Supplementary Material for From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering

Jiangtong Li[1], Li Niu[1*], Liqing Zhang[1*]
[1] Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{keep_moving-Lee,ustcnewly,lqzhang}@sjtu.edu.cn

In this document, we provide additional materials to supplement our main submission. In Sec. 1, we introduce the detailed process of our dataset annotation. In Sec. 2, we give a thorough comparison between our Causal-VidQA dataset and other existing datasets. In Sec. 3, we provide a series of ablation experiments in our Causal-VidQA dataset.

1. Dataset Annotation

Our annotation can be divided into three stages: instance segmentation annotation, rational video selection, and question-answer annotation. As mentioned in the main submission, before the annotation, we have 546,882 unbroken videos longer than 9s from Kinetics-700 [10].

Instance Segmentation Annotation. In traditional video question-answering, the instances in questions and answers are usually described by text, such as *the man in blue* or *the second woman from the right side*. However, in complex scene, the text would call for multiple attributes to pinpoint the described instances, which may direct the core of question-answering task to video object grounding. Since reasoning is the center task of our Causal-VidQA dataset, we perform instance segmentation annotation to assign the semantic labels to different instances in the video clips.

Considering that manually annotating the whole Kinetics-700 dataset is too expensive and time-consuming, we choose to combine the image instance segmentation (IIS) and video instance segmentation (VIS) together to finish the instance segmentation annotation, where the VIS requires the instance segmentation mask of the first frame in each video clip to proceed the instance segmentation of the rest frames. During instance segmentation annotation, we first convert each video clip into multiple frames with frame rate as 10 FPS. Then we employ the IIS model and VIS model to extract the instance segmentation mask and label of each instance in these frames. Finally, we reorganize these frames into a consecutive video clip with the same frame rate, 10 FPS. For the IIS model, we choose the widely used Mask R-CNN [4] with ResNet-101 [5] as backbone pre-trained on Microsoft COCO [12] instance segmentation dataset to infer the instance segmentation mask and label of the first frame. For the VIS model, we choose CFBI [18] with ResNet-101 [5] pre-trained on Youtube-VOS [17] to infer the instance segmentation masks of the following frames. To ensure that enough interactions exist in video clips, we keep the video clips with more than two segmented instances. After the instance segmentation annotation, we have 310,934 video clips left.

Rational Video Selection. After the automatic instance segmentation annotation, many videos are still irrational for video reasoning. Our main concerns can be summarized as: 1. there are some wrongly annotated videos (*e.g.*, *wrong label*, *incomplete segmentation*); 2. some video clips are unqualified for reasoning (*e.g.*, *the scene is too simple or too complex*, *the video is too vague*). Therefore, we arrange 20 undergraduate students to select the rational video.

For the first concern, the correct object segmentation is defined as: 1. correct labels are assigned in more then 80% video frames; 2. the segmentation covers more than 70% of the objects in all video frames. Based on the above definitions, the selection rules are defined as: if the number of the correctly segmented objects is larger than one, the corresponding video clip is selected as rational video clip. Note that, the wrong segmentations in the rational video clip are annotated and deleted. For the second concern, we design the following irrational rules to select rational video: 1. the scene contains no human-object and human-human interaction (too simple); 2. the scene contains more than 20 human-object and human-human interactions (too complex); 3. the scene contains more than 10 persons (too complex); 4. the segmented objects is too vague to be recognized by annotators. A video clip that is judged as rational in both concerns will be regarded as rational video. After the rational video selection, we have 27,183 video clips left.

Question-Answer Annotation. In the question-answer annotation stage, we hire 40 undergraduate students as annotators and randomly divide them into 20 groups. In each

*Corresponding author.

Dataset	Visual Type	Visual Source	Annotation	Description	Explanation	Prediction	Counterfactual	#Video/Image	#QA	Video Length (s)
Motivation [15]	Image	MS COCO	Man	✓	✓	✓	×	10,191	-	-
VCR [22]	Image	Movie Clip	Man	✓	✓	✓	×	110,000	290,000	-
MovieQA [14]	Video	Movie Stories	Auto	✓	✓	×	×	548	21,406	200
TVQA [11]	Video	TV Show	Man	✓	✓	×	×	21,793	152,545	76
TGIF-QA [8]	Video	TGIF	Auto	✓	×	×	×	71,741	165,165	3
ActivityNet-QA [20]	Video	ActivityNet	Man	✓	✓	×	×	5,800	58,000	180
Social-IQ [21]	Video	YouTube	Man	✓	✓	×	×	1,250	7,500	60
CLEVRER [19]	Video	Game Engine	Man	✓	✓	✓	✓	20,000	305,280	5
V2C [2]	Video	MSR-VTT	Man	✓	✓	×	×	10,000	115,312	30
NExT-QA [16]	Video	YFCC-100M	Man	✓	✓	×	×	5,440	52,044	44
Causal-VidQA	Video	Kinetics-700	Man	✓	✓	✓	✓	26,900	107,600	9

Table 1. Comparison between Causal-VidQA and other visual understanding benchmarks on images and videos. Causal-VidQA a new challenging video question-answering benchmark for real-world reasoning with manual annotations. It introduce a wide range of reasoning tasks including scene description, evidence reasoning and commonsense reasoning with four types of questions (*i.e.* description, explanation, prediction and counterfactual.

group, one annotator (questioner) is in charge of questions and the other (answerer) is in charge of answers and reasons. The questioners are expected to report videos that are hard to pose effective questions and raise four high quality questions (description, explanation, prediction, and counterfactual) for each video. The answerers are expected to check the quality of the questions first, answer the good questions with proper reason if needed, and return the bad ones back to the corresponding questioners for improvement. To guarantee that these annotators are qualified, we train and evaluate them before the annotation. Considering that each video clip in Kinetics-700 has its own action label, we assign the video clips with the same action category to the same group to ensure that the questions and answers do not overlap. For the description and explanation question, the questioner is asked to select a question type from a drop-down menu to balance different question types. Note that, for the main instances in the video clips, we have their fine-grained segmentation masks and labels. Therefore, the questioners are required to propose about 75% questions based on the segmentation labels (*e.g.*, *[person_1]* and *[person_2]* in Figure 1 of the main submission) and 25% questions without using the segmentation labels. Concerning whether to use segmentation labels when annotating answers, we adopt the same rule as for annotating questions. Further, in the multi-choice generation, we also require that all distractors have same segmentation labels as corresponding video clips. Finally, all the questions with the answers and reasons are further reviewed by the authors of this work. After the question-answer annotation, we have 26,900 video clips from 666 action categories along with 107,600 questions, 107,600 answers, and 53,800 reasons.

2. Dataset Comparison

In Table 1, we give a detailed comparison between our dataset and other existing visual reasoning dataset from data source, question type, and statistics information. From data source, we use the largest human action video dataset, where the videos are from daily life. However, some of existing datasets either use images as the visual input [22] or use the videos from unrealistic scenes [11, 14, 19] Compared our Causal-VidQA with other existing datasets in terms of question type, we can find that descriptive questions exist in all datasets, since descriptive questions are the base of visual understanding. Whereas, the explanatory questions also exist in most datasets, because some datasets, like MovieQA [14], TVQA [11], TGIF-QA [8], ActivityNet-QA [20], regard the spatio-temporal questions as explanatory questions, where we only include the questions started with "why" and "how" as the explanatory questions. For the predictive and counterfactual questions, only three datasets contain these types of questions, but they either focus on image or focus on virtual game environment, which are far from real-world video reasoning.

3. Ablation Study

In this section, we discuss several settings that would potential affect the performance of existing methods, including video input, question type, question length, segmentation label feature, guidance of question and dataset split.

3.1. Video Input

In this section, we analyze the effect of video sampling rates and the video representation based on HGA [9] with on-the-shelf BERT as text representation. For the effect of video sampling rates, we change the number of segments in range of [0, 16] with the step as 4 to plot the performance variance in Figure 1 (a), where 0 indicates BlindQA [6]. From Figure 1 (a), we can find that as the number of segments changes, the accuracy for all types of questions first increase and then get stable, where 8 is enough

Method	Acc_D	Acc_E	Acc_P	Acc_C	Acc
CoMem [3]	57.27 (-6.81)	51.08 (-11.71)	26.20 (-5.21)	26.24 (-6.31)	40.20 (-6.97)
HME [1]	59.05 (-4.31)	52.24 (-9.21)	24.18 (-4.74)	25.04 (-5.89)	40.12 (-6.03)
HGA [9]	62.46 (-3.21)	60.97 (-2.54)	26.69 (-5.53)	34.02 (-0.12)	46.07 (-2.85)
B2A [13]	62.44 (-3.77)	59.89 (-3.03)	25.38 (-5.77)	33.25 (-1.91)	45.49 (-3.62)

Table 2. Results of training on four types of questions separately. + (*resp.*, -) in brackets represents the improvement (*resp.*, drop) compared with the results in Table 2 of the main submission. D, E, P, and C stand for descriptive, explanatory, predictive and counterfactual questions.

Method	Acc_D	Acc_E	Acc_P	Acc_C	Acc
CoMem [3]	60.06 (-4.02)	59.07 (-3.72)	29.99 (-1.42)	31.01 (-1.54)	45.03 (-2.68)
HME [1]	58.97 (-4.39)	57.07 (-4.38)	27.93 (-0.99)	30.16 (-0.77)	43.53 (-2.63)
HGA [9]	61.45 (-4.22)	59.19 (-4.32)	31.33 (-0.89)	33.37 (-0.91)	46.33 (-2.59)
B2A [13]	62.10 (-4.11)	58.93 (-3.99)	30.14 (-1.01)	34.19 (-0.97)	46.59 (-2.52)

Table 3. Results of training without segmentation label feature. + (*resp.*, -) in brackets represents the improvement (*resp.*, drop) compared with the results in Table 2 of the main submission. D, E, P, and C stand for descriptive, explanatory, predictive and counterfactual questions.

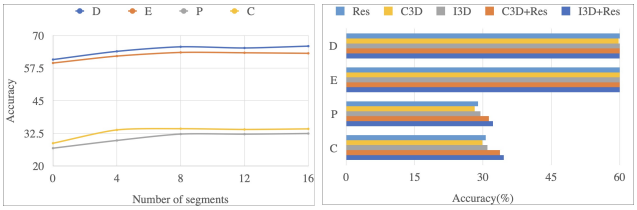

Figure 1. (a) Results with different numbers of clips. (b) Results with different video representations. D, E, P, and C stand for descriptive, explanatory, predictive and counterfactual questions.

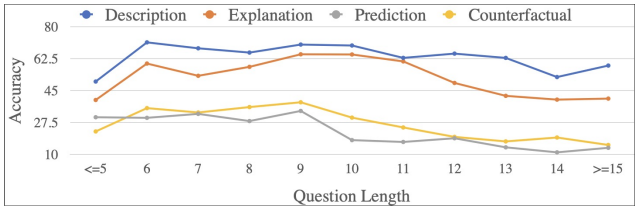

Figure 2. Results distribution along question length.

to achieve competitive results. Besides, comparing the improvement gained by different types of questions, we can find that the improvement for descriptive questions is the most significant, which reflects that the descriptive questions count more on visual representation. For the effect of video representation, we evaluate on five combinations of motion and appearance features, including motion and appearance (I3D+ResNet, C3D+ResNet), only motion (I3D, C3D), and only appearance (ResNet). The experiment results are shown on Figure 1 (b), where we can find that (1) the combination of motion and appearance feature always outperforms only using one of them, which indicates that the motion and appearance information can complement with each other; (2) the performance of I3D outperforms C3D with/without ResNet, which shows I3D as an inflated extension of 2D CNN can match better with ResNet in feature space. Similar observation was also found in [7, 16].

3.2. Question Type

These four types of questions emphasize different visual-language interaction, which may restrain the performance among each other. Therefore, we analyze the effect of question type based on CoMem [3], HME [1], HGA [9], and B2A [13] with on-the-shelf BERT as text representation. We train each method on four types of questions separately and show the performance of different methods on Table 2. Comparing the performance between Table 2 of the main submission and Table 2, we can find that all four types of questions suffer from obvious drop when training separately, which indicates that the description, evidence reasoning and commonsense reasoning capture similar information from language and video modality and are capable of promoting each other.

3.3. Question Length

In this section, we analyze the effect of question length based on HGA [9] with on-the-shelf BERT as text representation. As shown in Figure 2, the performance on descriptive question first increases and then gets stable as the length of question increases. However, for other three types of reasoning questions, the performance first increases and then drops severely as the length of question increases, which indicates that as the length of question increases, reasoning questions are hard to be understood by current methods and relations between language and video are hard to capture.

3.4. Segmentation Label Feature

To emphasize the reasoning instead of video object grounding, we annotate the segmentation labels in videos and use them during question-answering annotation. In this

Method	Acc_D	Acc_E	Acc_P	Acc_C	Acc
BlindQA [6]	20.33 (-40.45)	19.82 (-39.64)	4.22 (-22.59)	3.98 (-24.73)	12.09 (-31.85)
CoMem [3]	38.12 (-25.96)	34.73 (-28.06)	15.83 (-15.58)	11.62 (-20.93)	25.07 (-22.64)
HME [1]	37.62 (-25.74)	34.38 (-27.07)	15.91 (-13.01)	12.39 (-18.54)	25.08 (-21.08)
HGA [9]	38.05 (-27.62)	34.72 (-28.79)	15.60 (-16.62)	12.00 (-22.28)	25.09 (-23.83)
B2A [13]	38.56 (-27.65)	34.86 (-28.06)	15.72 (-15.43)	12.47 (-22.69)	25.40 (-23.71)

Table 4. Results of training and inference without questions. + (*resp.*, -) in brackets represents the improvement (*resp.*, drop) compared with the results in Table 2 of the main submission. D, E, P, and C stand for descriptive, explanatory, predictive and counterfactual questions.

Method	Acc_D	Acc_E	Acc_P	Acc_C	Acc
BlindQA [6]	62.44 (+1.66)	61.18 (+2.35)	29.52 (+2.71)	30.96 (+2.25)	46.03 (+2.08)
CoMem [3]	69.40 (+5.32)	69.91 (+7.12)	40.23 (+8.82)	41.10 (+8.55)	55.16 (+7.45)
HME [1]	68.65 (+5.29)	69.83 (+8.38)	38.26 (+9.34)	39.94 (+9.01)	54.17 (+8.01)
HGA [9]	72.40 (+6.73)	72.63 (+9.12)	41.88 (+9.66)	44.05 (+9.77)	57.74 (+8.82)
B2A [13]	72.65 (+6.44)	71.90 (+8.98)	40.86 (+9.71)	44.47 (+9.31)	57.72 (+8.61)

Table 5. Results of training on randomly split dataset. + (*resp.*, -) in brackets represents the improvement (*resp.*, drop) compared with the results in Table 2 of the main submission. D, E, P, and C stand for descriptive, explanatory, predictive and counterfactual questions.

section, we analyze the effect of segmentation label feature in text representation, based on CoMem [3], HME [1], HGA [9], and B2A [13] with on-the-shelf BERT as text representation. We train each method without using the segmentation label feature during text representation and the performance is shown in Table 3. For Table 3, we can find that segmentation label feature affects the model performance in a sense. Comparing these four types of questions, the descriptive questions is affected most and drops about 3-5 % among different methods, however, the other three types of questions only drop about 1-3% among different methods. We suspect that casual relations between video and language, and reasoning process cannot be well modeled by existing methods, therefore, removing the segmentation label feature does not affect these questions much.

3.5. Video to Answer

In this section, we study the effect of questions in our Causal-VidQA dataset, *i.e.*, what will happen if the questions do not exist? To this end, we conduct experiment without questions during training and inference based on CoMem [3], HME [1], HGA [9], and B2A [13] with on-the-shelf BERT as text representation. Experiment results of training and inference without questions are shown in Table 4. From the results, we have the following observations, 1) the performance of BlindQA is close to random selection, where the accuracy for descriptive and explanatory questions is around 20% and the accuracy for predictive and counterfactual questions is around 4%; 2) the performance on these existing method is similar. Based on the above two observations, on the one hand, the answers and reasons cannot be deduced simply based on semantic information among answers; on the other hand, without the guidance of question, the relations between video and answers are simplified and the gaps among existing methods are also erased.

3.6. Dataset Split

In Sec. 3.4 of the main submission, we split the dataset by action categories into training/validation/testing with a ratio of 7:1:2. In this section, we study the effect of random splitting the dataset into training/validation/testing set based on BlindQA [6], CoMem [3], HME [1], HGA [9], and B2A [13] with on-the-shelf BERT as text representation. We re-split dataset randomly with the same ratio as current split, and then extent the dataset to multi-choice version. Experiment results on the re-split dataset are shown Table 5. From Table 5, we can find that 1) the performance increases by a large margin for reasoning questions and remains for descriptive questions in methods except BlindQA; 2) the performance does not change much in BlindQA. This phenomenon proves that video scenes from same action category are similar, which would make the reasoning by simply correlating action categories with answers and reasons, however, for current split in main submission, answers and reasons are not simply correlated with action categories.

References

[1] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR 2019*, pages 1999–2007, 2019. 3, 4

[2] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. In *EMNLP 2020*, pages 840–860, 2020. 2

[3] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR 2018*, pages 6576–6585, 2018. 3, 4

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV 2017*, pages 2980–2988, 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778, 2016. 1

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 2, 4

[7] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *Int. J. Comput. Vis.*, 127(10):1385–1412, 2019. 3

[8] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR 2017*, pages 1359–1367, 2017. 2

[9] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI 2020*, pages 11109–11116, 2020. 2, 3, 4

[10] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, 2017. 1

[11] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: localized, compositional video question answering. In *EMNLP 2018*, pages 1369–1379, 2018. 2

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV 2014*, pages 740–755, 2014. 1

[13] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *CVPR 2021*, pages 15526–15535, 2021. 3, 4

[14] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR 2016*, pages 4631–4640, 2016. 2

[15] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *CVPR 2016*, pages 2997–3005, 2016. 2

[16] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR 2021*, pages 9777–9786, 2021. 2, 3

[17] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott Cohen, and Thomas S. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV 2018*, pages 603–619, 2018. 1

[18] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV 2020*, pages 332–348, 2020. 1

[19] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR 2020*, 2020. 2

[20] 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. In *AAAI 2019*, pages 9127–9134, 2019. 2

[21] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR 2019*, pages 8807–8817, 2019. 2

[22] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR 2019*, pages 6720–6731, 2019. 2