DREU Final Report: Compositional Spatio-Temporal Reasoning

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

Maneesh Agrawala
Stanford University
maneesh@cs.stanford.edu

Abstract

Existing Video Question Answering benchmarks are limited in the spatio-temporal reasoning skills they measure. We build a pipeline to combine Action Genome's spatio-temporal scene graph annotations with predefined templates to create a new benchmark AGQA. AGQA's question templates create questions that test a wide variety of spatio-temporal reasoning skills, including action sequencing, complex temporal localization, and compositional reasoning. Along with these questions, AGQA contributes an extensive suite of metrics to measure model's accuracy at different reasoning challenges, ability to generalize to new combinations of ideas, and consistency of an underlying worldview. We plan to test existing models on this new benchmark to measure their relative strengths and videos.

1. Introduction

The ability to perceive and reason about the world's actors, actions, objects, and relationships has been a longstanding goal for Computer Vision. Formally, we can benchmark progress towards this goal using tasks like question answering, in which a model's reasoning skills are evaluated by it's ability to answer questions about visual stimuli. A variety of benchmarks have been created to test a model's capabilities at answering questions about images [11, 7, 2, 32, 6, 17, 34, 14] and about videos [27, 20, 8, 15, 29, 22, 33, 30]. Ideally, models trained on these benchmarks should be capable of reasoning over both spatial relationships between objects [17, **?**] and temporal ordering of actions [31, 9]. Unfortunately, since most question-answering benchmarks operate over images, they are limited to only testing spatial relationships (e.g. "What is on top of the table?") [7, 17, 2]. The few existing video-only benchmarks have questions with simple temporal logic (e.g. "What does the bear on right do after sitting?") [8, 29, 22, 33, 30]. However, to answer questions that require models to jointly compose spatial and temporal reasoning over multiple steps (e.g."What did they do to the last object they put down before opening the window?"), we need newer benchmarks and a new class of models.

We introduce the video question answering benchmark Action Genome Question Answering (AGQA) and use it to evaluate models on compositional spatio-temporal reasoning. AGQA measures a wider variety of spatio-temporal reasoning skills than existing benchmarks. An ideal benchmark to measure spatio-temporal reasoning abilities should (1) be free of human annotation bias, (2) use videos of various lengths as input, (3) contain a large number of questions, and (4) provide not just a single accuracy score but a suite of new metrics that test various spatio-temporal capabilities. To achieve these goals, we developed an automatic approach that converts Action Genome's spatio-temporal scene graphs, containing objects, relationships, actors, and actions, into compositional questions [9].

Existing question-answering benchmarks are limited in the diversity of their questions. Since most benchmarks are image-based, they only test a model's reasoning over spatial relationships [11, 7, 2, 6, 17, 34], object attributes [11, 7, 2, 6, 17], and common sense understanding [32, 2, 17]. These benchmarks are unable to test reasoning over temporal relationships or activities. The questions asked by VideoQA benchmarks that do not include extra textual information like dialogue can be answered from a single frame of the video, apply to the entire video with no temporal localization, or ask "what happened before/after/while <action>?" [8, 29, 22, 33, 30]. Temporal localization refers to using the phrase "before/after/while <action>" to localize a relevant time in the video over which to reason. Beyond the three types of questions listed above, models should be able to perform more complex temporal localization, follow the changing state of objects over time, sequence actions, compare temporal qualities like length among different parts of the video (e.g. "Which activity did they do the longest?"), and generalize to new domains [18, 28]. Creating this diverse range of questions has previously been difficult. Questions automatically generated from video captions often lack diversity in structure [30, 8]. Human annotated datasets are too expensive to get a large enough sample to include a wide variety of categories [33, 30] The community does not know how existing models perform on more complex temporal reasoning and generalization to new content because existing benchmarks do not have questions requiring these abilities. By recursively generating questions with direct references, indirect references, and temporal localizations, our question generation pipeline creates a diverse set of questions that can be trained and tested all together or separated by type.

An ideal videoQA benchmark measures the ability to reason over visual stimuli instead of depending on the statistics of answer distributions [28]. However, many question answering datasets have a bias towards common answers that reduce dependence on visual input and inflate accuracy scores [6, 7]. Since we have question content information from our generation pipeline, we are able to smooth the answer distribution among different question types.

An ideal VideoQA benchmark would have videos with a variety of video lengths. Some VideoQA datasets have a variety of clip lengths [30, 29], but most VideoQA datasets use videos of less than 10 seconds long [8, 15, 29, 22, 33, 30]. Some models are better suited for longer or shorter videos, but an ideal model would be able to reason over a variety of lengths [23, 19]. Although it is expensive to annotate longer clips, we use two already annotated datasets on videos from 2.33 to 194.33 seconds long [25, 9].

An ideal VideoQA benchmark should be large. All current VideoQA sets have less than 350K question answer pairs [8, 15, 29, 22, 33, 30, 20, 27]. A larger training set can support a wider variety of questions while avoiding underfitting, and a larger testing set can be split into a wider variety of subsets [22]. On only 250 videos, our process generated over 8 million questions of diverse structures. Our final dataset will generate questions on over 9.8K videos.

Finally, an ideal VideoQA benchmark provides a wide range of metrics to judge a model's relative strengths and weaknesses. Most existing benchmarks measure accuracy on the overall dataset [22, 8, 29, 22, 33, 30]. Some also measure accuracy on categories of questions [8, 29, 30], or on questions without visual input [33]. For example [29] splits questions by the first word (”what”, ”where”, ”who”, ”how, ”when”). Reasoning complexly about videos requires multi-step reasoning, comparative and counting operators, flexibility with video length, appearance feature understanding, and fine and coarse motion understanding [19, 5]. The accuracy scores of existing benchmarks do not provide the granularity needed to measure success in these different aspects of video understanding [5, 19]. Furthermore, no existing benchmarks we know of measures a model's ability to generalize to new information. AGQA provides a suite of new metrics measuring accuracy by category, the ability to generalize, and the extent to which a model's answers reflect a consistent and realistic understanding of the world.

We have made much progress within the DREU timeline working towards creating a benchmark that has these qualities, and we plan to continue the project through the fall. This final report discusses the questions for 250 videos, generating over 8 million question and answer pairs. By the end of our project, our contributions will be: 1) a large VideoQA dataset of question answer pairs requiring multistep 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 benchmark datasets were created with the hopes of analyzing the reasoning abilities of ImageQA models [11, 7, 2, 32, 6, 17, 34, 14]. Benchmarks vary in input, from synthetic datasets [11], to cartoons [2], to charts [15], to real-world images [7, 17, 34, 6, 32, 2]. They also vary in the type of questions asked, from descriptive 7W's (who, what, where, when, which, why, how) [34], to commonsense reasoning [32], to compositional reasoning [11, 7], to spatial localization [34, 17, 7]. These benchmarks facilitated the development of many models made to tackle these challenges by measuring their spatial reasoning abilities [11, 7, 17, 28].

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 answer questions correctly without seeing visual input [7]. The GQA dataset addressed these biases by smoothing the answer distribution by question category. This balancing of the answer distribution retains but reduces the power of real world priors [7]. AGQA 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

A growing interest in Visual Question Answering has also lead to the development of some VideoQA benchmarks [27, 20, 8, 15, 29, 22, 33, 30]. Several of these prominent benchmarks rely on dialogue and plot summaries to

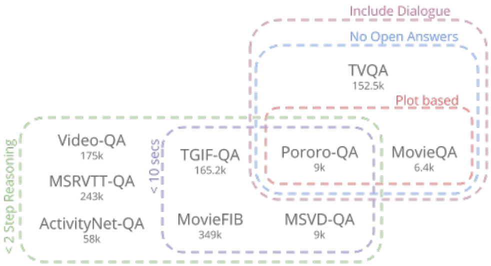


Figure 1. Existing VideoQA benchmarks have some drawbacks. Specifically, many have questions with less than 2 steps of reasoning, using short videos, having plot based questions, only multiple choice questions, and including dialogue.

reason over the contents of videos [20, 27, 15]. Models trained on these datasets have demonstrated a stronger dependence on the dialogue input than on the visual input, reducing these benchmarks' effectiveness at measuring visual spatio-temporal reasoning [27, 20]. Therefore, our project focuses pure video-only question answering benchmark.

Vision-only benchmarks exclusively ask questions that can be answered from a single frame of the video, questions that apply to the entire video with no temporal localization, or questions that ask ”what happened before/after/while <action>?”. They all, except [30, 29], refer to videos of less than 10 seconds long. Some have automatically generated questions from image descriptions [29, 33], some let human annotators choose from drop down menus [8], and others ask humans annotators to create questions [30, 27, 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 2-195 seconds long, and evaluates complex and multi-step temporal reasoning.

2.3. Compositional Reasoning

As questions become more complex, they often require multiple reasoning steps in order to find the answer. 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 of each action's occurrence. 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 are able to learn quickly by generalizing new information in existing contexts [26, 24]. Improving a model's ability to generalize will allow it to more quickly learn new domains, categories, and logical rules [18, 28]. .

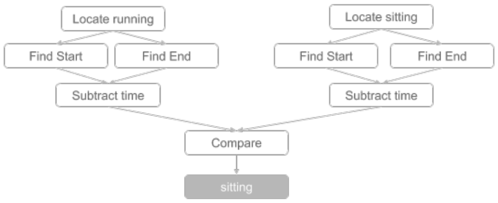


Figure 2. The substeps required to answer the question ”Was the person running or sitting for longer?”

Compositional questions have been used by ImageQA benchmarks to rigorously test models. Compositional questions tend to be longer, more complex, and use a wider variety of vocabulary. Furthermore, they can better test a model's reasoning ability both by requiring multiple steps of reasoning to answer the questions and by allowing for datasets to test generalization 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 to new concepts [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 model do after lower coat?”, requires finding when the model lowers her coat, then determining her activity after that. 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 those objects. 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

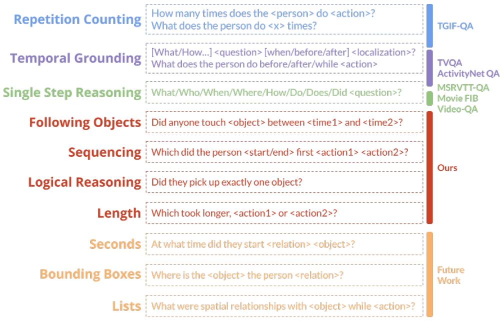

Repetition Counting
How many times does the <person> do <action>?
What does the person do <x> times?
TGIF-QA
Temporal Grounding
[What/How...] <question> [when/before/after] <localization>?
What does the person do before/after/while <action>
TVQA
ActivityNet QA
Single Step Reasoning
What/Who/When/Where/How/Do/Does/Did <question>?
MSRVTT-QA
Movie FIB
Video-QA
Following Objects
Did anyone touch <object> between <time1> and <time2>?
Sequencing
Which did the person <start/end> first <action1> <action2>?
Logical Reasoning
Did they pick up exactly one object?
Length
Which took longer, <action1> or <action2>?
Ours
Seconds
At what time did they start <relation> <object>?
Bounding Boxes
Where is the <object> the person <relation>?
Lists
What were spatial relationships with <object> while <action>?
Future Work

Figure 3. Various types of temporal reasoning.

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. 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 AGQA, we first established what types of temporal reasoning to explore. As mentioned previously, other vision-based VideoQA datasets have looked exclusively at questions that could be answered with one frame, questions that refer to the entire video with no temporal localization, and questions that ask ”what happened” before, after, or while an event was occurring [27, 20, 8, 15, 29, 22, 33, 30].

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, reasoning more deeply over a subject’s changing relationship with objects over time, and reasoning about when actions precede, succeed, or coexist with one another.

AGQA requires all the above types of reasoning to succeed. Future datasets may be able to expand even further by including questions with 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 be large, mitigate biases in the answer distribution, allow for novel compositions and 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, 25] on Charades videos. These videos are 30 seconds long on average and filmed by non-professionals who act out everyday activities using 36 objects. Charades annotates the times actions occur, and Action Genome annotates objects and 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 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 detailed spatio-temporal scene graphs 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 our own.

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>?”. These tags can be objects, relationships, actions, or temporal localization phrases (e.g. ”before putting down the dish”). 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.

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

3.3. 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 8 million questions on just 250 videos. Automatically generated datasets have previously been criticized for lacking diversity [30]. We have 32 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.

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. Our question generation pipeline allows for this balancing because we have the structural and semantic information associated with each template to split the questions into relevant categories before downsampling.

3.5. Novel Composition Tests

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.

3.6. 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 subsets of questions from the training and test sets. We are able to do this because our pipeline gives us control over the content of all the questions.

At the end of the 10 weeks of the DREU program, we have not yet balanced the dataset's open answer questions or divided the questions into a suite of metrics. However, we plan to incorporate these steps by the end of the summer.

4. Results

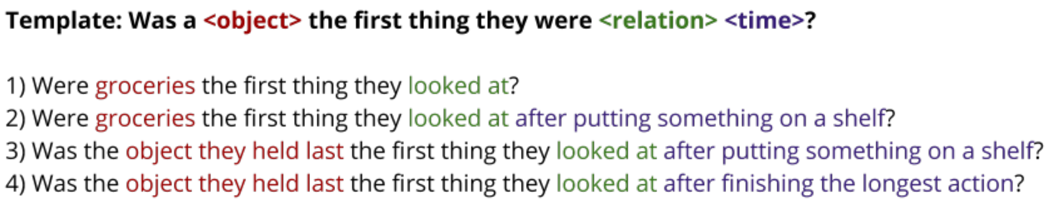
A the end of the 10 weeks of DREU, we have generated 8 million questions from 32 templates and 250 videos. Our final Video QA dataset consists of three parts: the Charades videos, without any annotations or the corresponding spatio-temporal scene graphs, the questions, the answers, and the suite of metrics.

In other words, AGQA = $\{V, Q, A, M\}$, 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 $q_{ij} \in q_i$ there is a corresponding answer $a_{ij} \in q_i$. M is a suite of metrics specifying which questions from the training and test set should be included to get a score for each metric.

Figure 6 shows that, like other question-answering datasets, our answer distribution has a very long tail. We plan to mitigate the effects of this long tail by balancing the distribution of answers for each category.

Certain question templates generate many more questions than others. Across the 250 videos, the template that asks "Did they <relation> a <object1> or a <object2>?" generates over 2,138,702 questions, while the template that asks "What is the last thing they did to the <object> only generates 345 questions. This disparity occurs because there are many more permutations to fill in all 4 tags of the first question than the 1 tag in the second question. Furthermore, many videos do not have a valid object to fill into the second question, because the person is doing multiple things to the object at once. The questions with the most answers are yes/no questions, so this disparity is also reduced through balancing.

Figure 9 shows that we have a large number of compositional steps in many of our questions. Our method of counting is described in more detail in Figure 4. The



Template: Was a <object> the first thing they were <relation> <time>?

1) Were groceries the first thing they looked at?
2) Were groceries the first thing they looked at after putting something on a shelf?
3) Was the object they held last the first thing they looked at after putting something on a shelf?
4) Was the object they held last the first thing they looked at after finishing the longest action?

Figure 4. Our pipeline uses indirect references to create a variety of questions with different required reasoning seps from the same template. 1) uses all direct references and no temporalization. It has two steps of reasoning (determining if looked at groceries, then finding if that is the first thing at which they looked). 2) Adds temporal localization for a total of 4 steps (localizing when putting something on a shelf, then shifting attention after). 3) Uses an indirect reference for the object, adding one additional step (the last object held) for a total of 5 steps. 4) adds an indirect reference within the temporal localization to add a step (find the longest action) for a question with a total of 6 compositional steps required to answer.

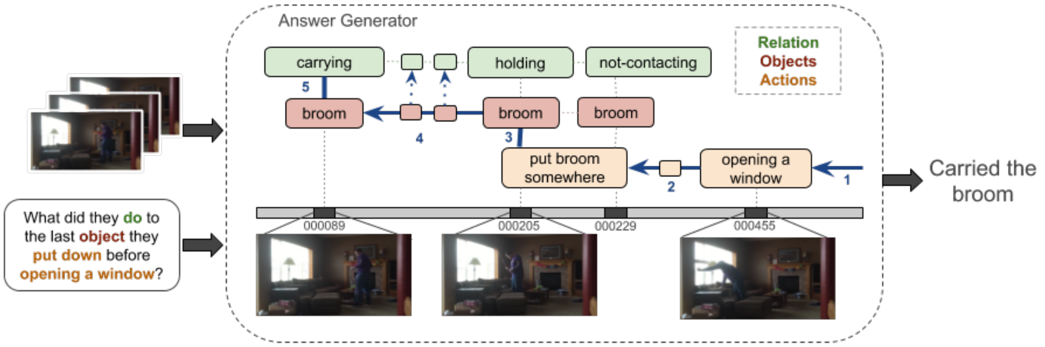


Figure 5. The sequence in which our answer generator traverses the spatio-temporal scene graph to automatically generate the answer to a question. The input question can be decomposed into the following spatio-temporal operations: 1) localize when the actor opened a window, 2) find the last event before then when the actor put an object down, 3) determine which object (the broom) was put down, 4) look for a different relationship between the actor and this broom, 5) Output that the actor “carried the broom”.

highest number of compositional steps in a single template is 5, suggesting that many questions gain additional compositionality from indirect references. These results show we have generated complex questions with many compositional steps.

4.1. Structural and Semantic Categories

Following the categorizations of [7] we break down the questions into structural and semantic categories.

Structural categories define the structure of the question. We have three structures, verify, choose, and compare. Verify questions have ”Yes” and ”No” answers verifying if a question is correct (e.g. Did they carry the blanket?). Choose questions choose between two answers (e.g. Did they carry the blanket or the dish?). Query questions have open answers (e.g. What did they carry?). As seen in Figure 4.1, there are significantly more ”verify” questions than the other categories. However, this distribution will shift after balancing. For example, after balancing only 22% of the verify questions remain. We plan to balance the other categories in the future.

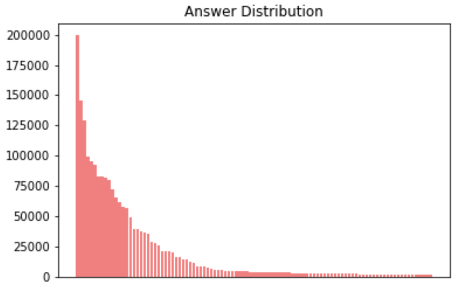


Figure 6. The distribution of the top 100 answers (excluding ”Yes” and ”No”).

Semantic categories define what ideas the question is mainly reasoning about. In our dataset, these semantic categories are objects, relationships, object-relationship pairs, and actions. As seen in Figure 8, the dataset consists mostly of questions asking about object-relationship pairs and objects.

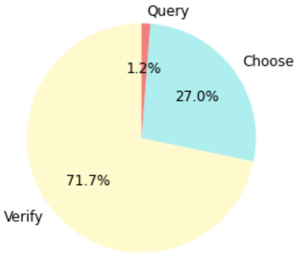


Figure 7. The structural distribution of questions.

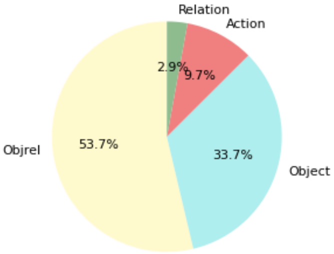


Figure 8. The semantic distribution of questions.

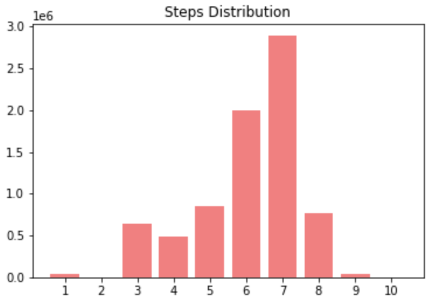


Figure 9. The distribution of the number of compositional steps.

5. Future work

To complete this project, we plan to expand upon our current progress in four ways through the fall. 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 balance the dataset on non-binary questions. Third, we will implement a suite of metrics. Fourth, we will evaluate existing VideoQA models on our dataset [19, 5, 21]

We have already balanced the dataset for yes/no questions. We plan to smooth the distribution of all open ended questions.

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 ≥ 30 seconds
Actions	Train on videos with < 5 actions. Test on videos with ≥ 5 actions
Compositional Steps	Train on questions with < 6 steps. Test on questions with ≥ 6 steps
Novel Compositions	See Table 2
Indirect Consistency	Accuracy on questions referring to the same ideas but using different numbers of indirect references
Direct Only	Questions with no indirect refs
% Training Data	Train on only 1, 5, 10 and 20 percent of training data
Answer Legitimacy	
Sequencing	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

We want AGQA 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 1. 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.

Table 2. Novel Composition

Name	Novel Combinations
object-relationship	touching table, touching food, beneath table, beneath food
repetition	taking a dish from somewhere, putting some food somewhere
first/last	looking at first, behind first, holding first
before/after	before standing up, before playing with phone, before throwing broom
longer	longer than standing up, longer than playing with phone, longer than throwing broom

6. Conclusion

Our project automatically generates a dataset of question-answer pairs for video and a suite of metrics on which to measure performance. AGQA 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 Conference 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] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 677–685, 2017.

[24] 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.

[25] 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.

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

[27] 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.

[28] Ben-Zion Vatashsky and Shimon Ullman. Vqa with no questions-answers training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10376–10386, 2020.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.