Video as Conditional Graph Hierarchy for Multi-Granular Question Answering

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, Tat-Seng Chua
Department of Computer Science, National University of Singapore

- **Motivation.** As presented in Figure 1, while videos are presented in frame sequences, the visual elements (objects, actions, activities and events) are not sequential but rather hierarchical (bottom-up view) in semantic space. To align with the multi-granular essences of linguistic concepts in language queries (top-down view), we propose to model the video as a conditional graph hierarchy to advance video question answering in a multi-granular fashion.

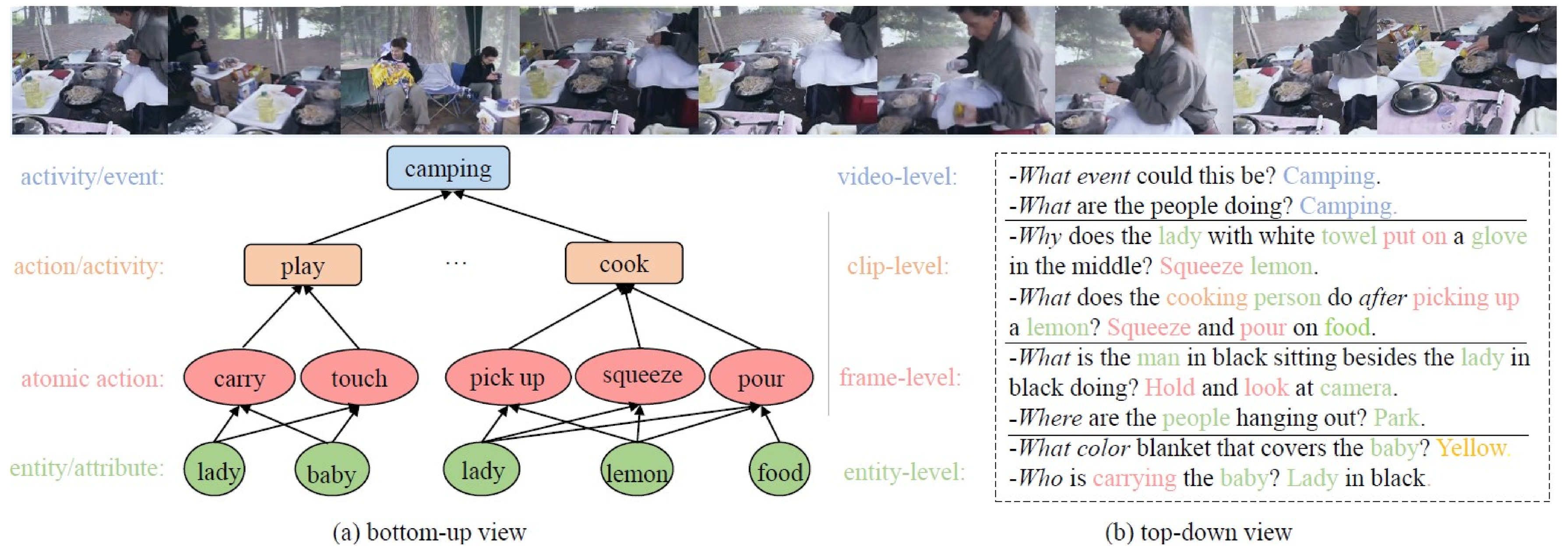

Figure 1. Illustration of the bottom-up and top-down views for VideoQA.

- **Method.** As shown in Figure 2, our model (HQGA) includes 3 graph hierarchies that operates at different levels to reason and aggregate visual elements of different granularities into a global representation.

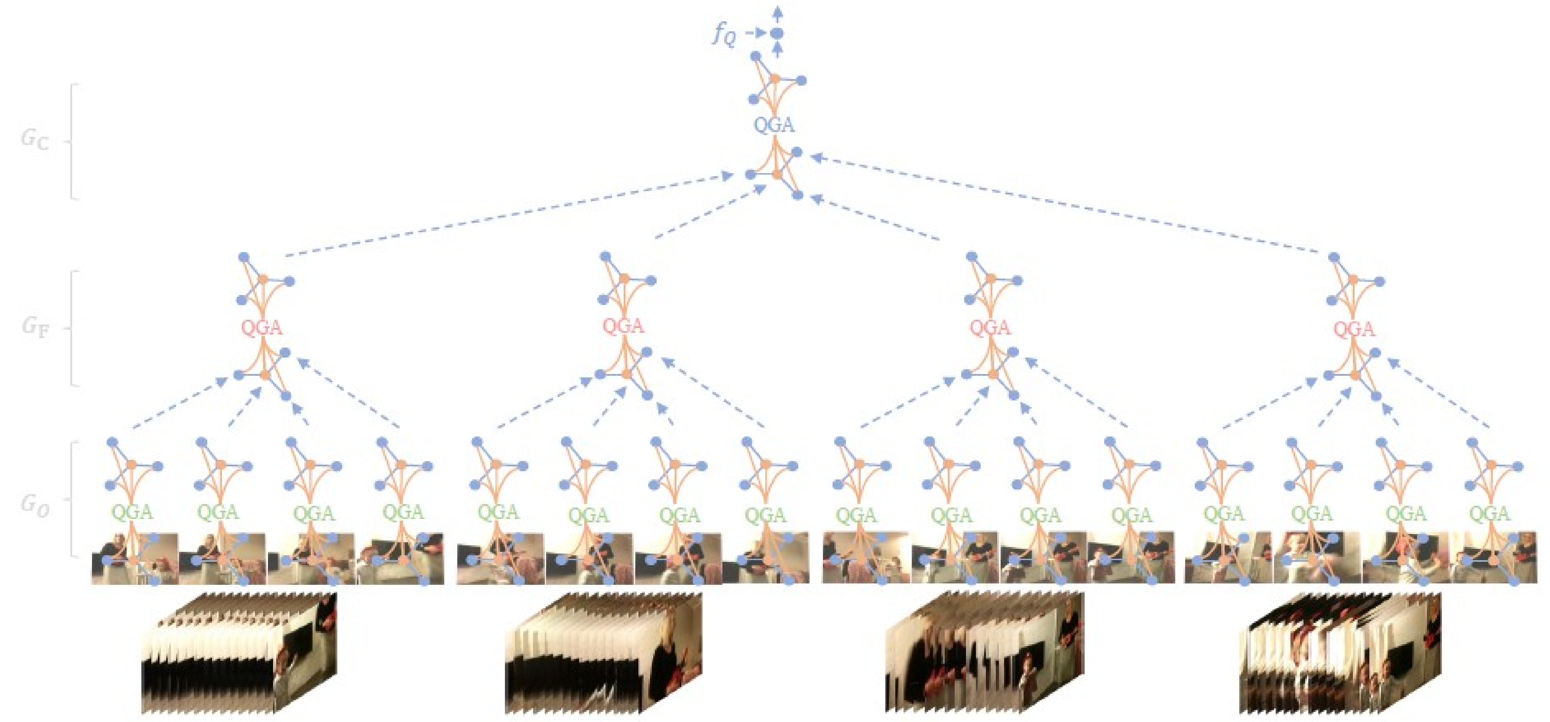

Figure 2. Overview of HQGA architecture.

- $\mathbf{G_O}$**:** Operate over regions to capture a snapshot of object interaction at frame level.
- $\mathbf{G_F}$**:** Operate over the outputs of G_O clip wisely, to model a short term interaction dynamics and to reason low-level elements int high-level components.
- $\mathbf{G_C}$**:** Operate over the outputs of G_F to aggregate the local, short term interactions into a global, video level representation.

- Our model architecture was achieved by level-wisely stacking a **Q**uery-conditioned **G**raph **A**ttention (**QGA**) unit as illustrated in Figure 3.
- QGA first contextualizes a set of input visual nodes X_{in} in relation to their neighbors under the condition of a language query Q, and then aggregates the contextualized output nodes X_{out} into a single global descriptor x_p.

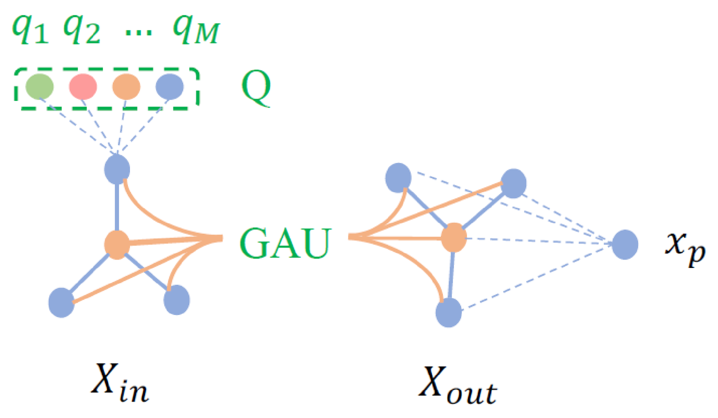

Figure 3. Illustration of QGA unit.

- **Experiments.** To validate our model's effectiveness, we experiment on four datasets that challenge the various aspects of video understanding from recognition of shallow object and activity, reason of action repetition and state transition, to deeper causal and temporal action interaction among multiple objects.

Datasets	Main Challenges	#Videos/#QAs	Train	Val	Test	VLen (s)	QA
MSRVTT-QA	Object & Action Recognition	10K/ 244K	6.5K/159K	0.5K/12K	3K/73K	15	OE
MSVD-QA	Object & Action Recognition	1.97K/ 50K	1.2K/30.9K	0.25K/6.4K	0.52K/13K	10	OE
TGIF-QA	Repetition Action	22.8K/22.7K	20.5K/20.5K	-	2.3K/2.3K	3	MC
	State Transition	29.5K/58.9K	26.4K/52.7K	-	3.1K/6.2K	3	MC
	Frame QA	39.5K/53.1K	32.3K/39.4K	-	7.1K/13.7K	3	OE
NExT-QA	Causal & Temporal Interaction	5.4K/48K	3.8K/34K	0.6K/5K	1K/9K	44	MC

- HQGA shows superior performances over previous methods on all 4 datasets. It also wins across per-question type as categorized in NExT-QA and TGIF-QA.

Models	Causal	Temp.	Descrip.	Overall
ST-VQA	44.76	49.26	55.86	47.94
Co-Mem	45.22	49.07	55.34	48.04
HME	46.18	48.20	58.30	48.72
L-GCN	45.15	50.37	55.98	48.52
HGA	46.26	50.74	59.33	49.74
HCRN	45.91	49.26	53.67	48.20
HQGA (Ours)	**48.48**	**51.24**	**61.65**	**51.42**

Accuracy on NExT-QA val set.

Models	Causal	Temp.	Descrip.	Overall
ST-VQA	45.51	47.57	54.59	47.64
Co-Mem	45.85	50.02	54.38	48.54
HME	46.76	48.89	57.37	49.16
L-GCN	47.85	48.74	56.51	49.54
HGA	48.13	49.08	57.79	50.01
HCRN	47.07	49.27	54.02	48.89
HQGA (Ours)	**49.04**	**52.28**	**59.43**	**51.75**

Accuracy on NExT-QA test set.

Models	TGIF-QA Action	TGIF-QA Trans.	TGIF-QA FrameQA	MSRVTT-QA	MSVD-QA
ST-VQA	62.9	69.4	49.50	30.9	31.3
PSAC	70.4	76.9	55.7	-	-
STA	72.3	79.0	56.6	-	-
MIN	72.7	80.9	57.1	35.4	35.0
QueST	75.9	81.0	59.7	34.6	36.1
AMU	-	-	-	32.5	32.0
Co-Mem	68.2	74.3	51.5	31.9	31.7
HME	73.9	77.8	53.8	33.0	33.7
L-GCN	74.3	81.1	56.3	33.7	34.3
HGA	75.4	81.0	55.1	35.5	34.7
DualVGR	-	-	-	35.5	39.0
GMIN	73.0	81.7	57.5	36.1	35.4
B2A	75.9	82.6	57.5	36.9	37.2
HCRN	75.0	81.4	55.9	35.6	36.1
HOSTR	75.0	83.0	58.0	35.9	39.4
HQGA	**76.9**	**85.6**	**61.3**	**38.6**	**41.2**

Accuracy on test sets of TGIF/MSRVTT/MSVD.

- The hierarchical structure contributes ~2.6% and 1.5% on MSRVTT-QA and NExT-QA respectively.
- The graph operation contributes ~2.4% and 1.1% on MSRVTT-QA and NExT-QA respectively.
- The multi-level token-wise condition contributes 1.2% and 0.8% on MSRVTT-QA and NExT-QA respectively.

Model Variants	NExT-QA	MSRVTT-QA
HQGA	**51.42**	**38.23**
w/o G_O	50.50	37.26
w/o G_F	50.00	37.05
w/o G_O & G_F	49.96	35.66
w/o G_C(s)	50.74	37.69
w/o G_C & G_F(ss)	50.44	36.94
w/o G_C & G_F & G_O(sss)	50.32	35.88
w/o Q_C	51.30	38.17
w/o Q_C & Q_F	51.08	37.62
w/o Q_C & Q_F & Q_O	50.62	37.03
w/ f_Q	50.16	37.52
w/o F_m	50.90	37.94
w/o F_a & F_m	50.34	37.86

- Our model works as a fully-differentiable, query-instantiated neural modular network. The fully-attention based implementation enables the visualization of the learned conditional attention weight with regard to the specific query & prediction.

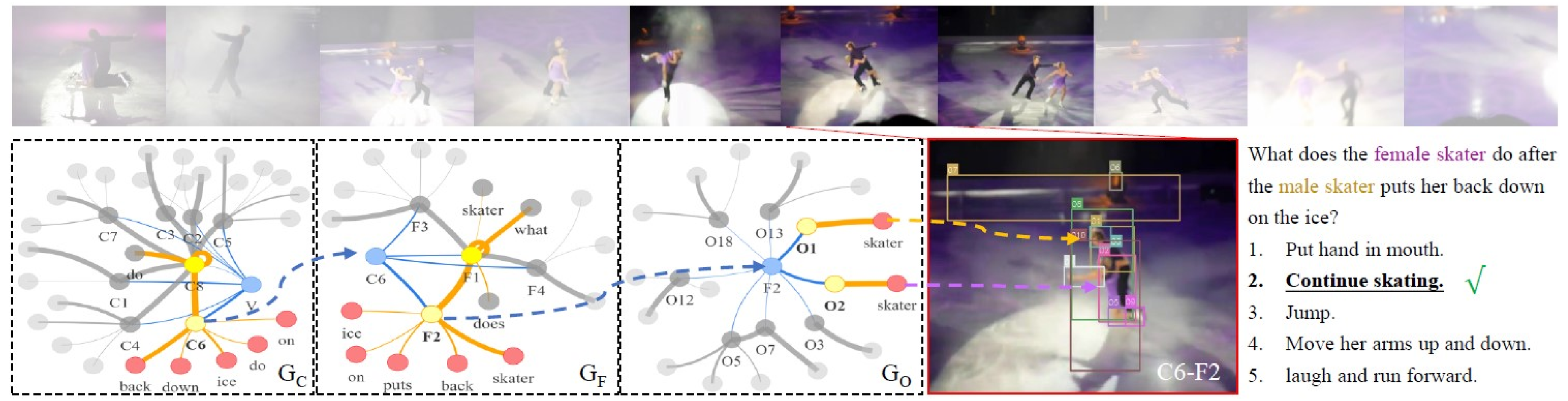

Figure 4. Visualization of the predictions and learned attention weights.

- Correctly find the relevant video moments (C_5&C_6) and also the related objects (man O_1 and female skater O_2) for the correct prediction.
- Graph nodes at high-levels response stronger to dynamic actions, while those at the bottom level response stronger to static things, e.g., objects & attributes.

- **Conclusion.**
- We provide the bottom-up and top-down insights to advance video question answering in a hierarchical, multi-granular fashion.
- We propose to model the video as a conditional graph hierarchy which is achieved by level-wisely stacking a query-conditioned graph attention module.
- Our model is effective, easy to understand, and is of enhanced generalizability; it shows superior performance to prior methods (w/o cross-model pre-training) across 4 datasets and also finds introspective evidences to understand the predictions.