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

Junbin Xiao, Xindi Shang, Angela Yao, Tat-Seng Chua
Department of Computer Science, National University of Singapore
{junbin,shangxin,ayao,chuats}@comp.nus.edu.sg

1. NExT-QA Dataset

1.1. Data Statistics

As shown in Figure 1, questions in NExT-QA mostly ask *'why did/does ...'*, and *'how/what did/does ...'*. This reveals that NExT-QA advances existing VideoQA datasets that pay attention to scene recognition (*what/who/where/which is/are ...*) towards the explanation of temporal actions. The rich causal and temporal questions make NExT-QA a unique QA dataset for video understanding. Other details of the dataset are shown in Figure 2 and Figure 3.

1.2. Dataset Comparison

In Figure 4, we compare NExT-QA with several related video QA datasets in terms of the distributions of QAs. From Figure 4 (a) we can see that most of the questions in NExT-QA are relatively long, with an average of 12 words per question. Questions in MSVD-QA [9] and MSRVTT-QA [9] on the other hand are the shortest among the compared datasets (mostly 5 words). TGIF-QA [3] and ActivityNet-QA [11] have about 8 words in most of their questions. Similarly, Figure 4 (b) shows that the answers in NExT-QA are longer, with an average of 3 words, whereas TGIF-QA and ActivityNet-QA are dominated by one-word answers. In addition, we also find that all the answers in MSVD-QA and MSVTT-QA are with only one word. The relatively longer questions and answers in NExT-QA enable much more interesting QA contents, *i.e.*, from recognition to explanation of video contents. In Figure 4 (c), we show the frequency of the answer words in terms of their part-of-speech (POS) tags, from which we can see that the answers in NExT-QA are much richer in verbs because it focuses on the causal and temporal action reasoning. Though ActivityNet-QA and TGIF-QA explore temporal actions as well, they emphasize action recognition and object/repetition count. As a result, their answers are dominated by nouns and numbers.

The above statistical comparisons demonstrate that our NExT-QA dataset opens new challenges and opportunities for deeper understanding of video contents that goes beyond

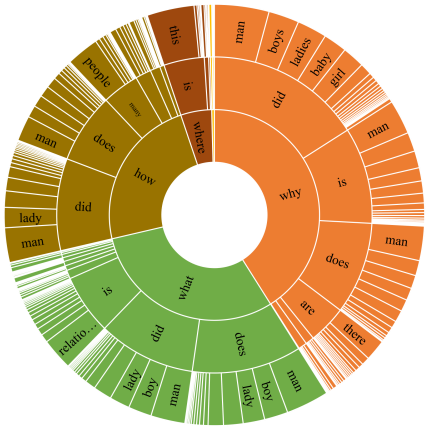


Figure 1: Distribution of NExT-QA questions by first three words. (The word 'the' in each question are ignored.)

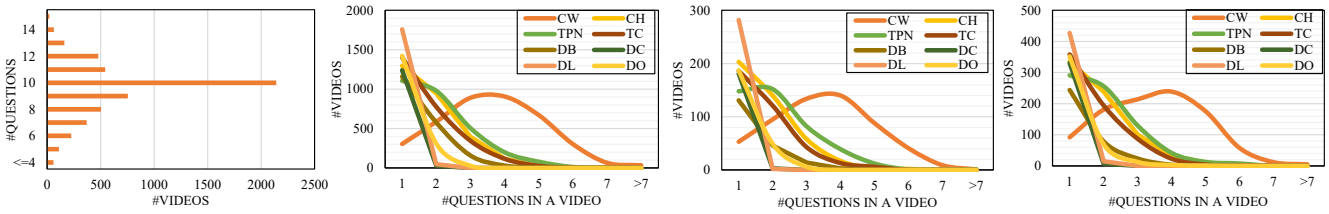
description. To better understand the dataset, we show some examples of the annotated question-answer pairs in Figure 6 (open-ended QA) and Figure 7 (multi-choice QA), from which we can also confirm that the answers to the questions can be visually inferred from the video content.

2. Analysis for open-ended QA

2.1. Evaluation

WUPS score is introduced in [5] to evaluate the generated answers. It is regarded as a soft version of accuracy that factors in synonyms and semantics. Specifically, given a predicted answer $P = \{p_1, p_2, ..., p_i, ...\}$ for a question whose reference answer (ground truth) is $R = \{r_1, r_2, ..., r_i, ...\}$, in which p_i and r_i are the ith tokens of the predicted and reference answers respectively, the WUPS score computes the similarity between two token sets as follows:

$$WUPS(P,R) = \min\{\prod_{p\in P}\max_{r\in R} WUP(p,r), \prod_{r\in R}\max_{p\in P} WUP(r,p)\} \times 100, \quad (1)$$



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

Figure 2: Distribution of the questions w.r.t videos. (a) the number of questions in most videos ranges from 4 to 14, and the vast majority of videos have 10 questions. In (b), (c) and (d), the distributions of questions are quite the same among the train/val/test data splits. For most of the videos, there are 2 to 6 causal questions that ask 'why' (CW); 1 to 3 causal questions that ask 'how' (CH); and 1 to 3 questions that ask temporal actions (either the previous/next (TPN) or the current (TC)). Aside from the causal and temporal questions, there are 1 or 2 descriptive questions asking either about binary-choice (DB), number-counting (DC), location (DL) or open-form (DO) in most of the videos.

(a) Train (b) Val (c) Test

Figure 3: Word clouds for frequent words in answers ('yes', 'no' and stop words are ignored.). The distributions vary little among train, val and test sets. This makes it possible to learn necessary information from training data for answering questions in val and test sets. Besides, the figures also show that there are various verbs in the answers in addition to nouns.

where $WUP(p, r)$ calculates the Wu-Parlmer similarity [2, 8] of two words based on their depth in the taxonomy [1, 6]: WUP(p, r) = 2*depth(lcs) / (depth(p) + depth(r)), in which lcs is the least common ancestor of the words p and r. If two words are semantically closer, they would be in same/nearer depths in the hierarchy and share more common ancestors, and thus get a higher WUP score.

2.2. Answer Decoders

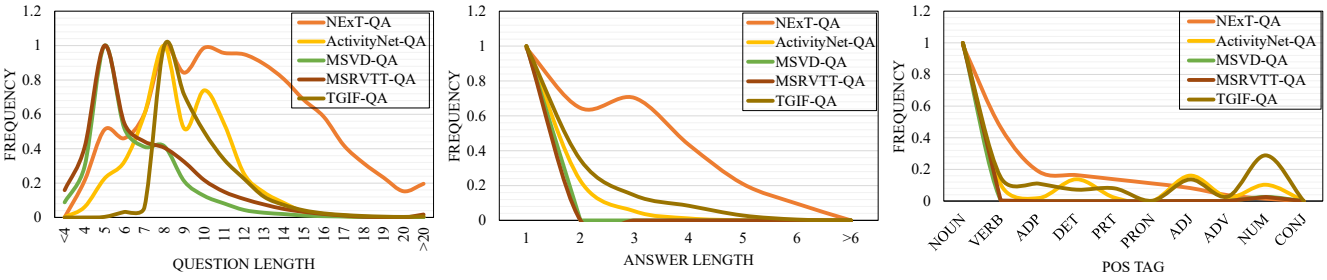
For answer decoders, we investigated several architectures as shown in Figure 5. The results in Table 1 are based on HGA [4] on validation set. From the results, we can see that the *naive* implementation performs the worst among other approaches. *naiveTrans* achieves the best result on descriptive question, but it still struggles on causal and temporal questions. *QnsAns* shows superior performance on temporal questions but is weak in answering causal questions featured in NExT-QA, and thus the overall WUPS score is still low. In contrast, *AttVid* and *AttQns* achieve better performances on causal questions and thus the better overall results. We attribute such strength of the attention-base decoders to the fact that they are better at determining which parts of the question or video should be attended for the answer. Since *AttQns* achieved the best overall result, we choose it as the default answer decoders for all other methods adapted from multi-choice QA.

Methods	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
Naive	12.95	15.04	45.65	20.44
NaiveTrans	12.74	15.15	**47.58**	20.77
QnsAns	12.50	**16.09**	46.82	20.77
AttVid	13.63	15.47	45.45	20.85
AttQns	**14.76**	14.90	46.60	**21.48**

Table 1: Results of different answer decoders.

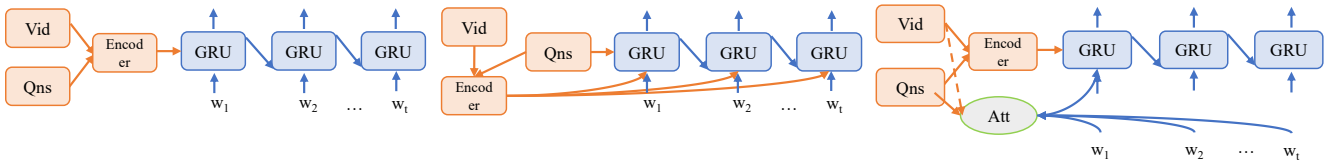
3. Results Analysis and Discussion

In Figure 6, we qualitatively analyze the models' performances on both multi-choice QA and open-ended QA tasks. According to the results, we make several main observations: 1) Answering causal and temporal questions requires much deeper understanding of both questions and videos that goes beyond a shallow description (refer to examples 1 to 6 *vs*. the last two), and the current models are still weak in this area. 2) When adapting models that are effective on multi-choice QA to open-ended QA, we find that they usually fail to correctly answer the questions, especially for causal and temporal questions (refer to examples 2, 3, 7 and 8). This suggest that the models either do not truly understand the video/questions, or struggle in generate the answers; both encumbering them from real-word application. 3) The models can correctly answer the questions to a cer-



(a) Question distribution w.r.t #words (b) Answer distribution w.r.t #words (c) Answer types

Figure 4: Detailed statistics of NExT-QA and popular VideoQA datasets.



(a) Naive & NaiveTrans (b) QnsAns (c) AttQns & AttVid

Figure 5: Architectures for answer decoders. (a) *Naive* [7], where the hidden state of the answer decoder is initialized with the output of the video-question (VQ) encoder. *NaiveTrans* is a variant of *Naive* by using transformation operation before RNNs. (b) *QnsAns*, in which the hidden state of the decoder is initialized with the last hidden state of the question encoder, and the output of the VQ encoder is concatenated with the input of the decoder at each time step. (c) An attention variant of the naive implementation, in which the attention can either be added to the question (*AttQns*) or the video (*AttVid*) [10] side.

tain extent in open-ended QA (refer to examples 2, 3 and 7), even if their WUPS scores are low with respect to the reference answers. 4) The predictions on some samples are semantically reasonable in answering the questions but are not relevant to the video contents (refer to examples 5, 6). This demonstrates that they can understand the questions, but are struggling in videos comprehension or language generation.

Overall, our NExT-QA dataset opens new challenges for deeper video understanding in that it benchmarks causal and temporal action reasoning, and is rich in object interactions in real-daily activities. Our extensive experiments show that existing models are weak in this area, which encourages future works for improvement. To facilitate research, the dataset and other related resources are available at https://github.com/doc-doc/NExT-QA.git

References

[1] Christiane Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012. 2

[2] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013. 2

[3] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 1

[4] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116, 2020. 2

[5] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NeurIPS*, pages 1682–1690, 2014. 1

[6] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

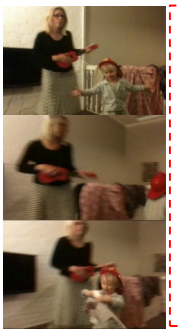
[7] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542, 2015. 3

[8] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138, 1994. 2

[9] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *MM*, pages 1645–1653. ACM, 2017. 1

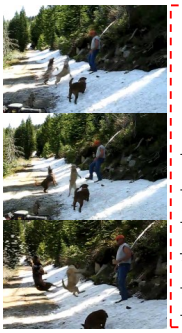
[10] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015. 3

[11] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. *AAAI*, 2019. 1



C: Why did the boy walk away from the woman after dancing for a while?
0. To take a paper. 1. Exploring the Christmas stuffs. 2. To take a picture. 3. Following boy. 4. Observe boy.

STVQA	2	to play with the. (6.72)
HME	1	to play the. (6.72)
HCRN	**0**	dance together. (3.17)
UATT	-	to the of. (0.00)
HGA	**0**	dancing action. (7.69)



C: Why do the dogs jump?
0. Bite the snow. **1.** Playing with kids. 2. Bite the toy. 3. Posing for camera. 4. Look at the cameraman.

STVQA	**0**	playing in snow. (4.94)
HME	**0**	playing. (2.61)
HCRN	1	playing. (2.61)
UATT	-	playing. (2.61)
HGA	**0**	chase the. (2.11)

T: What does the lady in singlet do as the lady next to her is sewing?
0. Record their process. 1. Pet gently. **2. Carve pumpkin**. 3. Help the baby walk . 4. Look at phone.

STVQA	3	take her. (1.19)
HME	3	hold her hand. (2.11)
HCRN	**2**	talk to the. (1.71)
UATT	-	null. (0.00)
HGA	1	inspect the baby. (13.64)

D: Where is this video taken?
0. Zoo. 1. Field. 2. Train station. 3. Classroom. **4. Beach.**

STVQA	**4**	beach. (100)
HME	**4**	outdoor. (0.00)
HCRN	**4**	beach. (100)
UATT	-	mountain. (72.73)
HGA	**4**	beach. (100)

C: How did the woman in yellow support the boy in blue at the start?
0. Caress baby**.** 1. Turn over his body. 2. Pull his finger. **3. Hold baby up.** 4. Rubbing baby s hair.

STVQA	**3**	hold her. (1.68)
HME	**3**	hold girl s hand. (8.79)
HCRN	**3**	hold her hand. (3.85)
UATT	-	hold her hand. (3.85)
HGA	**3**	hold onto harness. (5.71)

T: What did the lady in red do when the man in yellow first brought out the box?
0. Move body. 1. Take lollipop out. 2. Drags the box out. 3. Look at lady in white. **4. Point to lady in black.**

STVQA	2	talk. (0.26)
HME	2	put it. (0.12)
HCRN	2	him (0.00)
UATT	-	look at. (0.38)
HGA	2	adjust the in red (2.52)

T: How did the man react when the boy swung the stick toy towards him?
0. Catches it. 1. Stand up and point. 2. Kicks it. 3. Swing in the video. **4. Moved back.**

STVQA	0	pick up. (5.04)
HME	0	smile. (3.85)
HCRN	0	move the away. (13.33)
UATT	-	pick him up. (5.04)
HGA	0	laughing. (4.44)

D: What is the baby struggling to do in this video?
0. Smiling. 1. Get the toy giraffe. **2. Move body.** 3. Wipe the boy s mouth. 4. Swing himself.

STVQA	**2**	sleep. (3.85)
HME	**2**	sleeping. (3.85)
HCRN	**2**	lying. (3.85)
UATT	-	sleeping. (3.85)
HGA	**2**	sleep. (3.85)

Figure 6: Visualization of answer predictions for both multi-choice QA and open-ended QA. For multi-choice QA, the correct answers and predictions are highlighted in red. For open-ended QA, the WUPS score of each prediction is appended. 'null' means the methods fail to generate any effective words. (C: Causal. T: Temporal. D: Descriptive.)

C: Why was the woman in blue holding a red rope?
0.her snack.
1.protect her eyes.
2.guide her in rowing boat.
3.stylish.
4.hold the toy.

C: Why does the leopard jump up and down throughout the video?
0.excited
1.follow specific trail
2.to catch the toy
3.chase each other
4.photo taking

C: How does the woman in blue get the leopard jump?
0.wave toy around
1.blow whistle
2.walk around in front
3.clap both hands
4.snap fingers

C: Why is the man in yellow and the man in black carrying a shirtless man?
0.to see how to move the rod
1.stop him from falling
2.throw him into water
3.playing
4.shaking off snow

T: What does the shirtless man do after being thrown into water?
0.help the diver up
1.sit up
2.wriggle around
3.spit water out
4.fall down

T: What does the man in yellow and the one in black do after bring the shirtless man into water?
0.sit down
1.swim back towards the man
2.swim around
3.laugh
4.touch his head

C: Why did the woman bend down and run towards the baby?
0.to jump over him
1.exercises
2.entertain the baby
3.for fun
4.the dog bit her hand

T: What happened to the baby after the lady kicked towards baby from back?
0.walk away
1.run after it
2.make it move
3.flip onto back
4.fall to the ground

T: How did the lady in skirt react after the baby fall down?
0.hold baby and bend down more
1.laugh
2.hug woman's leg
3.jump
4.continue performance

C: Why did the boy jump onto the green disc at the start?
0.to break it.
1.slide down slope.
2.wide hole on the ground.
3.to keep afloat
4.for art.

T: How does the boy in black react while the boy on the green disc going down?
0.look at him
1.imitate the girl's movement
2.running point
3.stabilise himself
4.fall down

C: Why did the black jacket boy run after seeing the boy slide down?
0.retrieve the ball.
1.running after ball.
2.want to try slide.
3.check nobody behind.
4.let the dog chase.

D: Where is this happening?
0.car
1.hospital
2.forest
3.bowling alley
4.skate park

D: What is the animal shown in the video?
0.owl
1.rabbit
2.swan
3.sheep
4.dog

Figure 7: Examples of multi-choice QA. Each question has 5 options in which the correct answer is highlighted.