DREU Final Report

Madeleine Grunde-McLaughlin
University of Pennsylvania
Philadelphia, Pennsylvania
mgrund@sas.upenn.edu

Maneesh Agrawala
Stanford University
Palo Alto
maneesh@cs.stanford.edu

1. Introduction

mgm: Looking into the actual lengths, a lot esp in test set are shorter than 30 seconds. may want to find average

mgm: High level comment: is this too much or too little detail for an intro? anything else we should include?

mgm: Is this a useful thing to include: Few questions in existing benchmarks focus specifically on object manipulation through videos.

The ability to perceive and reason about human activity has been a long standing goal for Computer Vision. Research in Cognitive Science and Neuroscience suggests that humans decompose actions into a hierarchy of parts [29]. Therefore, in order to reason about activities, models should be capable of spatial and temporal reasoning over a subject's changing relationships with objects [9]. One task developed to measure visual reasoning abilities is Visual Question Answering (VQA), in which a model answers questions about a visual input.

A variety of benchmarks have been created to test a model's abilities at the VQA task. ImageQA benchmarks take images as input [11, 7, 2, 30, 6, 17, 32, 14] and VideoQA benchmarks take videos as input [26, 20, 8, 15, 27, 22, 31, 28]. Since most VQA benchmarks are image-based, they can test reasoning on spatial relationships, object attributes, and common sense understanding, but they cannot test reasoning over temporal relationships or activities.

For this reason, a growing number of VideoQA benchmarks have been release, but current benchmarks exhibit some weaknesses. First, many involve short videos ($<$ 10 seconds) [8, 15, 27, 22], or contain fewer than 200k questions mgm: what counts as small? [26, 20, 15, 27, 8, 28, 31].

Second, current benchmarks require at most two reasoning steps to answer. For example, the question "What happened to the woman before playing violin?" requires first finding when the woman played violin, then looking at what happened before then. [28]). Existing benchmarks also do not use logical or comparative operators in their questions, which leaves out interesting and relevant ideas (e.g. "Which activity did they do for longer?"). Leaving out questions that require more varied reasoning steps creates weaker benchmarks since many VideoQA models get stuck when faced with questions that require multiple reasoning steps [5].

Third, several of these benchmarks integrate visual cues with dialogue and plot summaries [20, 26, 15]. Their analysis found that question answering models depend more heavily on the dialogue input than on the visual input, reducing these benchmarks' effectiveness at measuring visual temporal reasoning [26, 20].

Fourth, after the release of the first large ImageQA datasets, investigations found that their answer distributions were heavily skewed. This bias towards very common answers effectively reduces dependence on visual input and inflates accuracy scores [6, 7]. Some ImageQA datasets have worked to balance answer distributions, but to our knowledge no VideoQA datasets have systematically balanced answer distributions [6, 7].

Finally, existing benchmarks have accuracy measurements only for the overall dataset and sometimes for each question type (e.g. Repetition Count, Repeating Action, State Transition, and Frame-based questions in [8]). Reasoning about videos requires multi-step reasoning, comparative and counting operators, flexibility with video length, appearance feature understanding, and fine and coarse motion understanding. A wider metrics suite would improve measurement of models' relative strengths and weaknesses across these many categories.

We propose a VideoQA benchmark that addresses these concerns. Our benchmark is automatically generated in a pipeline inspired by [7] to combine question templates with Charades action annotations and Action Genome's spatio-temporal scene graphs. This process combines a variety of linguistic structures with the content of over mgm: 10,000 videos, producing mgm: add here questions. These questions have a wide variety of lengths, complexities, and structures. With this generation pipeline, we have tight control over the contents of each question, allowing us to balance the answer distribution and create a wide suite of metrics.

We have made much progress within the DREU timeline,

and we plan to continue the project through the fall. This final report discusses the questions for 500 videos, mgm: ? template and ? question-answer pairs By the end of our contributions will be: 1) a large VideoQA dataset of question answer pairs requiring multi-step reasoning and 2) a wide suite of metrics to thoroughly evaluate current models.

2. Prior Work

2.1. ImageQA

At first, the Visual Question Answering task was mostly restricted to ImageQA. A wide variety of benchmarking datasets were created with the hopes of analyzing the reasoning abilities of ImageQA models [11, 7, 2, 30, 6, 17, 32, 14]. Benchmarks vary in input, from synthetic datasets [11], to cartoons [2], to charts [15], to real-world images [7, 17, 32, 6, 30, 2]. They also vary in the type of questions asked, from the 7W's (who, what, where, when, which, why, how) [32], to commonsense reasoning [30], to compositional reasoning [11, 7], to spatial localization [32, 17, 7]. These benchmarks facilitated the development of many models made to tackle these challenges by measuring their spatial reasoning abilities [11, 7, 17].

However, many of these datasets contained real world priors exacerbated by human annotation bias, resulting in inflated accuracy scores and a lack of understanding of the actual reasoning abilities of these models [6, 7]. For example, in the VQA1.0 dataset, 41% of answers for questions starting with "What sport is..." were "Tennis". These types of priors meant that models could answer over 50% of the questions correctly without considering the visual input [6]. Once these issues came to light, some new benchmarks attempted to mitigate these biases. VQA2.0 took many of the questions from VQA1.0 and added a similar picture leading to a different answer. This procedure helped, but was only applied to 71% of questions due to annotation difficulties. Models measured on VQA2.0 could still answer 67% of binary questions and 27% of open questions correctly without seeing visual input [7]. The GQA dataset addressed these biases by creating equal numbers of binary yes/no questions per category, and smoothing the answer distribution in open ended questions. This balancing of the answer distribution retains but reduces the power of real world priors [7]. Our benchmark is similar to GQA in that it mitigates biases by smoothing answer distributions. However, it will move beyond just spatial reasoning and into the temporal domain.

2.2. VideoQA

mgm: how to include new information here without being too repetitive with the introduction?

A growing interest in Visual Question Answering on videos has lead to the development of some VideoQA benchmarks as well [26, 20, 8, 15, 27, 22, 31, 28]. Some

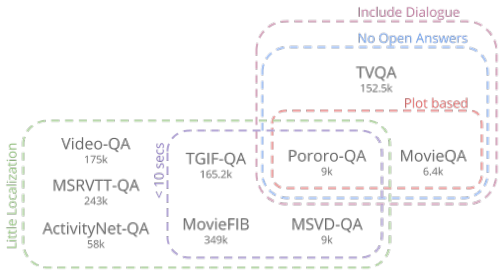

Figure 1. The shortcomings of current VideoQA models.

incorporate both visual input and textual input from a TV show or movie [26, 20, 15], several of which ask plot-based, not exclusively vision based, questions [26, 15]. Models trained on these datasets tend to depend heavily on dialogue for long-term temporal reasoning [26, 20].

Vision-only benchmarks exclusively ask questions that can be answered from a frame image of the video, questions that apply to the entire video, or questions that ask "what happened before/after/while <action>?" mgm: is there a better way to explain this type of question?. They all, except [28, 27], refer to videos of less than 10 seconds long. Some have automatically generated questions from image descriptions [27, 31], some let people choose from drop down menus [8], and others use humans to create questions [28, 26, 8, 20]. The largest dataset with solely human generated questions is [20] with 152.5K question answer pairs, and the largest dataset with solely automatically generated questions is [22] with 349K question answer pairs. Our dataset is purely vision based, works on videos of mgm: 30-90 seconds long, generates mgm: add here question answer pairs and evaluates complex and multi-step temporal reasoning.

2.3. Compositional Reasoning

As questions become more complex, they often require of multiple reasoning steps in order to answer them. For example, the question "Was the person running or sitting for longer?" first requires finding the start and end of when the person was running and sitting, subtracting the start from the end, then comparing the lengths. A constrained set of logical steps (e.g. action localization, subtraction, comparison, etc) can be used as building blocks that are reordered to respond to a wide variety of different questions. This multi-step reasoning is called compositional reasoning because the overall understanding is composed of a series of smaller reasoning steps. Humans use compositional reasoning to learn quickly and generalize to novel combinations of ideas [25, 18, 23].

Compositional questions have been used by ImageQA benchmarks to rigorously test models. Compositional questions tend to be longer, more complex, and use a wider

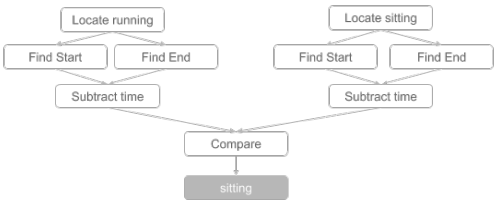

Figure 2. The substeps required to answer the question "Was the person running or sitting for longer?".

Figure 3. Various types of temporal reasoning.

variety of vocabulary. Furthermore, they can better test a model's reasoning ability by both requiring multiple steps of reasoning to answer the questions and by allowing for datasets to test generalizaton to novel compositions [11, 7, 18]. For example, on the CLEVR dataset, the training questions could include the phrases "right of cube" and "behind sphere", but not "right of sphere" as a combination. Testing on this novel combination "right of sphere" in the test set tests a model's ability to generalize [18, 11]. Many questions relevant to reasoning over videos require multi-step reasoning. A simple example covered by current datasets involves first localizing in time, then reasoning about that specific time. For example, this question from [8], "What does the bear on the right do after sitting?", requires finding when the bear stops sitting, then determining his activity after that. mgm: Is this actually 3 b/c have to look at the bear? Another way to reason compositionally is to refer to subjects indirectly, as the ImageQA benchmark [7] does with "What color is the food on the red object to the left of the girl?".

Some ImageQA benchmarks use compositional reasoning extensively [11, 7], but no VideoQA datasets go beyond two steps of compositional reasoning. Although these logical building block steps have been defined for ImageQA [4], there are no compositional reasoning steps defined for temporal reasoning. Current models have struggled with multi-step reasoning, motivating a benchmark like ours that specifically explores compositional reasoning [5].

2.4. Scene Graphs

Scene graphs are a symbolic representation of an image [17]. The graph consists of nodes representing the objects in the image and edges representing relationships between these nodes. For each object node there are associated attributes. For example, an image with a man wearing white shorts would have an object node for "shorts" with "white" as an attribute and an object node for "man". These object nodes would be connected by an edge representing the relationship "wearing". Creating this symbolic representation of an image reduces it to its semantic parts. Using this representation has improved performance on many visual tasks such as visual question answering [12], visual question answer generation [7], relationship modeling [16], image captioning [1], image generation [10, 3], and image retrieval [3, 13].

Inspired by Visual Genome's scene graphs, [9] annotated spatio-temporal scene graphs to create Action Genome. Spatio-temporal scene graphs consist of nodes representing objects and edges representing relationships between them mgm: (add in that these relationships are 3 categories?). Each of these objects are associated with a specific frame in the video. The resulting effect is that a subject's relationships with the objects changes over time. Our project will use these spatial-temporal scene graphs to generate questions about videos.

3. Methods

3.1. Types of Temporal Reasoning

To create our benchmark, we first established what types of temporal reasoning to explore. As mentioned previously, other vision-based datasets have looked exclusively at questions that could be answered with one frame, questions that refer to the entire video, and questions that ask "what happened" before, after, or while an event was occuring [26, 20, 8, 15, 27, 22, 31, 28].

However, there are more types of temporal reasoning yet to be explored and tested on models. These new types include action sequencing, reasoning over the length of time activities occurred, reasoning with comparative and logical operators, and reasoning about when actions precede, succeed, or coexist with one another. Other questions could more deeply explore a subject's changing relationship with objects over time.

Our benchmark requires all the above types of reasoning to succeed. Future datasets may be able to expand even more by including sequences, bounding boxes, and time identifiers (e.g. she started running at 10.5 seconds) as answers.

3.2. Our Generation Pipeline

Beyond exploring a wider variety of temporal concepts, an ideal benchmark would 1) be large, 2) mitigate biases in

the answer distribution, 3) allow for novel compositions and 4) include a large suite of metrics. In order to achieve these goals, we automatically generate questions using a pipeline inspired by [7].

The question generation pipeline consists of two parts: 1) augmenting spatio-temporal scene graphs to create detailed video representations and 2) building question templates with spaces for video-specific content. The first part of our pipeline uses dataset annotations from Action Genome and Charades [9, 24] on Charades videos. These mgm: 30-90 second videos are filmed by non-professionals who act out everyday activities using 36 objects. Charades annotates the times actions occur, and Action Genome annotates attention, spatial, and contact relationships on a sample of frames.

To make these scene graphs applicable for question answering, we first combine the annotations from both datasets into one data structure. We then created object and relationship nodes for action segmentation, adding in charades actions as a relationship category. We then supplement current annotations with their logical extensions. For example, if a person is carrying something, they are also holding it and touching it. With these adjustments, we have a filled out spatio-detailed spatio-temporal scene graph combining two datasets worth of annotations into a structure that can be fed into the question templates. An important limitation to note is that although these scene graphs are detailed, our questions depend on their annotations, so any errors in those datasets affects ours.

The second part of our pipeline consists of question templates that each have a natural language sentence with tags representing different elements categories in the video, such as in the question "What were they <contact relationship>?". Each template also contains information about the question's structural and semantic contents.

To generate a question-answer pair, our system combines a spatio-temporal scene graph with these templates by replacing the tags with elements of the corresponding types. For a video where the person is holding a blanket while sitting on a chair, our pipeline would create both questions "What were they holding?" and "What were they sitting on?". Given the scene graph information, we then automatically determine the answers "blanket" and "chair" respectively.

These tags can be filled with both direct and indirect references. For example, an <object> tag's direct reference would be "blanket" but an indirect reference would be "the object they were holding". We ensure that there are no questions that give away the answers (e.g. "Were they holding the object they were holding?"). Indirect references can also be recursive. The object tag listed above could also be replaced with "the object they were holding after closing the door" or "the object they were holding after the shortest action". Indirect references require locating the referred object, action, relationship, or time in video before reasoning about the question as a whole, leading to complex questions that require multiple reasoning steps.

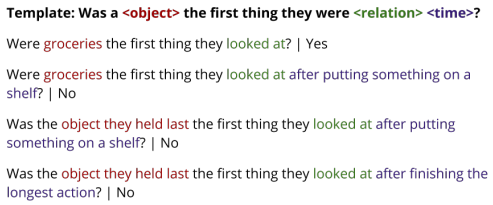

Figure 4. Various types of temporal reasoning.

This procedure to generate questions allows us to work towards the qualities of an ideal benchmark.

3.3. Is large

Human annotation of question-answer pairs for video is time consuming and expensive. Automatically generating questions allows us to create a much larger dataset of mgm: add here. Automatically generated datasets have previously been criticized for lacking diversity [28]. We have mgm: add templates asking a wide variety of questions. Each template is also associated with multiple natural language versions of the same question. Indirect reference tags expand complexity even further.

mgm: make some sort of graphic comparing size?

3.4. Mitigates biases in answer distribution

Many ImageQA benchmarks do not accurately determine a model's reasoning ability because skews in their answer distributions reduce dependence on visual input and increase reliance on the relative prevalence of different answers in the data [7, 6]. Although there has not been an in depth analysis of skewed answer distributions in Video QA biases, our dataset, if left unbalanced, would be skewed. For example, a person is covered by a blanket 6721 times but standing on a blanket only 32 times. Consequently, the answer to the question "What were they doing to the blanket?" is much more likely to be "covered by" rather than "standing on". To avoid inflated accuracy scores, we want to retain the same distribution (i.e. still have more instances where the answer is "covered by"), but smooth the answer distribution to make the differences less extreme. We balance the answer distribution for each category using the same process as [7] to downsample highly popular answers to shift probability into the tail of the distribution. mgm: Talk about the process in detail for doing that? Global vs local variables. mgm: Include graphs. Our question generation pipeline allows for this balancing because we have the structural and semantic information associated with each

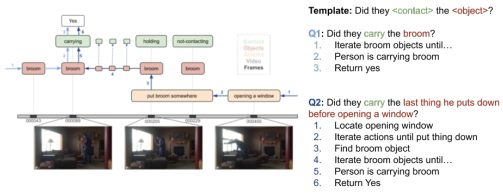

Figure 5. Various types of temporal reasoning.

template to split the questions into relevant categories before balancing.

3.5. Tests novel composition

An important part of human reasoning is the ability to generalize concepts of the same category to novel compositions. For example, in the training set the model may see the phrases "before running" and "the action they did the longest". Then, in the test set, we can test if they understand the idea of "before the action they did the longest". We are able to create such a training and testing split to judge these novel compositions because we have knowledge of what temporal localization, logical reasoning, and indirect tags are in each question. mgm: Potentially VG slide 18 for how to solve them?

3.6. Has a suite of new metrics

Finally, we want to be able to complexly analyze a model's various abilities across different types of temporal and logical reasoning. We can achieve better measurements by choosing different types of questions from the training and test sets based on qualities of the questions. We are able to do this because our pipeline gives us control over the content of all the questions. We have not accomplished this step thus far in the 10 weeks of the DREU program, however we plan to incorporate it by the end of the summer.

4. Results

A this point, we have completed mgm: this many templates and generated them for 500 videos, creating mgm: this many question-answer pairs. Our final Video QA dataset consists of three parts: the Charades videos, without any annotations or the corresponding spatio-temporal scene graphs. the questions, the answers.

In other words, Our Benchmark = $\{V, Q, A\}$, where every $v_i \in V$ corresponds to a list of questions $q_i \in Q$ and a list of corresponding answers $a_i \in A$. For every question $q_i j \in q_i$ there is a corresponding answer $a_i j \in q_i$. Each $q_i j$ is associated with binary flags determining if it is included in each metric.

We have implemented most of this pipeline, besides the metrics which we plan to implement soon.

Break down the dataset into better stats.

Pre and post balancing

Do things like average length of question vs average length of other questions in other datasets

Example questions

Note that final dataset does not have access to the Charades and AG annotations

5. Future work

To complete this project, we plan to expand upon our current progress in three ways. First, we will create a wider variety of templates, both in the subject of what they ask and in the natural language question. Second, we will develop a suite of metrics. Third, we will evaluate existing VideoQA models on our dataset [19, 5, 21]

We want our benchmark to be able to provide a complex analysis of the abilities of VideoQA models. Therefore, we will create a suite of metrics, each of which measures a different aspect of performance, as specified in Table mgm: how to refer to this?. Content Category metrics ask how well the model performs overall and on different question and answer categories. Generalization metrics ask if given a small portion of basic concepts, can the model generalize to more complex questions. Answer legitimacy metrics ask if the model has a consistent and plausible understanding of the contents of the video. Our suite of metrics will provide insight on these questions and a more detailed understanding of the model's strengths and weaknesses.

6. Conclusion

Our project contributes a dataset of question-answer pairs for video, a pipeline for creating question answer pairs from spatio-temporal scene graphs, and a suite of metrics on which to measure performance. Our benchmark will facilitate the creation of VideoQA models by providing a challenging task.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569, 2019.

[4] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Confer-*

Table 1. Suite of Metrics

Content Category	
All	Accuracy on all questions
Question Type	Accuracy on each question category (e.g. Counting, Length, First, ...)
Answer Type	Accuracy on each answer category (e.g. binary, and open)
Generalization	
Video length	train on videos of length $<$ 30 seconds. Test on videos of length $\geq$ 30 seconds
Actions	train on videos with $<$ 5 actions. Test on videos with $\geq$ 5 actions
Compositional Steps	train on videos with $<$ 6 steps. Test on videos$\geq$ 6 actions
Novel Compositions	TODO
Indirect ref Consistency	TODO
% training data	train on only 1, 5, 10 and 20 percent of training data
Answer Legitimacy	
TODO	Something where sequencing is consistent
Consistent	If answers a question correctly, answers all logical entailments correctly
Validity	Answer type is of correct genre (e.g. object, relation, action, count, yes/no)
Plausibility	Answer exists in distribution for that question
Distribution	Predicted answers follow same distribution as ground truth

ence on Human Factors in Computing Systems, pages 4061–4064, 2015.

[5] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007, 2019.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[8] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017.

[9] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[10] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[11] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.

[13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[14] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[15] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.

[16] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6867–6876, 2018.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[18] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882, 2018.

[19] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9972–9981, 2020.

[20] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[21] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019.

[22] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017.

[23] Eric Schulz, Josh Tenenbaum, David K Duvenaud, Maarten Speekenbrink, and Samuel J Gershman. Probing the compositionality of intuitive functions. In *Advances in neural information processing systems*, pages 3729–3737, 2016.

[24] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[25] Jun Tani. Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE*, 102(4):586–605, 2014.

[26] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[27] 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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[28] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[29] J Zacks, B Tversky, and G Iyer. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130:201–213, 2018.

[30] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[31] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *arXiv preprint arXiv:1611.04021*, 2016.

[32] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.