This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning

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

Ranjay Krishna
Stanford University
ranjaykrishna@cs.stanford.edu

Maneesh Agrawala
Stanford University
maneesh@cs.stanford.edu

Abstract

Visual events are a composition of temporal actions involving actors spatially interacting with objects. When developing computer vision models that can reason about compositional spatio-temporal events, we need benchmarks that can analyze progress and uncover shortcomings. Existing video question answering benchmarks are useful, but they often conflate multiple sources of error into one accuracy metric and have strong biases that models can exploit, making it difficult to pinpoint model weaknesses. We present Action Genome Question Answering (AGQA), a new benchmark for compositional spatio-temporal reasoning. AGQA contains $192M$ *unbalanced question answer pairs for* $9.6K$ *videos. We also provide a balanced subset of* $3.9M$ *question answer pairs,* 3 *orders of magnitude larger than existing benchmarks, that minimizes bias by balancing the answer distributions and types of question structures. Although human evaluators marked* 86.02% *of our question-answer pairs as correct, the best model achieves only* 47.74% *accuracy. In addition, AGQA introduces multiple training/test splits to test for various reasoning abilities, including generalization to novel compositions, to indirect references, and to more compositional steps. Using AGQA, we evaluate modern visual reasoning systems, demonstrating that the best models barely perform better than non-visual baselines exploiting linguistic biases and that none of the existing models generalize to novel compositions unseen during training.*

1. Introduction

People represent visual events as a composition of temporal actions, where each action encodes how an actor's relationships with surrounding objects evolves over time [44, 46, 30, 37]. For instance, people can encode the video in Figure 1 as a set of actions like putting a phone down and holding a bottle. The action holding a bottle can be further decomposed into how the actor's relationship with the bottle evolves – initially the actor may be twisting the bottle and then later shift to holding it. This ability to decompose

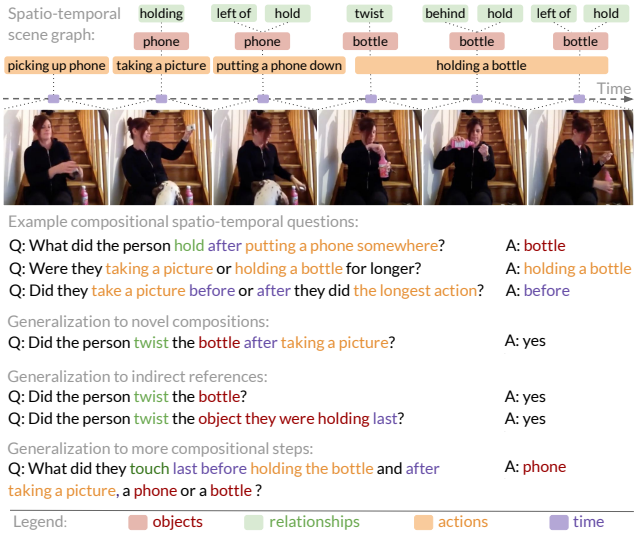

Figure 1. We introduce AGQA: a new benchmark to test for compositional spatio-temporal reasoning. AGQA contains a balanced $3.9M$ and an unbalanced $192M$ question answer pairs associated with $9.6K$ videos. We design handcrafted programs that operate over spatio-temporal scene graphs to generate questions. These questions explicitly test how well models generalize to novel compositions unseen during training, to indirect references of concepts, and to more compositional steps.

events is reflected in the language people use to communicate with one another [8, 42], so tasks involving both vision and language comprehension, such as answering questions about visual input, can test models' compositional reasoning capability. We can ask questions like "What did the person hold after putting a phone down?" and expect a model capable of compositional spatio-temporal reasoning to answer "bottle." While such behavior seems fundamental to developing vision models that can reason over events, the vision community has only developed compositional question answering benchmarks using static images [17] or synthetic worlds [31, 56] which either are not spatio-temporal or do not reflect the diversity of real-world events.

Although questions and visual events are composed of multiple reasoning steps, existing video question answering benchmarks conflate multiple sources of model errors

Table 1. AGQA is 3 orders of magnitude larger than all existing VideoQA benchmarks. It contains real-world videos and compositional open-answer questions with action, object, and relationship grounding. AGQA's questions focus on visual comprehension and do not require common sense or dialogue understanding.

Dataset	Video			Question answers				Grounding		
	Avg. length (s)	# videos (K)	Real-world	Not dialogue related	Open answer	Compositional	# questions	objects	relationships	actions
MarioQA [43]	3-6	188		✓	✓		188K			
CLEVRER [56]	5	20		✓	✓	✓	282K	✓	✓	✓
Pororo-QA [27]	1.4	16.1	✓				9K			
MovieQA [48]	202.7	6.77	✓				6.4K			✓
SocialIQ [59]	99	1.25	✓				7.5K			
TVQA [33]	76.2	21.8	✓			✓	152.5K			✓
TVQA+ [34]	7.2	4.2	✓			✓	29.4K	✓		✓
MovieFIB[40]	4.9	118.5	✓	✓	✓		349K			
TGIF-QA [18]	3.1	71.7	✓	✓	✓		165.2K			
MSVD-QA [53]	<10	1.97	✓	✓	✓		50.5K			
Video-QA [61]	45	18.1	✓	✓	✓		175K			
MSRVTT-QA [53]	10-30	10	✓	✓	✓		243K			
ActivityNet-QA [58]	180	5.8	✓	✓	✓		58K			
AGQA	30	9.6	✓	✓	✓	✓	**192M**	✓	✓	✓

into a single accuracy metric [18, 53, 40, 61, 58]. Consider this stereotypical question-answer pair, Q: " What does the bear on the right do after sitting?" A: "stand up" [18]. A model's inability to answer such questions does not afford any deeper insights into the model's capabilities. Did the model fail because it is unable to identify objects like bear or relationships like sitting or does it fail to reason over the temporal ordering implied by the word after? Or did the model fail for a combination of these reasons?

Not only are failure cases difficult to analyze, but the inputs where the model correctly guesses the answer are equally difficult to dissect. Due to biases in answer distributions and the non-uniform distribution of occurrences of visual events, models may develop "cheating" approaches that can superficially guess answers without learning the underlying compositional reasoning process [36, 54]. To effectively measure how well models jointly compose spatio-temporal reasoning over objects, their relationships, and temporal actions, we need newer benchmarks with more granular control over question composition and the distribution of concepts in questions and answers.

To measure whether models exhibit compositional spatio-temporal reasoning, we introduce Action Genome Question Answering (AGQA)[1]. AGQA presents a benchmark of $3.9M$ balanced and $192M$ unbalanced question answer pairs associated with $9.6K$ videos. We validate the accuracy of the questions and answers in AGQA using human annotators for at least 50 questions per category and find that annotators agree with 86.02% of our answers. Each question is generated by a handcrafted program that outlines the necessary reasoning steps required to answer a question. The programs that create questions operate over Charades' action annotations and Action Genome's spatio-temporal scene graphs, which ground all objects with bounding boxes and actions with time stamps in the video [19, 45]. These programs also provide us with granular control over which reasoning abilities are required to answer each question. For example, some questions in AGQA only require understanding the temporal ordering of actions (e.g. "Did they take a picture before or after they did the longest action?") while some others require understanding actions in tandem with relationships (e.g. "What did the person hold after putting a phone somewhere?"). We control bias using rejection sampling on skewed answer distributions and across families of different compositional structures.

With our granular control over the question generation process, we also introduce a set of new training/test splits that test for particular forms of compositional spatio-temporal desiderata: generalization to novel compositions, to indirect references, and to more compositional steps. We test whether models (PSAC, HME, and HRCN [11, 32, 35]) generalize to novel compositions unseen during training — the training set can contain the relationship twist and the object bottle separately while the test set requires reasoning over questions such as "Did the person twist the bottle after taking a picture?" with both concepts paired together in a novel composition. Similarly, we test whether models generalize to indirect references of objects by replacing objects like bottle in "Did the person twist the bottle?" with an indirect reference to make the question "Did the person twist the object they were holding last?" Finally, we test whether models generalize to questions with more reasoning steps by constraining the test set to questions with more reasoning steps than those in the training set (e.g. "What did they touch last before holding the bottle but after taking a picture, a phone or a bottle?").

Using AGQA, we evaluate modern visual reasoning systems (PSAC, HME, and HRCN [11, 32, 35]), and find that they barely perform better than models that purely exploit linguistic bias. The highest performing model achieves only 47.74% accuracy, and HCRN performs only 0.42% better than a linguistic-only version. While there is some evidence that models generalize to indirect references, all of them decrease in accuracy when the number of compositional steps increase and none of them generalize to novel compositions.

[1]Project page: https://tinyurl.com/agqavideo

2. Related Work

Our work lies within the field of video understanding using language and is targeted towards the question answering task. We use spatio-temporal scene graphs to generate our questions, and we provide a suite of new evaluation metrics to measure compositional spatio-temporal reasoning.

Image question answering benchmarks. A wide variety of visual question answering benchmarks have been created over the past five years [21, 17, 2, 60, 14, 29, 62, 26]. These benchmarks vary in input, from synthetic datasets [21], to cartoons [2], charts [27], or real-world images [17, 29, 62, 14, 60, 2]. They also vary in the type of questions asked, from descriptive questions (who, what, where, when, which, why, how) [62], to ones requiring commonsense reasoning [60], spatial compositional reasoning [21, 17], or spatial localization [62, 29, 17]. These benchmarks facilitated the development of many model architectures and learning algorithms that demonstrate spatial compositional reasoning abilities [39, 49, 6]. However, none of these measure temporal reasoning beyond guessing common sense actions that usually require external knowledge [60].

Video question answering benchmarks. As shown by the benchmarks in Table 1, there is a growing interest in measuring video reasoning capabilities using question answering [48, 33, 18, 27, 53, 40, 61, 58, 56]. Several of these prominent benchmarks rely on dialogue and plot summaries instead of a video's visual contents [33, 48, 27, 59], resulting in models with a stronger dependence on the dialogue than on the visual input and therefore reducing the benchmark's effectiveness at measuring visual reasoning [48, 33].

Some video-only question-answering benchmarks are synthetically generated [56, 43], which affords the granular control necessary to measure model abilities like causality [56], or counting [43]. However, these benchmarks use short video clips, utilize only a handful of objects, focus on questions that require commonsense or external knowledge (Figure 2), and lack the visual diversity of real-world videos. Other video-only benchmarks suffer from the biases and simplicity associated with human generated questions [58, 48, 18, 33] or descriptions [53, 61]. The largest human-annotated [33] and generated [40] datasets contain $152.5K$ and $349K$ questions. In comparison, our corpus is purely vision based, is three orders of magnitude larger, and evaluates complex and multi-step reasoning.

Scene graphs. Scene graphs were first introduced as a Cognitive Science [4, 52] inspired representation for static images [29, 23]. Each scene graph encodes objects as nodes in the image and pairwise relationships between objects as directed edges connecting nodes. The Computer Vision community has utilized the scene graph representation for a variety of tasks including visual question answering [22], relationship modeling [38], object localization [28], evaluation [1], generation [20, 3], retrieval [3, 23] and few-shot learning [7, 10]. Of particular interest to our project is how scene graphs from Visual Genome [29] were used to create GQA, a benchmark for compositional spatial reasoning over an image [17]. Our work is a generalization of GQA's pipeline. While GQA uses indirect references to objects with attributes (e.g. "red") and spatial relations (e.g. to the left of), we also use temporal localizations (e.g. before), indirect action references (e.g. the longest action), and changes in a subject's relationship with objects over time (e.g. before holding the dish). Our programs operate over Action Genome's spatio-temporal scene graphs to automatically generate question-answer-video pairs [19].

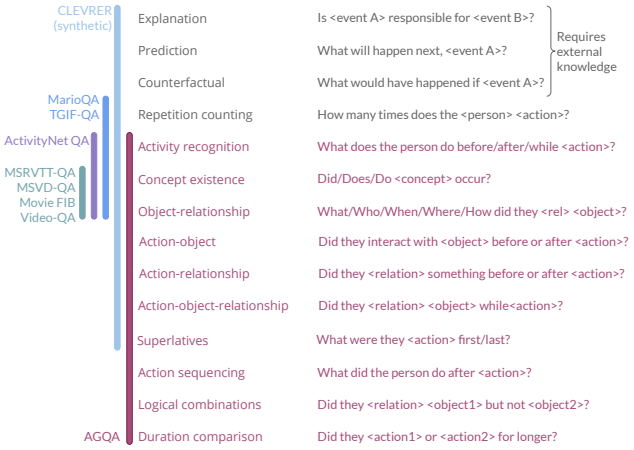

Figure 2. AGQA contains a variety of compositional spatio-temporal reasoning types that are absent from existing real-world video-only corpuses, including duration of actions, interactions between relationships and actions, action sequencing, and logical combinations. We focus on questions that require visual understanding, so we do not have questions that require external knowledge.

Compositional reasoning. While there are numerous definitions of compositionality, we in particular use what is more colloquially referred to as bottom-up compositionality — "the meaning of the whole is a function of the meanings of its parts" [9]. In our case, reasoning about the question "Was the person running or sitting for longer?" requires finding the start and end of when the person was running and sitting, subtracting the start from the end, then comparing the resulting lengths. Unfortunately, the most popular benchmarks and metrics defined to study compositional behavior have been limited to synthetic environments [25, 31, 21, 56] or to static images [17]. Recent work has argued the importance of compositionality in enabling models to generalize to new domains, categories, and logical rules [31, 49] and has discovered that current models struggle with multi-step reasoning [11]. These studies motivate a benchmark like ours that defines multiple metrics to explore compositional reasoning in real-world videos.

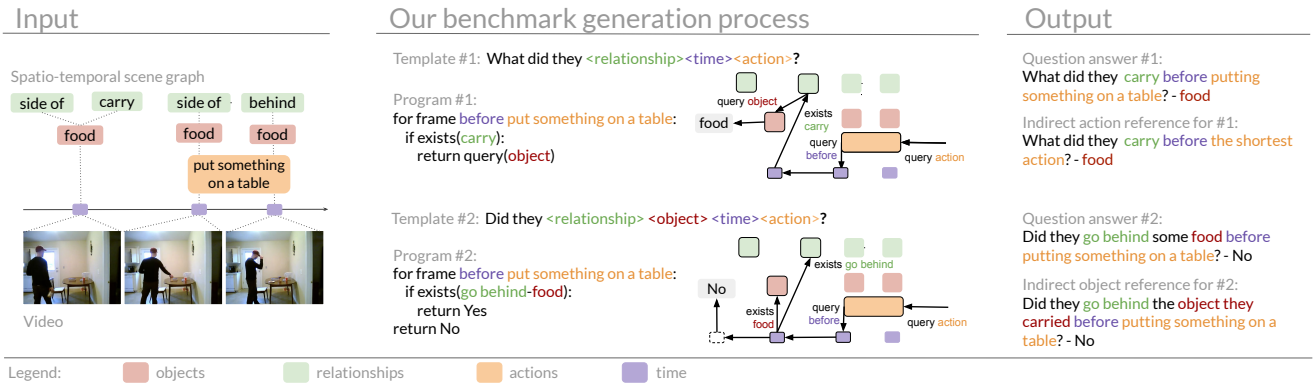

Figure 3. (Left:) Our benchmark generation process expects a dataset of videos with spatio-temporal scene graphs as input. (Middle:) We handcraft programs that operate over the scene graphs to generate questions and answers. (Right:) We balance generated questions and their corresponding answers using rejection sampling to avoid biases that models can exploit. We can control the number of reasoning steps required to answer a question by either developing more complex programs or by referencing visual concepts using indirect references (e.g. referring to a specific action as the shortest action or object as the object they carried).

3. The AGQA benchmark

Our benchmark generation process takes videos with annotated spatio-temporal scene graphs [19] as input and produces a balanced corpus of question-answer pairs (Figure 3). First, we consolidate and augment Action Genome's spatio-temporal scene graphs [19] and Charades' action localizations [45] into a symbolic video representation. Next, we handcraft programs that operate over the augmented spatio-temporal scene graphs and generate questions using probabilistic grammar rules. Then, we reduce biases in answer distributions and by question structure types, resulting in a balanced benchmark that is more robust against "cheating." Finally, we create new evaluation metrics that allow us to test how well models generalize to novel compositions, to indirect references, and to more compositional steps.

3.1. Augmenting spatio-temporal scene graphs

AGQA is generated using programs that operate over Action Genome's spatio-temporal scene graphs. Each spatio-temporal scene graph is associated with a video and contains objects (e.g. food, bottle) that are grounded in video frames, and the spatial relationships (e.g. above, behind), and contact relationships (e.g. carry, wipe) that describe an actor's interactions with the objects [19]. We augment Action Genome's spatio-temporal scene graphs with actions (e.g. running) from the Charades dataset, localized using time stamps for when the action starts and ends [45].

To use these scene graphs for question generation, we augment them by specifying entailments between actions and relationships, incorporating prior knowledge about action sequencing, merging synonymous annotations, and removing attention relationships. Some actions and relationships, such as carrying a blanket and twisting the blanket, entail other relationships such as holding and touching. We augment the scene graphs with such entailment relationships to avoid generation of degenerate questions like "Were they touching the blanket while carrying the blanket?" We created heuristics that adjust the start and end times of actions to avoid logical errors. For example, the action taking a pillow from somewhere would often end after the next action, holding a pillow, would start. To be able to generate questions that reason over the temporal ordering of these events, we modified the events so that the first action ends before the next one starts. To avoid generating simple questions with only one answer, we use co-occurrence statistics to prune relationships that only occur with one object category (e.g. turning off a light). We also consolidate references to similar objects and actions (e.g. eating a sandwich and eating some food) so that each concept is represented by one phrase. Finally, we remove all attention relationships (e.g. looking at) from Action Genome's annotations because our human evaluations indicated that evaluators were unable to accurately discern the actor's gaze.

The resulting spatio-temporal scene graphs have more clean, unified, and unambiguous semantics. Our final ontology uses 36 objects, 44 relationships, and 157 actions. There are $7,787$ training and $1,814$ test set scene graphs.

3.2. Question templates

To generate question and answer pairs from spatio-temporal scene graphs, we handcraft a suite of programs, each associated with a template (see Figure 3). Each template has a variety of natural language question frames that can be filled in by scene graph content. For example, a template "What did they <relationship><time><action>?" can generate questions like "What did they tidy after snuggling with a blanket?" and "What did they carry before putting something on a table?" To answer this question, the associated program finds the action put something on a

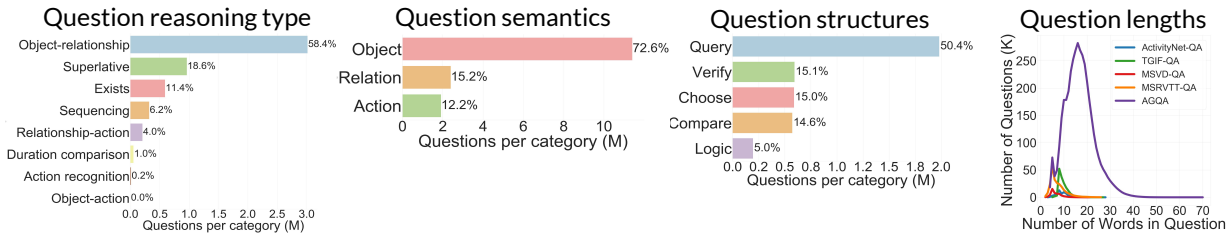

Figure 4. We classify each question in AGQA in three category types. Reasoning types distinguish which reasoning steps are required to answer the question. Semantic types split questions based on whether they are asking about an object, relationship, or action. Question structure types indicate the question's form. We also compare the distribution of question lengths of AGQA and existing video question answering benchmarks.

table, attends to events before that action, finds where the relationship carry occurs, and finally queries for the object.

This generation process associates each question with the reasoning skills and number of reasoning steps used to answer it. While some of the spatio-temporal reasoning skills required to answer our questions are inspired from existing corpora, successfully answering AGQA's questions requires a variety of new spatio-temporal reasoning absent in existing benchmarks (see Figure 2). Along with incorporating more reasoning skills, we increase the number of compositional reasoning steps necessary to answer a question by allowing question templates to use phrases that localize a time within the video and indirect references to objects, relationships, and actions. For example, we can replace food with the indirect reference the object being carried or walking through a doorway with the shortest action.

For each question, we also keep track of its answer type, semantic class, and structure. Open answer questions have many possible answers, while binary questions have answers that are Yes/No, before/after, or are specified as one of two options (e.g. carrying or throwing) within the question. A question's semantic class describes its main subject, a (1) object; (2) relationship; or (3) action. AGQA classifies questions into five structure categories: (1) query for all open questions; (2) compare for comparisons (3) choose for questions that present two alternatives from which to choose; (4) verify questions that respond yes or no to the question's contents; and (5) logic questions with logical conjunctions. We display the distribution of questions across these categories in Figure 4.

Before adding a question to the benchmark, we ensure that there is no ambiguity in answers by removing questions for which multiple elements could satisfy the constraints of the question. We avoid nonsensical compositions (e.g. "Were they eating a mirror?") by only asking about object-relationship pairs that occur at least 10 times in Action Genome. We also delete questions that answer themselves (e.g. "What did they hold while holding a blanket?"). Finally, we remove questions that always have one answer across all our videos (e.g. "Are they wearing clothes?").

We handcraft 269 natural language question frames that can be answered from a set of 28 programs. Using these programs, we generate $192M$ question-answer pairs, with over $45M$ unique questions and 174 unique answers.

3.3. Balancing to minimize bias

Machine learning models are notoriously adept at exploiting imbalances in question answering datasets [14, 17, 21]. We mitigate inflated accuracy scores by balancing our benchmark's answer distributions for each reasoning category and by the distribution of question structures.

We balance answer distributions with an approach inspired by the method described in GQA [17]. We first balance all answer distributions for each overall reasoning type and then for each concept within that reasoning type. For example, we first balance the answer distribution for the "exists" category, then that of the "exists-taking-dish-and-picture" category. For binary questions, we ensure that each answer is equally likely to occur. For open answer questions, we iterate over the answers in decreasing frequency order, and re-weight the head of the distribution up to the current iteration to make it more comparable to the tail.

Second, we use rejection sampling to normalize the distribution of question structures. Our templates generate more binary questions than the more difficult query questions. We balance the benchmark such that query questions constitute at least 50% of the benchmark. We further balance the binary answer questions such that approximately 15% are comparisons, 15% are choose questions, 15% are verify questions, and 5% use a logical operator. This new distribution of question structures increases the benchmark's difficulty and makes the distribution of required reasoning skills more varied.

Our balancing procedure reduced AGQA from an unbalanced set of $192M$ question answer pairs to a balanced benchmark with $3.9M$ question answer pairs. We provide a detailed algorithm in supplementary materials.

Table 2. Although humans verify 86.02% of our answers as correct, modern vision models struggle on a variety of different reasoning skills, semantic classes, and question structures. In fact, most of the increase in HCRN's performance comes from exploiting linguistic biases instead of from visual comprehension.

	Question Types	Most Likely	PSAC [35]	HME [11]	HCRN (w/o vision)[32]	HCRN[32]	Human
Reasoning	object-relationship	8.82	34.75	**43.91**	42.33	43.00	80.65
	relationship-action	50.00	56.84	57.84	**58.06**	56.75	90.20
	object-action	50.00	58.33	50.00	51.67	**63.33**	93.75
	superlative	10.29	30.51	**41.10**	36.83	37.48	81.25
	sequencing	49.15	59.95	59.60	**62.11**	61.28	90.77
	exists	50.00	69.94	70.01	72.12	**72.22**	79.80
	duration comparison	23.70	29.75	44.19	**45.24**	45.10	92.00
	activity recognition	4.72	3.78	3.23	7.57	**11.21**	78.00
Semantic	object	9.38	32.79	**42.48**	40.74	41.55	87.97
	relationship	50.00	65.51	66.10	**67.40**	66.71	83.58
	action	32.91	57.91	58.12	**60.95**	60.41	86.45
Structure	query	11.76	27.20	36.23	36.50	**37.18**	83.53
	compare	50.00	56.68	58.06	**59.65**	58.77	92.53
	choose	50.00	33.41	**49.32**	39.52	40.60	83.02
	logic	50.00	67.48	69.75	69.47	**69.90**	70.69
	verify	50.00	68.34	68.40	70.94	**71.09**	88.26
Overall	binary	50.00	54.19	**59.77**	57.93	58.11	86.65
	open	11.76	27.20	36.23	36.50	**37.18**	83.53
	all	10.35	40.40	**47.74**	47.00	47.42	86.02

3.4. New compositional spatio-temporal splits

With control over our generated set of questions, we measure how well models perform across different reasoning skills, semantic classes, and question structures. We also introduce a new set of train/test splits to test for particular forms of compositional spatio-temporal reasoning that require generalization to novel and more complex concepts.

Novel compositions: To test whether models can disentangle distinct concepts and combine them in novel ways, we hand-select a set of concept pairs to only appear in the test set. For example, we remove all training questions that contain the phrase before standing up, but retain only questions with the specified phrases in the test set.

Indirect references: The semantic categories in a question can be referred to directly (e.g. blanket, holding, and eating something) or indirectly (e.g. the object they threw, the thing they did to the laptop, and the longest action). Indirect references make up the core method through which we increase compositional steps. This metric compares how well models answer a question with indirect references if they can answer it with the direct reference.

More compositional steps: To test whether models generalize to more compositional steps, we filter the training set to contain simpler questions with $\leq M$ compositional steps, such as "What did they touch?" then reduce the test set to contain only questions with $> M$ compositional steps, such as "What did they touch last before holding the bottle but after taking a picture, a phone or a bottle?"

4. Experiments and analysis

We begin our experiments with scores from a human validation task on the AGQA benchmark that evaluates the correctness of our benchmark generation process. Next, we compare state-of-the-art question answering models on AGQA, revealing a large gap between model performance and human validation of our dataset. We report how well models perform on spatio-temporal reasoning, for different semantics, and for each structural category. Finally, we report how well models generalize to novel compositions, to indirect references, and to more compositional steps. All experiments run on the balanced version of AGQA.

Models: We evaluate three recent video question answering models: PSAC [35], HME [11], and HCRN [32]. PSAC uses positional self-attention and co-attention blocks to integrate visual and language features [35]. HME builds memory modules for visual and question features and then fuses them together [11]. HCRN, a current best model, stacks a reusable module into a multi-layer hierarchy, integrating motion, question, and visual features at each layer [32]. We use identical feature representations, from the ResNet *pool5* layer and ResNeXt-101, for all models.

We compare performance against a "Most-Likely" baseline that reports the accuracy of always guessing the most common answer after balancing (Section 3.3). Binary questions have a Most-Likely accuracy of 50% because they ask a Yes/No or before/after question, or they list answers in the question (e.g. "What did they hold, a bag or a dish?").

Table 3. We introduce new training/test splits to measure whether models generalize to novel compositions and to more compositional steps. B and O refer to binary and open questions. Overall, none of the models generalize.

		Most Likely	PSAC	HME	HCRN
Novel	B	50.00	43.00	**52.39**	43.40
Composition	O	15.87	14.80	19.46	**23.72**
	All	10.55	32.49	**40.11**	36.06
More	B	50.00	35.39	**48.09**	42.46
Compositional	O	14.51	28.00	33.47	**34.81**
Steps	All	12.81	31.13	**39.70**	38.00

Table 4. Here we break down the models' accuracy when generalizing to novel compositions of different reasoning types.

	Most Likely	PSAC	HME	HCRN
Sequencing	13.67	38.35	**44.77**	42.91
Superlative	12.60	31.97	**41.48**	34.01
Duration	10.96	38.65	48.19	**48.90**
Obj-rel	35.63	19.12	22.17	**25.71**

4.1. Human evaluation

To quantify the errors induced by our benchmark generation process, we hire subjects at a rate of $15/hr in accordance with fair work standards on Amazon Mechanical Turk [51]. We present at least 50 randomly sampled question per question type from AGQA to our subjects. We used the majority vote of three subjects as the final human answer. Human validation labeled 86.02% of our answers as correct, implying that about 13.98% of our questions contain errors. These errors originate in scene graph annotation errors and ambiguous relationships. We describe in supplementary materials the sources of human error and a second validation task. To put this number in context, GQA [17] and CLEVR [21], two recent automated benchmarks, report 89.30% and 92.60% human accuracy, respectively.

4.2. Performance across reasoning abilities

Each question is associated with the one or more reasoning abilities necessary to answer the question. By analyzing performance on each reasoning category, we get a detailed understanding of each model's reasoning skills. Overall, we find that across the different reasoning categories HME and HCRN perform better than PSAC (Table 2). HME outperforms the others on questions asking about superlatives, while HCRN outperforms the others on questions involving sequencing and activity recognition.

However, for most reasoning categories, HCRN does not outperform a language-only version of itself (HCRN w/o vision) by more than 1.5%. In fact, HCRN performs worse than its language-only counterpart on questions that involve sequencing actions as well as questions that reason over the length of time actions occurred. The only two reasoning categories in which the HCRN model outperforms the language-only baseline by more than 1.5% are on questions that focus on activity recognition and on questions comparing object-action interaction. Although HCRN improves on questions that require activity recognition, these questions are very challenging for all models and for humans. A more detailed breakdown of each section split by binary and open question types is in supplementary materials.

Table 5. We evaluate performance on questions with indirect references. Precision values are accuracy on these indirect questions when the corresponding question with only direct references was answered correctly, while recall values are accuracy on all questions with that kind of indirect reference.

	PSAC		HME		HCRN	
	Precision	Recall	Precision	Recall	Precision	Recall
Object	64.82	38.64	79.16	**47.32**	**81.03**	46.29
Relationship	40.84	24.12	**48.6**	29.39	46.77	**29.82**
Action	64.53	34.62	**81.68**	**45.15**	80.22	43.05
Temporal	66.48	33.15	80.71	**42.91**	**83.92**	42.13

4.3. Performance across question semantics

We also compare how models perform across different question semantic categories (Table 2). HCRN only improves over the language-only variant for questions that revolve around objects. However, object-related questions were the most difficult for all three models.

4.4. Performance across question structures

Different question structures also appear more challenging than others (Table 2). Open-ended query questions are very challenging and have the lowest accuracy for all models. HCRN outperforms the language-only variant in this category by only 0.68%. The models have similar performances for each structural category, with the exceptions that PSAC struggles the most with open-ended questions and HME outperforms the rest on choosing questions.

4.5. Generalization to novel compositions

All models struggle when tested on novel compositions unseen during training (Table 3). HME outperforms the others overall and on binary questions, while HCRN performs best for open-ended questions. However, no model performs much better than the Most-Likely model on open questions. Only HME outperforms 50% on binary questions with 52.39% accuracy.

We further break down the performance on novel compositions by composition type (Table 4). For example, in the sequencing category we remove compositions like before standing up from the training set and test how well models perform on questions with those compositions in the test set. We find that models perform the worst on novel compositions that involve new object and relationship pairs and best on reasoning about the length of novel actions. HME generalizes best to novel sequencing and superlative compositions, while HCRN generalizes best to novel compositions of the duration of actions and object-relationship pairs.

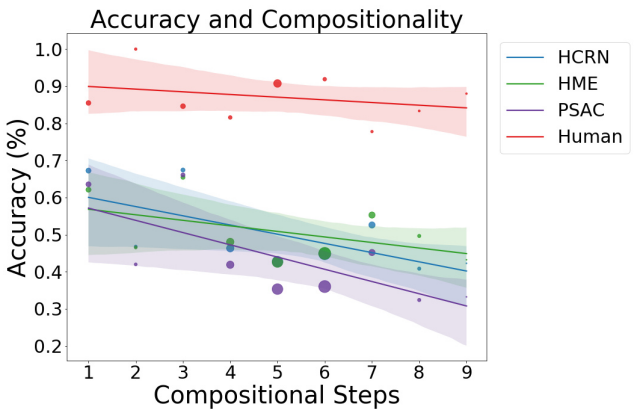

Figure 5. For all three models, we fit a linear regression and find that accuracy is negatively correlated with the number of compositional reasoning steps used to answer the question. However, the R^2 scores are relatively weak for all three: HCRN (.43), HME (.24), and PSAC (.51). This is likely because all three models barely outperform the Most-Likely baseline, even for small compositional steps. The human validation study's R^2 score is .09. The size of the dots correlates with the number of questions with that many steps, with the model's test set size scaled to be 1000x smaller. The shaded area is the 80% confidence interval.

4.6. Generalization to indirect references

We report precision and recall for how well models generalize to indirect references in Table 5. HCRN generalizes best to indirect object and temporal references, while HME generalizes best to relationship and action indirect references. However, the models still fail on at least nearly a fifth of questions with indirect references, even when it correctly answers the direct counterpart.

4.7. Generalization to more compositional steps

When trained on simple questions and tested on questions with more compositional steps, the models outperform the Most-Likely baseline on open questions. However, they still achieve less than 50% accuracy on binary questions. HCRN performs the best on open-ended questions, but HME generalizes better overall to questions with more compositional steps than the other models. This is likely because HME's architecture was explicitly designed to answer semantically complex questions, as it has a memory network for reasoning over question features [11].

Despite some aptitude at generalizing to more complex questions, these models' accuracy scores decrease as the number of compositional steps increase (Figure 5).

5. Discussion and future work

In conclusion, we contribute AGQA, a new real-world compositional spatio-temporal benchmark that is 3 orders of magnitude larger than existing work. Compositional reasoning is fundamental to understanding visual events [30, 37] and has been sought after recently by a number of papers [47, 55, 57, 13, 24]. However, to the best of our knowledge, AGQA is the first benchmark to use language to evaluate visual compositional desiderata: generalization to novel compositions, to indirect references, and to more compositional steps. Our experiments paint a grim picture — modern visual systems barely perform better than variants that exploit linguistic bias, and no models generalize to novel compositions. Although these models demonstrated some capability to generalize to more compositional steps, the overall trend was negative; model accuracy decreased as the number of reasoning steps increased.

While the results may appear grim, they also suggests multiple directions for future work to pursue. We expect researchers to utilize AGQA as a benchmark to make progress in the following directions:

Neuro-symbolic and semantic parsing approaches: We believe that the fundamental component missing in current models is the ability to extract systematic rules from the training questions. A model might perform better if it can operate in the "rule-space" using an explicit representation, either using neuro-symbolic [41] or semantic parsing [22, 15] to convert a question into an executable program. As AGQA provides ground truth scene graph annotations for all questions, it naturally leads into this line of work.

Meta-learning and multi-task learning: Since none of the models exhibited generalization to novel compositions, meta-learning might be a promising objective, which requires models to discover shared underlying compositional rules [12]. Such an approach can expose models to a number of learning environments with varying sets of novel compositions during the meta-train step. Another formulation worth exploring is multi-task learning, where models also simultaneously learn to detect objects, classify relationships, and recognize actions [5].

Memory and attention based approaches: HME outperformed other models in generalizing to more compositional steps. Perhaps this improvement is due to its explicit usage of memory when processing the question features. Future work can explore methods to keep track of each reasoning step with memory networks [50], or even use attention based approaches [16] to iteratively reason over the steps outlines in a question.

AGQA contributes a benchmark evaluating compositional spatio-temporal reasoning in visual systems along a variety of dimensions. The structure of this benchmark provides the computer vision community with multiple directions for future work.

Acknowledgements. This work was partially supported by the CRA DREU program, the Stanford HAI Institute, and the Brown Institute. We also thank Michael Bernstein, Li Fei-Fei, Lyne Tchapmi, Edwin Pan, and Mustafa Omer Gul for their valuable insights.

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] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[5] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[7] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2580–2590, 2019.

[8] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[9] MJ Cresswell. Logics and languages. 1973.

[10] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationships as functions: Enabling few-shot scene graph prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] 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.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[13] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.

[14] 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.

[15] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017.

[16] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018.

[25] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2019.

[26] 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.

[27] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022, 2017.

[28] 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.

[29] 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.

[30] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008.

[31] 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.

[32] 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.

[33] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018.

[34] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019.

[35] 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.

[36] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[37] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 812–819, 2014.

[38] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[39] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[40] 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.

[41] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

[42] Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.

[43] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[44] Jeremy R Reynolds, Jeffrey M Zacks, and Todd S Braver. A computational model of event segmentation from perceptual prediction. *Cognitive science*, 31(4):613–643, 2007.

[45] 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.

[46] Nicole K Speer, Jeffrey M Zacks, and Jeremy R Reynolds. Human brain activity time-locked to narrative event boundaries. *Psychological Science*, 18(5):449–455, 2007.

[47] Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5058–5067, 2017.

[48] 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.

[49] 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.

[50] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[51] Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206, 2019.

[52] Jeremy M Wolfe. Visual memory: What do you know about what you saw? *Current biology*, 8(9):R303–R304, 1998.

[53] 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.

[54] Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. What gives the answer

away? question answering bias analysis on video qa datasets. *arXiv preprint arXiv:2007.03626*, 2020.

[55] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10353–10362, 2019.

[56] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[57] Ting Yu, Jun Yu, Zhou Yu, and Dacheng Tao. Compositional attention networks with two-stream fusion for video question answering. *IEEE Transactions on Image Processing*, 29:1204–1218, 2019.

[58] 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.

[59] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[60] 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.

[61] 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. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4334–4340, 2017.

[62] 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.