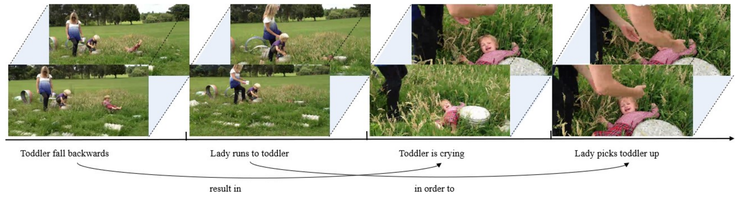
NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions

Junbin Xiao, Xindi Shang, Angela Yao, Tat-Seng Chua
Department of Computer Science, National University of Singapore

➢ Introduction

- Actions in videos are often not independent but rather related with causal and temporal relationships.
- Recognizing/describing the independent object/action in a video is now attainable with advanced neural network models.
- Understanding the causal and temporal interaction lies at the core of human intelligence, but remains a great challenge for AI and is also much less explored.
- **NExT-QA** is proposed to benchmark such a problem; it hosts causal and temporal action reasoning in VideoQA and is rich in object interaction in daily activities.

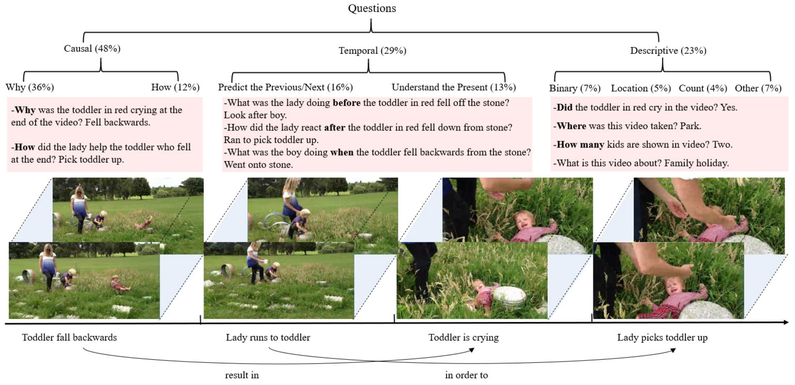


➢ Attractive Property

- The 1st VideoQA benchmark challenges comprehending the causal & temporal action interaction, advancing video understanding from recognition to explanation.
- Comprehensive baselines and analysis for both multi-choice and open-ended QA, along with detailed question types to help analyze the VQA models.
- High quality (with multi-turn manual annotation & check).

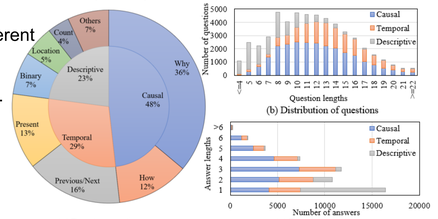
➢ Dataset Overview

- **Causal questions** explain why something happen or how to bring a visual effect.
- **Temporal questions** challenge understanding the order of the actions.
- **Descriptive Questions** aim at a scene-level recognition of the visual facts in videos.

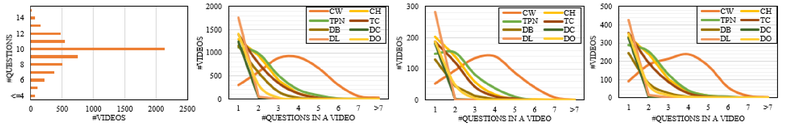


➢ Data Statistic

- 5440 Videos (train/val/test: 3870/570/1000) with 52*K* question-answer pairs (48% of causal questions, 29% of temporal actions and 23% of descriptive questions).
- ~10 questions in each video covering different type of topics.
- Average video length is 44s, question length is 12, and answer length is 3.
- Different type of questions are evenly distributed on the train/val/test sets.



(a) Distribution of question types (b) Distribution of questions (c) Distribution of answers



(a) Distribution of questions per video. (b) Train (c) Val (d) Test

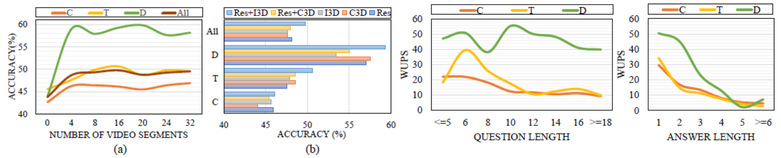
➢ Experiment

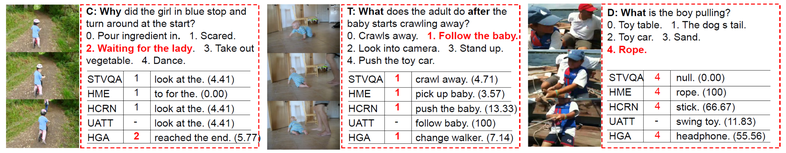
Methods	Text Rep.	Acc_C Why	Acc_C How	Acc_C All	Acc_T Prev&Next	Acc_T Present	Acc_T All	Acc_D Count	Acc_D Location	Acc_D Other	Acc_D All	Acc
EVQA [2]	GloVe	28.38	29.58	28.69	29.82	33.33	31.27	43.50	43.39	38.36	41.44	31.51
PSAC [28]	GloVe	35.81	29.58	34.18	28.56	35.75	31.51	39.55	67.90	35.41	48.65	35.57
PSAC+ [28]	GloVe	35.03	29.87	33.68	30.77	35.44	32.69	38.42	71.53	38.03	50.84	36.03
CoMem [12]	GloVe	36.12	32.21	35.10	34.04	41.93	37.28	39.55	67.12	40.66	50.45	38.19
STVQA [20]	GloVe	37.58	32.50	36.25	33.09	40.87	36.29	45.76	71.53	44.92	55.21	39.21
HGA [21]	GloVe	36.38	33.82	35.71	35.83	42.08	38.40	**46.33**	70.51	46.56	55.60	39.67
HME [9]	GloVe	39.14	34.70	37.97	34.35	40.57	36.91	41.81	71.86	38.36	51.87	39.79
HCRN [25]	GloVe	39.86	36.90	39.09	37.30	43.89	40.01	42.37	62.03	40.66	49.16	40.95
EVQA [2]	BERT-FT	42.31	42.90	42.46	46.68	45.85	46.34	44.07	46.44	46.23	45.82	44.24
STVQA [20]	BERT-FT	45.37	43.05	44.76	47.52	51.73	49.26	43.50	65.42	53.77	55.86	47.94
CoMem [12]	BERT-FT	46.15	42.61	45.22	48.16	50.38	49.07	41.81	67.12	51.80	55.34	48.04
HCRN* [25]	BERT-FT	**46.99**	42.90	45.91	48.16	50.83	49.26	40.68	65.42	49.84	53.67	48.20
HME [9]	BERT-FT	46.52	**45.24**	46.18	47.52	49.17	48.20	45.20	**73.56**	51.15	58.30	48.72
HGA [21]	BERT-FT	**46.99**	44.22	**46.26**	**49.53**	**52.49**	**50.74**	44.07	72.54	**55.41**	**59.33**	**49.74**

Results of multi-choice QA on val set

Methods	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
Popular	9.73	8.95	28.39	13.40
BlindQA	12.14	14.85	40.41	18.88
STVQA [19]	12.52	14.57	45.64	20.08
HME [9]	12.83	14.76	45.13	20.18
HCRN [25]	12.53	15.37	45.29	20.25
UATT [56]	13.62	**16.23**	43.41	20.65
HGA [21]	**14.76**	14.90	**46.60**	**21.48**

Results of open-ended QA on val set





➢ Discussion & Conclusion

- SOTA methods perform well on descriptive questions but are weak in causal and temporal action reasoning
- Methods that are effective in multi-choice QA struggle in generating the answers in open-ended QA scenario.
- Future efforts can be made in understanding the rich object interactions and capturing the causal and temporal relationship.
- Data and code is available at: https://github.com/doc-doc/NExT-QA.