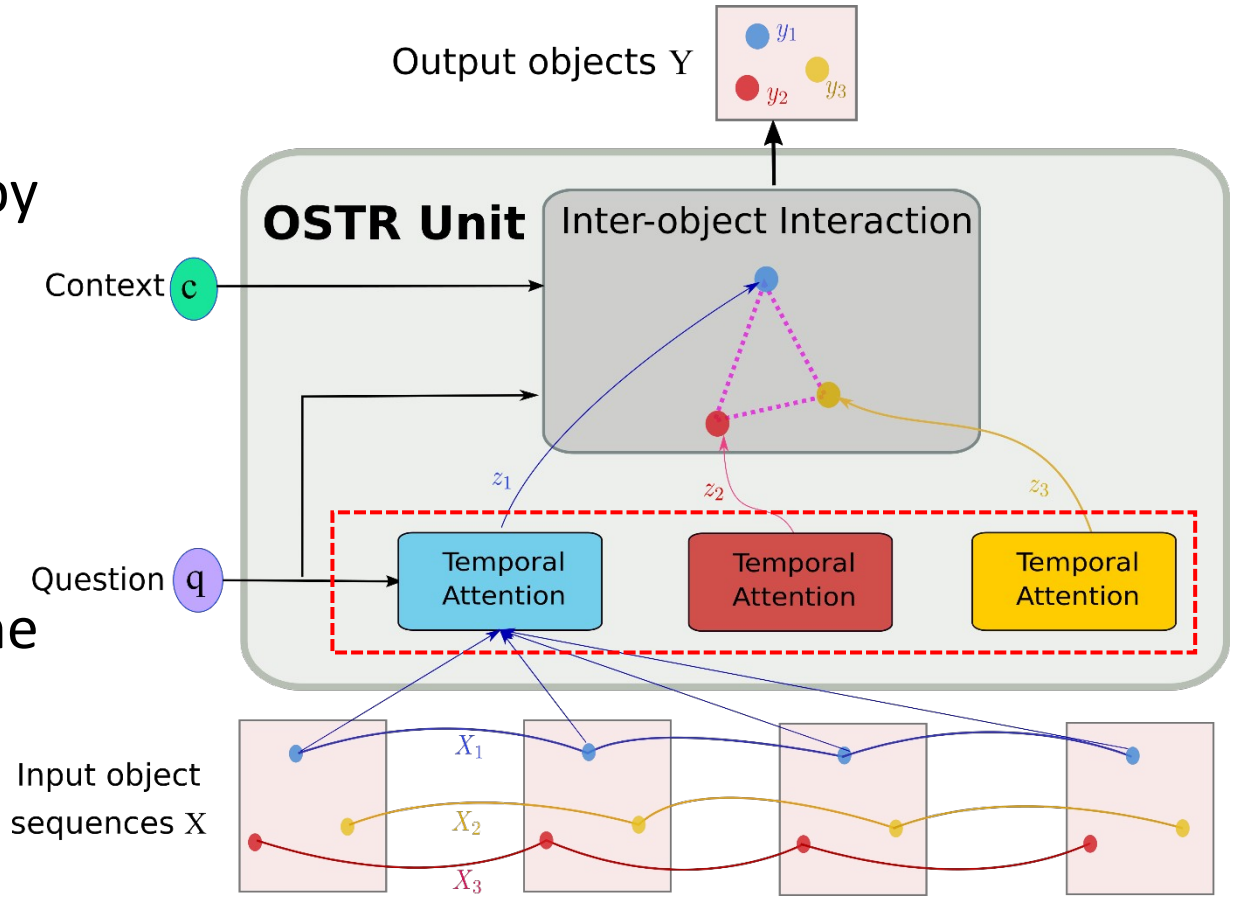
# Object-oriented Spatio-Temporal Reasoning (OSTR)

***Intra-object Temporal Attention***

- Each object sequence is summarized by a temporal attention module:

,

Withis a binary mask vector to handle the null values caused by missed detections.



# Object-oriented Spatio-Temporal Reasoning (OSTR)

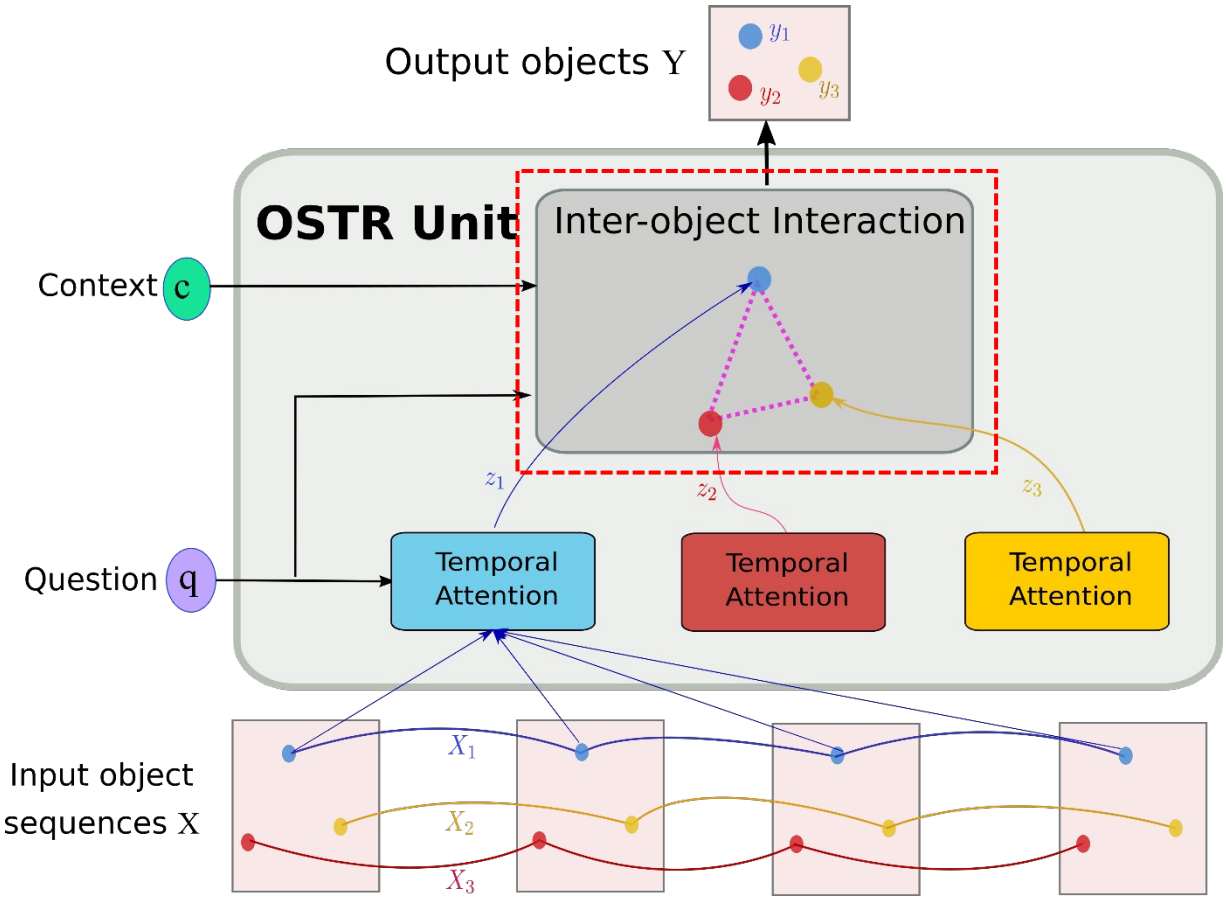
***Inter-object Interaction***

- The inter-object graph :
  - The summarized objects
  - The *query-induced correlation matrix*

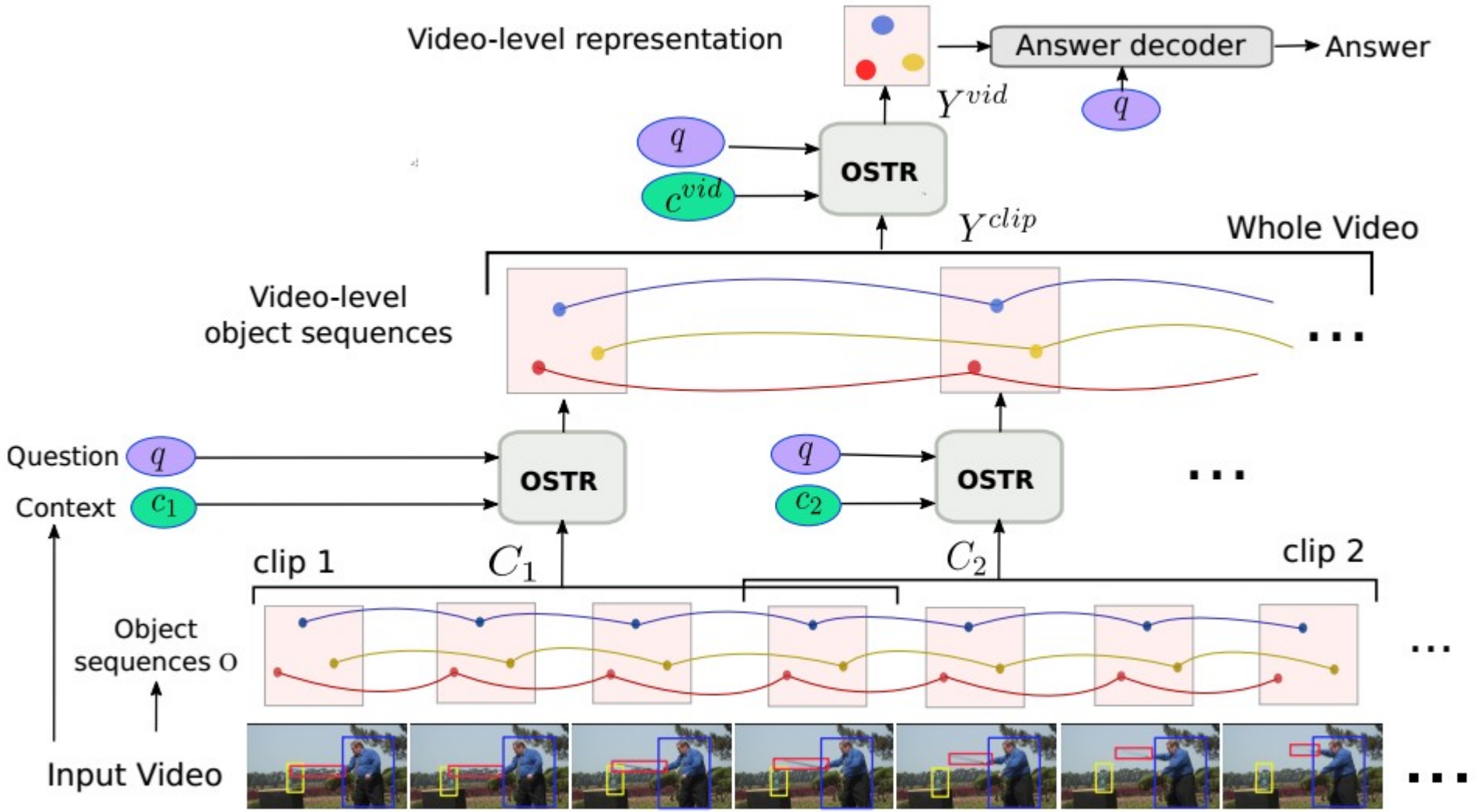
- Augment the object representations with the global context :

,

is the hidden states of the final GCNs layer.



# Hierarchical Object-oriented Spatio-Temporal Reasoning (HOSTR)



# Results

| Model | TGIF-QA | | | |
|---|---|---|---|---|
| | Action↑ | Trans.↑ | Frame↑ | Count↓ |
| ST-TP (R+C) | 62.9 | 69.4 | 49.5 | 4.32 |
| Co-Mem (R+F) | 68.2 | 74.3 | 51.5 | 4.10 |
| PSAC (R) | 70.4 | 76.9 | 55.7 | 4.27 |
| HME (R+C) | 73.9 | 77.8 | 53.8 | 4.02 |
| HCRN (R) | 70.8 | 79.8 | 56.4 | 4.38 |
| HCRN (R+F) | 75.0 | 81.4 | 55.9 | 3.82 |
| HOSTR (R) | **75.6** | 82.1 | **58.2** | 4.13 |
| HOSTR (R+F) | 75.0 | **83.0** | 58.0 | **3.65** |

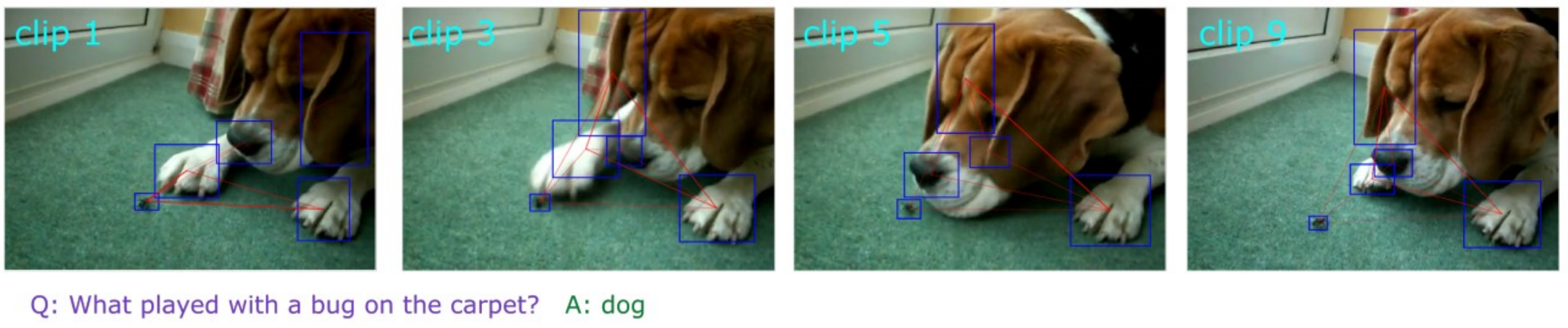
Performance on TGIF-QA

# Results (Cont.)

| 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 |
| HCRN | 36.1 | 35.4 |
| HOSTR | **39.4** | **35.9** |

Performance on MSVD-QA and MSRVTT-QA

# Qualitative Analysis




- HOSTR attended mostly on the objects related to the concepts relevant to answer the question.
- It intuitively agrees with how human might visually examine the scene given the question.

# Conclusion

- Introduce a general-purpose neural reasoning unit with dynamic object interactions per context and query.
- Design a hierarchical network that produces reliable and interpretable video question answering.

# Thank you !

Long Hoang Dang

*Email: hldang@deakin.edu.au*

**Applied Artificial Intelligence Institute,**

Deakin University

75 Pigdons Rd, Waurn Ponds VIC, Australia