Time-Evolving Conditional Character-centric Graphs for Movie Understanding

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

Introduction

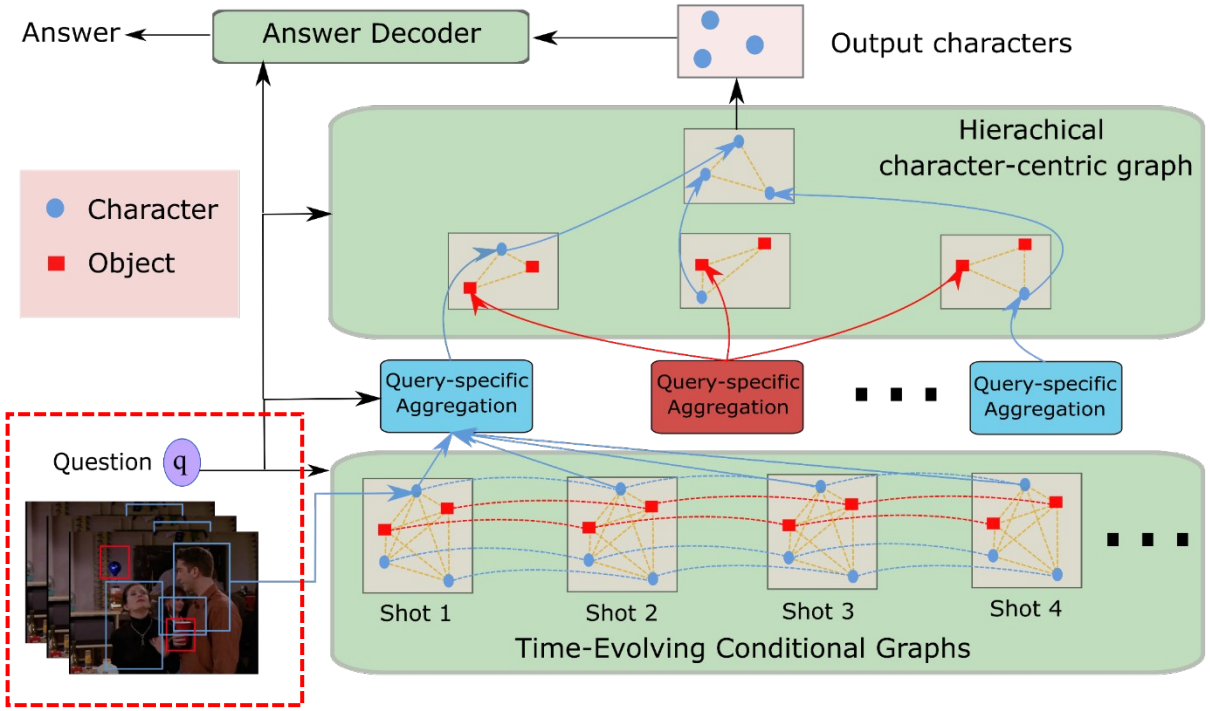
- Capturing the dynamic story in long-term human-centric video (Movie QA) presents a powerful testbed for temporal graph modeling.
- Challenge: Learning the dynamic spatio-temporal interactions of human actors and other objects implicitly from visual information.

Time-Evolving Conditional Character-centric Graph (TECH)

Input

- A set of human characters and non-human object sequences over *S* shots

- A contextual embedding .
- A global vector .



Time-Evolving Conditional Character-centric Graph (TECH)

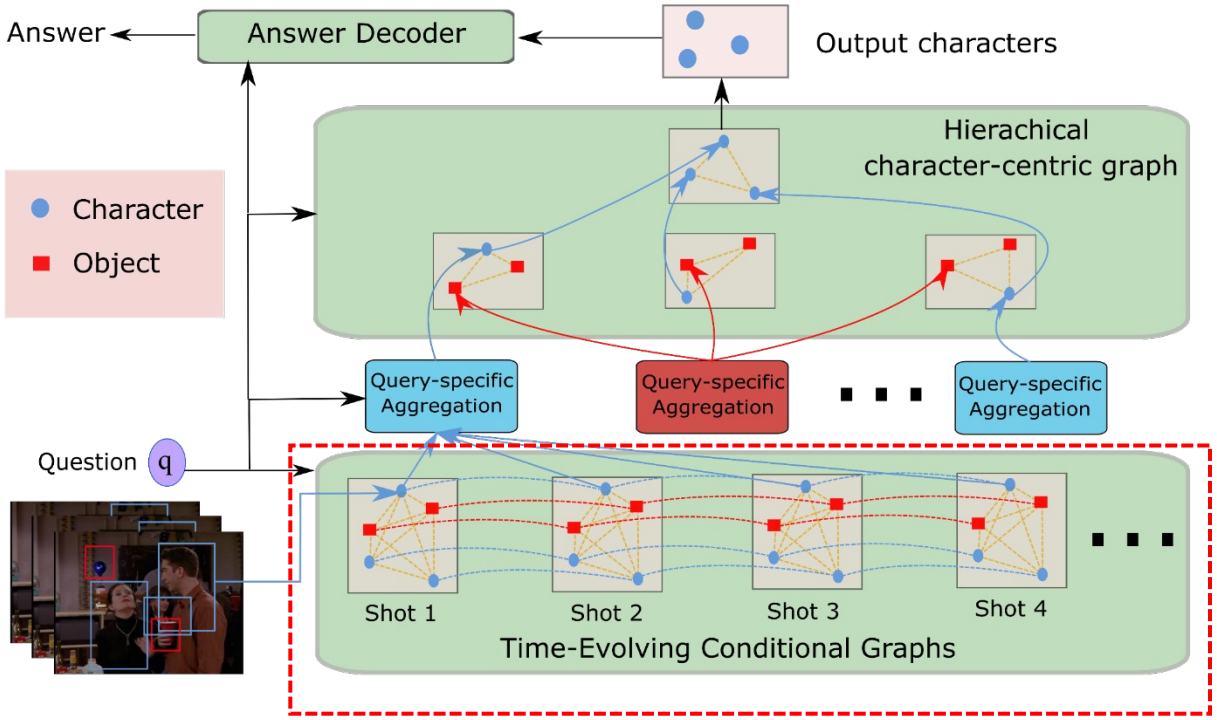
Shot-based Entities Graph

- TECH is a recurrent system of query-conditioned dynamic graphs.
- We build a dynamic graph for each video shot s:

;

for

and ;

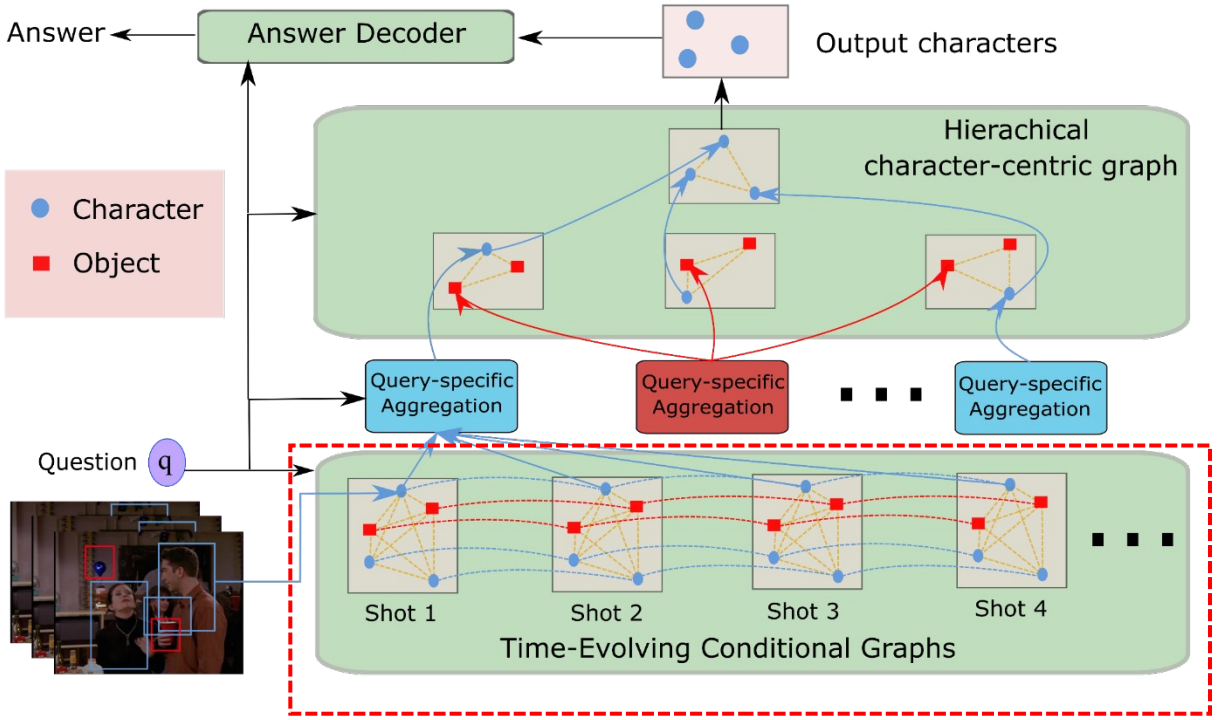


Time-Evolving Conditional Character-centric Graph (TECH)

Entity-based Graph Evolving

- Refining the representation of each vertex at the present shot s by:
  - Spatial relationship with neighbors.
  - Temporal clues from the previous shots.
- Assuming as initial representation:

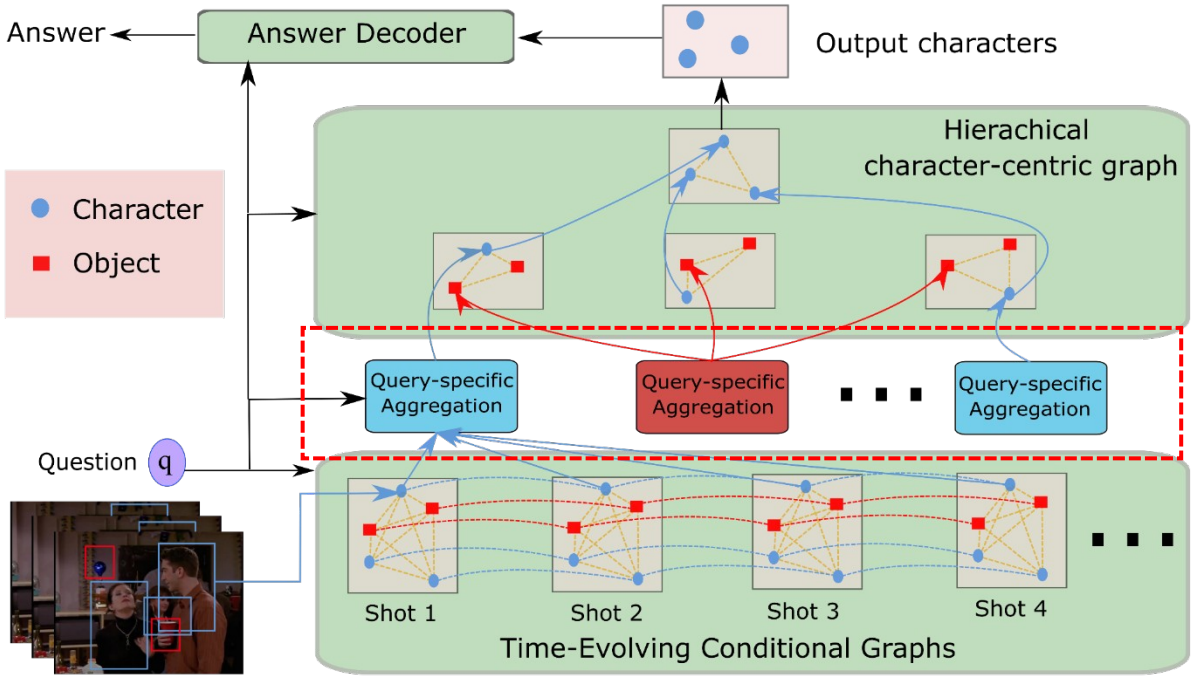
;



Time-Evolving Conditional Character-centric Graph (TECH)

Query-specific Aggregation

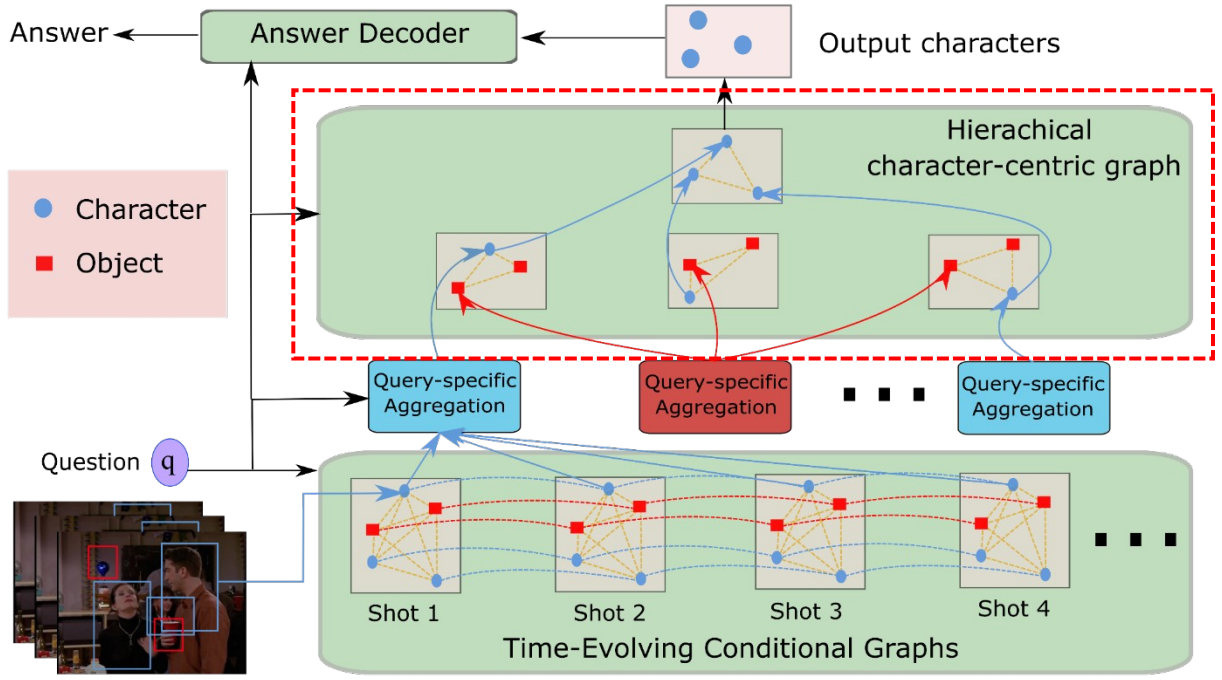
- A movie may contain a large amount of information, the information relevant to a query is more specific.
- We design a *query-specific temporal information aggregation* module to retrieve visual moments that are only relevant to the query.



Time-Evolving Conditional Character-centric Graph (TECH)

Hierarchical Character-centric Graph

- There are two types of entity-level relations of interest:
  - Character-character
  - Character-object
- Two stages of feature refinement:
  - Object-to-character refinement
  - Character-to-character refinement



Results

Models	Val. Acc↑
TVQA w. CNN feat. [10]	42.01
TVQA w. visual concept [10]	44.27
BERT Video QA [16]	44.63
STAGE [11]	45.83
DenseCap* [4]	45.85
TECH	**47.79**

Performance on TVQA dataset

Qualitative Analysis

Question: What did Monica do after she walked in the door ?

Answer candidates:

A. grabbed a bottle of water B. set her purse down C. took her phone out

D. showed Rachel a check E. jumped up and down with joy

Ground truth: B. set her purse down TECH: B. set her purse down DenseCap: D. showed Rachel a check

Question: What color mug does Joey have next to him when Monica sits down next to him ?

Answer candidates:

A. Green B. Red C. Yellow D. Blue E. White

Ground truth: D. Blue TECH: D. Blue DenseCap: C. Yellow

Qualitative examples show advantages of TECH in handling long-term temporal relationships in video while DenseCap struggle.

Conclusion

- Designing TECH as a recurrent system of query-dependent dynamic graphs that allow information to effectively flow from early points to later points in time.
- Showing the benefits of paying attention to human characters and their interactions within a movie clip over the interactions with other non-human objects.

Thank you !

Long Hoang Dang

Email: hldang@deakin.edu.au

Applied Artificial Intelligence Institute,

Deakin University

75 Pigdons Rd, Waurn Ponds VIC, Australia