6. Supplementary

The following sections cover more details regarding the results of our experiments and benchmark generation process. First, we separate each model's performance on binary questions that have two possible answers versus on questions that have open answers. Next, we detail how we create high quality questions, including how we augment the spatio-temporal scene graphs, ignore unchallenging and ambiguous questions, and balance answer and question structure distributions. We then explain which compositions were held out when generating the novel composition training/test split. Finally, we describe the two human studies that validate the correctness of our generation process, analyze the source of errors, and provide recommendations for future work.

Although all our experiments were conducted on the balanced version of the dataset, we will also release two additional versions of the benchmark: a full unbalanced version and a smaller subsampled unbalanced version. Since the full unbalanced AGQA is 192M questions, training models with data at this scale might be prohibitive for smaller research groups. The smaller unbalanced set can be used to serve as a benchmark under resource constraints.

6.1. Binary versus open answer results

Since the number and distribution of possible answers for a question affects the model's likelihood of guessing the correct answer, we split the experimental results by binary and open answer questions. Models achieve lower accuracy on open answer questions than they do on binary questions in each reasoning category (Table 4). However, we find that models generalize to indirect references better on open answer questions than on binary questions (Table 2).

We separate our analysis on binary versus open answer questions to appropriately compare each model's performance against the Most-Likely baseline. The Most-Likely baseline for binary questions (e.g. Yes/No or before/after answers) is usually higher than that for open answer questions. In the balanced benchmark, the Most-Likely baseline has a 50% accuracy on Yes/No binary questions. Other binary questions compare the attributes of two elements (e.g. "Were they fixing a light or consuming some medicine for longer?") or offer a choice between two elements (e.g. "Did they open a closet or a refrigerator?"). All questions with the answer choice have a 50% chance of correctness if the model chooses one of the provided options. However, the category-wide Most-Likely baseline may be lower than 50% because the category-wide Most-Likely answer may not be one of the two presented options in a question. For open answer categories, the Most-Likely accuracy baseline is the percent of all questions in the category with the most common answer.

In this section, we explore models' accuracy on binary and open questions in each reasoning, semantic, and structural category (Table 4). We then further analyze the results from the metrics measuring generalization to novel compositions (Table 1), indirect references (Table 2), and compositional steps (Table 3).

Table 1. Results from the novel compositions metric split by binary and open answer questions. Questions with novel compositions in the superlative, sequencing, and object-relationships categories struggle with open-answered questions. B rows and O rows show results for binary and open answer questions respectively.

		Most Likely	PSAC	HME	HCRN
Sequencing	B	50.00	51.26	**56.94**	51.66
	O	9.83	23.99	31.24	**33.19**
	All	13.67	38.35	**44.77**	42.91
Superlative	B	50.00	40.28	**51.62**	40.68
	O	11.75	9.74	14.32	**16.15**
	All	12.60	31.97	**41.48**	34.01
Duration	B	50.00	54.61	**60.32**	56.00
	O	19.33	25.27	38.02	**42.94**
	All	10.96	38.65	48.19	**48.90**
Obj-rel	B	50.00	35.93	**44.81**	37.85
	O	66.32	2.66	0.00	**13.82**
	All	35.63	19.12	22.17	**25.71**

Reasoning categories (Table 4): Across all reasoning categories, models perform much worse on open answer questions than binary questions. HCRN is generally stronger on open answer questions than the other models, especially for questions involving sequencing, duration, and action recognition. In fact, HCRN improves upon its blind counterpart for all open-ended question categories with the exception of questions involving superlatives, in which it performs only 0.02% worse. In contrast, HCRN performs worse than its blind counterpart for binary questions in the duration, sequencing, and relationship-action categories.

Semantic categories (Table 4): The non-blind HCRN model outperforms all others on open-ended questions that reason over objects and actions. For binary questions on objects, HME performs over 5% better than all other models, but for questions reasoning over relationships, all models perform within 2% of one another.

Structural categories (Table 4): Each structural class contains either only binary or only open questions, so these results are identical to those in the main paper. For HCRN, using visual features improves accuracy for all structural categories besides compare.

Novel composition (Table 1): We split element pairs in the novel compositions metric into four different types. We provide more detail on the makeup of each type in Section 6.6. For each category we pair a phrase with several objects, relationships, or actions to create novel compositions for the test set. For example, the superlative row looks at questions that contain the concept first paired with different relationships, including behind and holding.

Table 2. Results from the indirect references metric split by binary and open answer questions. Generally, open answer questions see a greater increase in accuracy on questions with indirect references when the model can correctly answer the equivalent question without indirect references. B rows and O rows show results for binary and open answer questions respectively. The Precision values for indirect relationship questions are N/A because none of the direct counterpart questions were answered correctly.

		PSAC		HME		HCRN	
		Precision	Recall	Precision	Recall	Precision	Recall
Object	B	57.71	52.14	**70.44**	**59.62**	68.87	56.01
	O	72.91	26.19	87.26	35.97	**91.94**	**37.32**
	All	64.82	38.64	79.16	**47.32**	**81.03**	46.29
Relationship	B	40.84	43.03	**48.60**	**53.39**	46.77	51.39
	O	N/A	0.98	N/A	0.00	N/A	3.41
	All	40.84	24.12	**48.60**	29.39	46.77	**29.82**
Action	B	53.32	48.25	**68.73**	**59.66**	63.92	54.81
	O	75.69	23.75	**92.47**	33.57	92.18	**33.66**
	All	64.53	34.62	**81.68**	**45.15**	80.22	43.05
Temporal	B	50.06	47.16	**64.50**	**58.42**	61.19	53.63
	O	74.49	27.18	87.29	36.29	**92.27**	**37.23**
	All	66.48	33.15	80.71	**42.91**	**83.92**	42.13

Table 3. Results from the compositional steps metric split by binary and open answer questions. The models generalize very poorly to questions with more compositional steps.

		Most Likely	PSAC	HME	HCRN
More Compositional Steps	B	50.00	35.39	**48.09**	42.46
	O	14.51	28.00	33.47	**34.81**
	All	12.81	31.13	**39.70**	38.00

The models struggle to generalize to novel compositions for all categories except duration. They all perform the worst at generalizing to novel object-relationship compositions. Across all categories, HCRN performs the best with novel compositions in open-answer questions, but HME performs the best with novel compositions in binary questions.

Indirect references (Table 2): This metric measures a model's accuracy on questions with indirect references and phrases that specify one part of the video. The Recall score shows the overall accuracy on these questions. Many questions with indirect references (e.g. "Did they contact the object they were watching?") have an equivalent question with no indirect references (e.g. "Did they contact a television?"). The Precision score shows the accuracy of questions with indirect references when the model answered the equivalent question with no indirect references correctly.

The larger increase in accuracy from Recall to Precision on open answer questions than on binary questions implies that a model better generalizes to open answer questions with indirect references. For the questions for which the model answered the direct version correctly, HME performed better than the other models on binary questions and HCRN performed better on open-ended questions.

Compositional steps (Table 3): Models struggle to generalize to questions with more compositional steps when they train on questions with fewer compositional steps. HME performs the best overall and on binary questions, but HME performs better on open answer questions.

6.2. Scene graph augmentation details

Action Genome's spatio-temporal scene graphs [1] annotate five sampled frames from each Charades action [4]. Each frame contains object annotations with lists of the contact, spatial, and attention relationships between the subject and the object [1]. We generate questions and answers based on these spatio-temporal scene graphs. Inaccurate or incomplete scene graph data can lead to uninformative and incorrect question generation. Since the scene graph annotations in Action Genome are often noisy, inconsistent, and sparse, we augment them using the following techniques to minimize errors:

Duplication: When Action Genome contains multiple annotations for the same object, for example if both food and sandwich refer to the same object, questions become artificially hard to answer. A person or model who identifies the correct answer to the question "What were they eating?" would have 50% chance of answering the question incorrectly if the object is annotated twice.

We first tried replacing duplicate references with the hypernyms using WordNet [3], in this case replacing food with sandwich. However, the Amazon Mechanical Turk annotators who recorded the Charades videos in their homes used many different objects, including items that do not appear to be sandwiches to the human eye, to represent sandwiches. Therefore, merging duplicate annotations into the more specific annotation of sandwich still lead to inaccuracies. To reduce this sort of error, we resorted to use hypernyms instead of hyponyms. For example, we replace annotations of sandwich and groceries with food throughout the entire dataset.

Sometimes an object is annotated with with multiple references that are not synonymous (e.g blanket and clothes) and therefore can not be replaced throughout the entire dataset. We identify all such objects by identifying objects of different categories with an intersection over union of ≥ 0.5. We merged such object pairs by manually re-annotating each instance.

We similarly merge action annotations from Charades to remove duplicate action references. We also replace vague actions, like eating something with their most specific counterparts, like eating some food.

Inconsistency: There are also inconsistencies in annotations for the spatial relationships beneath and above. We define a person as beneath an object if the object was at head level or above, and above an object otherwise. However, the Action Genome annotations do not use a consistent rule between or even within object classes. We go through beneath and above relationships for every object, remove annotations for objects that have $\leq 95\%$ intra-class consistency, and flip annotations when necessary to make them

Table 4. Results split by binary and open questions. B rows and O rows show results for binary and open answer questions respectively.

	Question Types		Most Likely	PSAC	HME	HCRN (w/o vision)	HCRN	Human
Reasoning	obj-rel	B	50.00	47.91	**57.24**	52.30	52.88	78.95
		O	11.96	27.49	36.55	36.82	**37.49**	90.90
		All	8.82	34.75	**43.91**	42.33	43.00	80.65
	rel-action	B	50.00	56.84	57.84	**58.06**	56.75	90.20
	obj-act	B	50.00	58.33	50.00	51.67	**63.33**	93.75
	superlative	B	50.00	43.35	**56.77**	49.69	50.81	81.81
		O	8.46	12.22	**18.52**	18.51	18.49	80.77
		All	10.29	30.51	**41.10**	36.83	37.48	81.25
	sequencing	B	50.00	60.38	60.05	**62.51**	61.77	94.73
		O	5.57	4.60	1.29	9.91	**10.92**	85.18
		All	49.15	59.95	59.60	**62.11**	61.28	90.77
	exists	B	50.00	69.94	70.01	72.12	**72.22**	79.80
	duration	B	50.00	30.49	45.25	**46.22**	46.05	91.89
		O	5.60	3.26	6.25	10.33	**11.68**	92.31
		All	23.70	29.75	44.19	**45.24**	45.10	92.00
	activity recognition	O	4.72	3.78	3.23	7.57	**11.21**	78.00
Semantic	object	B	50.00	45.10	**56.18**	50.00	51.13	87.39
		O	11.85	27.27	36.33	36.59	**37.26**	90.90
		All	9.38	32.79	**42.48**	40.74	41.55	87.97
	relationship	B	50.00	65.51	66.10	**67.40**	66.71	83.58
	action	B	50.00	57.91	58.87	**61.68**	61.09	90.21
		O	4.20	3.68	3.84	8.12	**11.31**	80.95
		All	32.91	57.91	58.12	**60.95**	60.41	86.45
Structure	query	O	11.76	27.20	36.23	36.50	**37.18**	83.53
	compare	B	50.00	56.68	58.06	**59.65**	58.77	92.53
	choose	B	50.00	33.41	**49.32**	39.52	40.60	83.02
	logic	B	50.00	67.48	69.75	69.47	**69.90**	70.69
	verify	B	50.00	68.34	68.40	70.94	**71.09**	88.26
	Overall	B	50.00	54.19	**59.77**	57.93	58.11	86.65
		O	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

Table 5. A list of the overall reasoning types of questions in AGQA.

Reasoning type	Templates	Unbalanced (M)	Balanced (K)	Answering question involves	Example templates
Obj-Rel	11	81.237	3014.86	A specific interaction with a specific object	Was the person <relationship><object>? What were they <relationship>? Were they <relationship><object>first?
Rel-Action	1	0.392	206.11	A relationship compared to an action	Were they <relationship>something before or after <action>?
Obj-Act	1	0.006	0.48	An object in comparison to an action	Where they contacting <object>before or after <action>?
Superlative	10	8.877	961.65	An extreme instance of an attribute	Were they <relationship><object>first? What was the person doing for the most time?
Sequencing	3	0.927	320.39	The sequence in which two actions occur	Did they <action>before or after they <action? What did they do after <action>? What did they do before <action>?
Exists	6	176.485	590.35	Verifying if some concept exists	Were they <relationship><object>? Did they <action>? Did they <relationship>something?
Duration comparison	6	0.160	53.20	The length of time of actions	What did they spend the most amount of time doing? Was <action>something they spent less time doing than <action>? Did they <action>or <action>for more time?
Activity recognition	2	0.012	11.65	Determining what action occurs	What did they do after <action>? What did they do before <action>?

consistent with our definition.

Sparsity: Action Genome annotations are also sparse. Since Action Genome annotates 5 frames per Charades action, objects and relationships from one action are not always annotated in the frames sampled from other co-occuring actions. We propagate annotations to surrounding frames using simple heuristic rules. Since spatial relationships do not have entailments, we do not ask questions us-

ing spatial relationships for videos with sparse annotations. We consider the 30% of videos in which fewer than 60% of object annotations had spatial relationships to be sparse.

Entailments: Sometimes, Action Genome annotations do not always include all occurring relationships, leading to incorrect and uninformative questions like Q: "Were they touching the object they were carrying?" A: "No." To address this problem, we curate a list of relationship entailments; for instance, if someone is carrying something, they are also holding and touching it. Actions also entail particular relationships. For instance, snuggling with a pillow entails that someone is snuggling (a verb) with a pillow (an object). Therefore, we create a new class of "verb" relationships from the Charades action annotations.

Uncertainty in action localization: Since the exact time an action begins and ends is often ambiguous, actions that actually occur in sequence are often incorrectly annotated with some overlap, resulting in nonsensical questions. We curate heuristics of common sequences of actions to avoid issues with uncertain action localization. For example, a person must take an object before holding it and finish holding it before putting it somewhere. If these annotations overlap, we automatically adjust the time stamps such that they do not overlap. Using the same entailments, we assume actions must occur if they are missing. For example, if someone begins holding an object in the middle of the video, we can assume that just before they were taking an object from somewhere.

6.3. Question quality checks

To create challenging and quality questions, we audit each question before it is added to the benchmark using the following filtering processes:

Rare combinations: We remove questions with object and relationship pairs that occur less than 10 times (e.g. "Were they twisting the doorway?").

Blacklisted object-relationship pairs: Questions also cannot involve object-relationship pairs from a list we manually curated of pairs that are likely to occur in the video but not be annotated in the spatio-temporal scene graph (e.g. "Were they above the floor?" or "Were they wearing clothes?").

Answer in the question: If questions indicate the answer (e.g. "What were they twisting while twisting a blanket?"), we remove them.

Realistic decoy questions: For questions that ask if some action occurred, we create realistic decoys by only asking about actions in which the action's relevant object or verb exists in the video. For example, if a video includes the action opening a door, we include questions that have the same verb (e.g. "Did they open a refrigerator?") or the same object (e.g. "Did they fix a door?"). However, we do not include questions that ask about actions that do not overlap with any objects or verbs in the video (e.g. "Did they wash a window?").

Confusing relationships: Action Genome contains attention relationships (looking at and not looking at) and also sometimes annotates the lack of a contact relationship (not contacting). Our human evaluations uncovered that questions with these relationships, such as "Were they looking at the object they were behind before walking through the doorway?" are hard to answer correctly. Therefore we do not ask questions with these relationships in our benchmark.

Similar objects: We avoid asking about similar pairs of words like door and doorway within the same question.

Multiple possible answer: For each question, we check that the constraints of the question only lead to one possible answer. For example, if they are holding multiple things we cannot ask "What were they holding?"

Grammatical Correctness: We ensure grammatical correctness by specifying in each template the necessary tense for any relationship or action and the relevant articles for each object.

Only one possible answer: We ignore questions where there is only one possible answer in the entire dataset (e.g. since the verb turning off is only ever associated with the object light, we do not allow the question "What were they turning off?").

Action localization checks: We only allow an action to be referenced as the "longest" or "shortest" action if it is 7 seconds longer or shorter than all other actions. Similarly, we only allow questions comparing the length of two actions if they have more than a 7 second difference between them. Since localizing the beginning and end of actions is noisy, the 7 second buffer ensures that actions have a large enough difference in length reduce incorrect questions.

6.4. Templates

Our 28 templates generate AGQA's question-answer pairs (Table 9). Each template has multiple natural language options that can be filled with scene graph information to create a diverse set of questions (Figure 5). The templates are also each associated with a program that automatically generates the answer to questions using the spatio-temporal scene graph. These programs are composed of a discrete set of reasoning steps that can be combined in numerous ways to answer a wide variety of questions (Table 6). Each question has several group labels describing the question's required reasoning skills (Table 5), semantic class (Table 8), and structural category (Table 7).

6.5. Balancing

Our balancing process occurs in two rounds (Figure 3). First, we smooth the answer distributions for open answer questions and make each possible answer of binary questions equally likely (Algorithm 1). Then, we change the

Table 6. Listed are the reasoning steps used to generate an answer from a spatio-temporal scene graph. Items in these inputs and outputs are object, relationship, action, and frame nodes in the spatio-temporal scene graphs.

Category	Reasoning step	Inputs	Outputs
Filtering	query	item, attribute type	attribute
	getFrames	frame, scene graph, before/after	frames before or after indicated index
	exists	items, query item	true if query item in items, false otherwise
	objectRelation	frame, object, relationship	true if object relationship exists in the frame, false otherwise
	chooseOne	items, query item1, query item2	which of the two items is present in the list
	iterate	items, function, integer x	the first x items in data structure to return True in function
Verification	verify	boolean	"Yes" if true, "No" if false
	and	list of booleans	true if all items in list are true, false otherwise
	xor	two booleans	true if exactly one boolean is true
Comparison	equals	item1, item2	true if item1 equals item 2 false otherwise
	comparative	item1, item2, attribute, more/less	item that is more or less in reference to a certain attribute
	superlative	items, attribute, most/least	the item with the most/least in a certain dimension
	difference	value1, value2	the difference between the two values
Item Sets	overlap	items1, items2	true if overlap between items1 and items2, false otherwise
	containedIn	items1, items2	true if items1 contained in items2, false otherwise
	sort	items, attribute	sorted concepts by attribute

Table 7. Every question in AGQA is associated with one of these five question structures.

	Templates	Unbalanced (M)	Balanced (M)	Description	Example templates
Query	10	2.2	1.98	Open ended questions	Which object did they <relationship>? What did the person do <time><action>? What did they spend the longest amount of time doing?
Compare	7	1.5	0.57	Compare attributes of two options	Compared to <action>, did they <action>for longer? Did the person contact <object>before or after <action>?
Choose	3	6.1	0.59	Choose between two options	Was <object>or <object>the thing they <relationship>? Did they <relationship><object>or <object>first? Which did they <relationship>last <object>or <object>
Verify	6	131.0	0.59	Verify if a statement is true	Does someone contact <object>? Did they <relationship><object>last? Was the person <relationship>something? Did they <action>?
Logic	2	52.0	0.19	Use AND or XOR logical operator	Were they <relationship>both a <object>and <object>? Were they <relationship><object>but not <object>?

proportion of questions of each structure type to create a more diverse and challenging benchmark (Algorithm 2).

Across all balancing steps in this process, the algorithm deletes questions from a specified question category. For example, the exists-paper category includes all questions asking if the person contacts some paper in the video. If at any point a category only has one possible answer, all questions from that category are deleted. Within a category with multiple possible answers, we split questions further by the effect of their temporal localization phrase. A temporal localization phrase combines <time> and <action> phrases to focus the reasoning process on a segment of the video (e.g "before washing a dish"). Many questions with temporal localization phrases (e.g. "What did they put down last before washing a dish?") correspond to an identical question without the temporal localization phrase (e.g. "What did they put down last?"). Sometimes adding the temporal localization phrase changes the answer of the question. In this example, the answer would change if they put down something after washing a dish. In other cases, adding the temporal localization phrase does not change the answer. Although many more of the generated questions are in the latter category, where the temporal localization does not change the answer, questions where the temporal localization does change the answer are more difficult. We delete questions such that the number of instances in which a temporal localization phrase changes the answer is close to the number of times it does not.

Answer distributions: Binary answer distributions are first split into very specific content categories. For example, questions that ask "Did they lie on a bed or the floor first?" have the content category first-lie-bed-floor, with two answers bed and floor. We delete questions from the answer that is more frequent until both answers occur an equal number of times. We balance each individual content category, rather than binary questions overall, to reduce a model's ability to guess the right answer based on the question.

We then smooth answer distributions for open answer questions such as "What were they holding?" We define a

Table 8. Questions in AGQA are categorized as reasoning primarily about an object, relationship, or action.

	Templates	Unbalanced (M)	Balanced (M)	Example templates
Object	11	38.1	2.9	Were they contacting <object>before or after <action>? Which were they <relationship>, <object>or <object>? Was <object>the first thing they were interacting with?
Relationship	5	87.2	0.6	Was the person <relationship><object>? Did they <relationship>something before or after <action>? Was the person <relationship> something?
Action	12	67.6	0.4	Did the person <action>? Compared to <action>, did they <action>for longer? Did they <action>before or after <action>?

proportion b to represent the proportion of the "head," or the side of a specified index in the ordered distribution with the more frequent answers, to the "tail," or side of the same specified index in the distribution with the less frequent answers. To avoid errors from very infrequent answers, we ignore the answers that cumulatively represent at most 5% of the question-answers pairs in the distribution. Then, we place the splitting index at the most frequent answer in the distribution and randomly sample to delete questions from the head until either the head to tail ratio is equal to or less than b or deleting any more questions would change the frequency ordering. The splitting index moves down the distribution, and we delete more questions each round. This process smooths the distribution enough such that it reaches a condition that no more than 30% of question-answer pairs have as answers the most frequent 20% of answers types.

Algorithm 1: Answer distribution smoothing

Input: Q: Unbalanced question-answer pairs
Output: Question-answer pairs with smoothed answer distributions
for *reasoning type in reasoning types* **do**
  if *reasoning type is binary* **then**
    d_{reason} = questions to delete to make both answers equally plausible
  else
    d_{reason} = questions to delete so 20% of answers represent at most 30% of all questions
  delete d_{reason} questions from Q_{reason}
  for *content category in reasoning type* **do**
    if *content category is binary* **then**
      $d_{content}$ = questions to delete to make both answers equally plausible
    else
      $d_{content}$ = questions to delete so 20% of answers represent at most 30% of questions
    delete $d_{content}$ questions from $Q_{content}$

Algorithm 2: Structural type balancing

Input: Q: a set of questions with smoothed answer distributions
Input: P: a map from structural category to a percentage
Output: A set of questions balanced by structural type
for *struct in structural categories* **do**
  d_{struct} = number to delete from Q_{struct} to get P_{struct}
  N_{templ} = number of templates in struct
  split d_{struct} into a d_{templ} for each template such that $Q_{templ} - d_{templ} = Q_{struct}/N_{templ}$
  for *templ in structural category* **do**
    $N_{content}$ = number of content categories in templ
    split d_{templ} into a $d_{content}$ for each content category such that $Q_{content} - d_{content} = Q_{templ}/N_{content}$
    for *content category in template* **do**
      split $d_{content}$ into a d_{ans} to retain answer distribution.
      delete d_{ans} questions from Q_{ans}

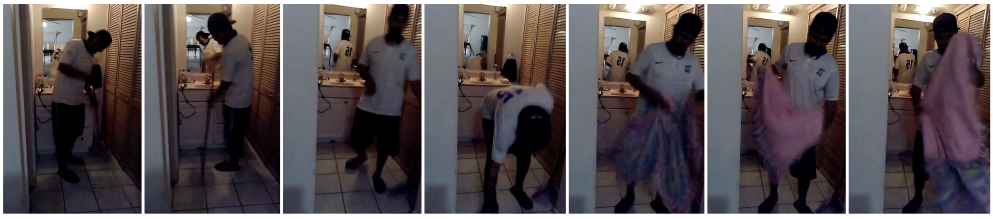
We defined this condition after experimenting with different parameters to empirically evaluate which condition created the smoothest answer distribution on a wide variety of distribution shapes without deleting more than 90% of the questions. If the splitting index reaches the tail and this condition is not met, the process repeats with a lower b proportion.

We first smooth overall reasoning categories, such as "superlative," and then smooth individual content categories, such as first-holding for the questions that ask "What were they holding first?"

At the end of this balancing round, each general and specific question category has balanced answer distributions that create a more challenging benchmark by reducing, though not eliminating, the model's ability to guess the an-

Table 9. AGQA's questions come from these 33 templates. Most templates optionally allow phrases that localize within time.

Template	Unbalanced (K)	Balanced (K)	Reasoning	Structural	Semantic	Steps	Natural language example
objExists	2316.0	99.0	exists	verify	object	1	Did they contact <object>?
objRelExists	14297.9	98.8	exists, obj-rel	verify	relationship	1	Was the person <relationship><object>?
relExists	20465.7	97.6	exists	verify	relationship	1	Did they <relationship>something?
actExists	66541.1	98.6	exists	verify	action	1	Did they <action>?
andObjRelExists	26010.4	93.4	exists, obj-rel	logic	relationship	3	Did they <relationship><object>and <object>?
xorObjRelExists	26010.4	97.6	exists, obj-rel	logic	relationship	3	Did they <relationship><object>but not <object>?
objWhatGeneral	34.2	31.2		query	object	1	What did they interact with?
objWhat	1796.6	1574.2	obj-rel	query	object	2	Which object were they <relationship>?
objWhatChoose	4254.1	194.3	obj-rel	choose	object	3	Which object were they <relationship>, <object>or <object>?
actWhatAfterAll	4.1	4.0	sequencing, activity-recognition	query	action	1	What did they do after <action>?
actWhatBefore	1.4	1.4	sequencing, activity-recognition	query	action	1	What did they do before <action>?
objFirst	146.2	136.9	superlative, obj-rel	query	object	2	Which object were they <relationship>first?
objFirstChoose	992.5	197.8	superlative, obj-rel	choose	object	3	Which object were they <relationship>first, <object>or <object>?
objFirstVerify	1213.4	99.0	superlative, obj-rel	verify	object	3	Were they <relationship><object>first?
actFirst	5.3	5.0	superlative, activity-recognition	query	action	1	What were they doing first?
objLast	246.5	223.9	superlative, obj-rel	query	object	2	Which object were they <relationship>last?
objLastChoose	929.8	196.2	superlative, obj-rel	choose	object	3	Which object were they <relationship>last, <object>or <object>?
objLastVerify	5339.0	98.8	superlative, obj-rel	verify	object	3	Were they <relationship><object>last?
actLast	1.3	1.2	superlative, activity-recognition	query	action	1	What were they doing last?
actLengthLongerCompare	39.3	12.6	duration-comparison	compare	action	5	Was the person <action>or <action>for longer?
actLengthShorterCompare	39.3	12.6	duration-comparison	compare	action	5	Was the person <action>or <action>for less time?
actLengthLongerVerify	39.3	12.6	duration-comparison	compare	action	5	Did they <action>for longer than they <action>?
actLengthShorterVerify	39.3	12.6	duration-comparison	compare	action	5	Did they <action>for less time than they <action>?
actLongest	2.3	2.3	superlative, duration-comparison	query	action	1	What were they doing for the most amount of time?
actShortest	0.5	0.5	superlative, duration-comparison	query	action	1	What were they doing for the least amount of time?
actTime	921.9	315.0	sequencing	compare	action	5	Was the person <action>before or after <action>?
relTime	391.8	206.1	rel-act	compare	relationship	5	Was the person <relationship>something before or after <action>?
objTime	6.4	0.5	obj-act	compare	object	5	Did the person contact a <object>before or after <action>?



Q1: After walking through a doorway, which object were they interacting with? A1: blanket

Q2: Was a broom one of the things they were contacting while holding the thing they went in front of? A2: Yes

Q3: While tidying something on the object they were touching first, of everything they went on the side of, what was the person on the side of last? A3: broom

Q4: Did the person touch the thing they took before or after they tidied up with the first thing they went behind? A4: after

Q1: After eating some food, did they touch a table or a chair? A1: chair

Q2: Between holding a cup of something and washing their hands, did they touch both some food and the object they were above before starting to sit at a table? A2: No

Q3: Which did they go on the side of before washing their hands but after sitting in a chair, a blanket or the last thing they took? A3: blanket

Q4: Of everything they went on the side of before washing a dish but after eating something, what did they go on the side of first? A4: blanket

Q1: What did they start to do first after holding some clothes? A1: playing with a phone

Q2: In the video, did they go behind the last thing they went in or the object they were putting down last first? A2: laptop

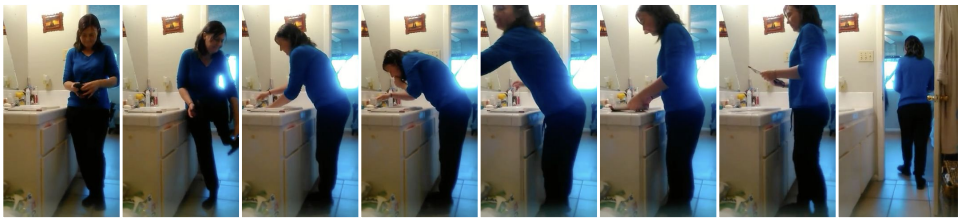
Q3: Did they watch something before or after throwing the object they were in front of last somewhere? A3: before

Q4: Which object were they in between watching a laptop or something on a laptop and taking the object they were putting down first from somewhere? A4: clothes

Figure 1. Examples of Questions in AGQA.

swer of a large number of questions based on just the questions themselves.

Question Structures: After the first round of balancing answer distributions, there are more binary questions than the more difficult open answer questions. We use rejection sampling again to change the distribution of question structure types to increase the proportion of open answered query questions. First, we determine how many questions of each structural type need to be deleted to get close to an ideal structural distribution. However, instead of randomly picking any question of that structural type to delete, we balance the amount of questions to delete to make the dis-



Q1: Were they interacting with a shoe before or after walking through a doorway? A1: before

Q2: Before watching some paper but after smiling, were they touching the object they were taking before holding some paper but after smiling but not the object they were in after taking a blanket from somewhere? A2: No

Q3: What did they go behind while walking through a doorway, a shoe or some paper? A3: paper

Q4: After smiling, was the person touching both a dish and the last thing they went behind? A4: Yes

Q1: Did the person put a dish somewhere before sitting in a bed? A1: Yes

Q2: Before taking something from a box but after sneezing, did they touch the thing they went in front of before starting to drink from a cup but not the object they were behind before opening a box but after walking through a doorway? A2: No

Q3: Before opening a box, were they holding a box? A3: No

Q4: Was the person holding a dish but not the object they were in front of first before sitting in a bed? A4: Yes

Q1: Did they put a dish somewhere before or after they started to watch outside of a window? A1: before

Q2: Before taking the object they were on the side of first from somewhere but after holding a blanket, was washing their hands or opening a refrigerator the activity they did for longer? A2: washing their hands

Q3: Between throwing a blanket somewhere and holding the first thing they washed, did they go in front of a laptop or some food? A3: laptop

Q4: Did they wipe the object they were behind first before starting to wash their hands? A4: No

Figure 2. Examples of Questions in AGQA.

tribution of templates and individual question categories as equally spread as possible. Within each distribution questions are deleted such that the original answer distribution holds.

After both rounds of balancing, the benchmark contains a larger percent of open answered and challenging questions, with less skew in the answer distribution.

6.6. Novel compositions

We explore several types of compositional pairs when constructing the training/test split for the novel compositions metric (Table 1).

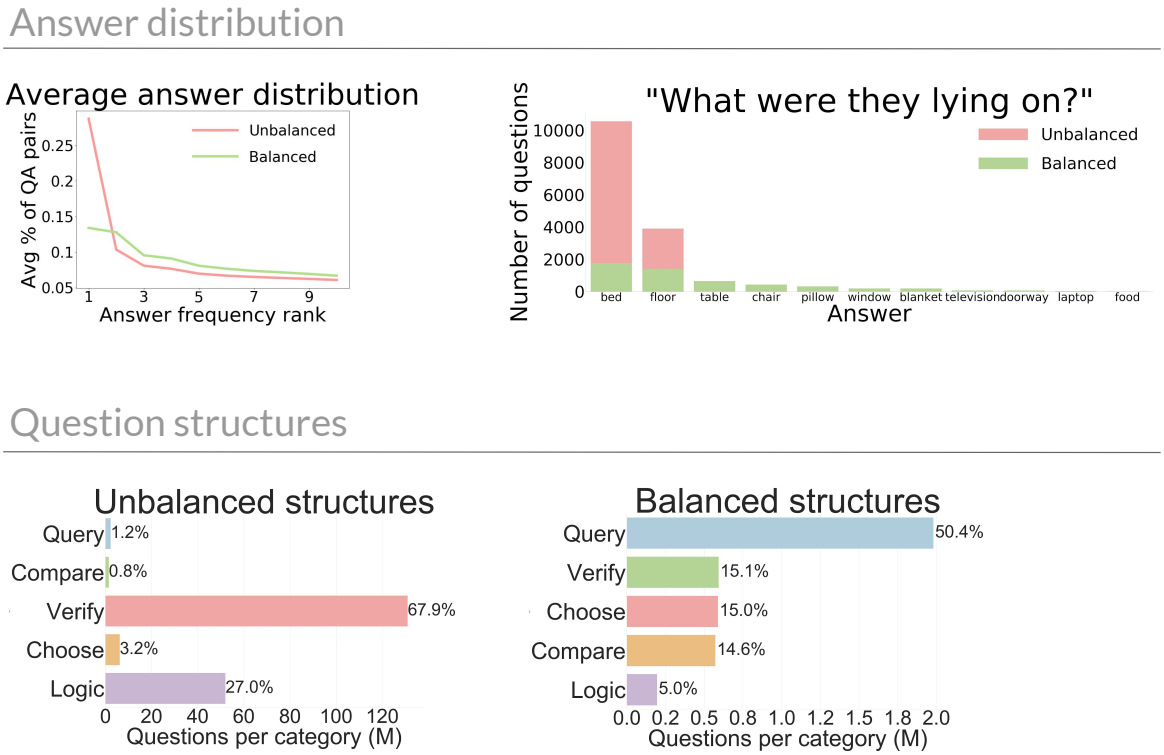


Figure 3. **Top Row:** The first round of balancing smooths the answer distributions of each question category. The top left figure shows the percentage of questions in the top 10 frequency ranks of each open answer category. The top right figure shows the answer distribution for all questions with the base "What were they lying on?" Before balancing, 64% of all answers were bed. After balancing, 34% of all answers were bed. **Bottom Row:** The second round of balancing deletes questions to change the distribution of question structures. Since query questions are more varied and more difficult, we make them the largest portion of the benchmark.

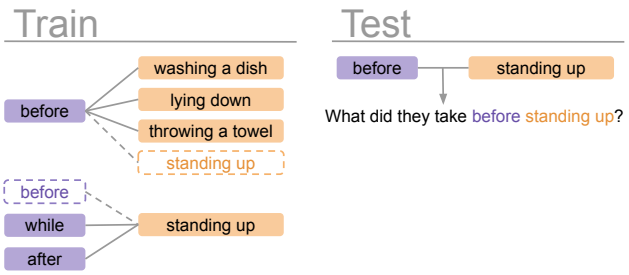


Figure 4. We design the training split for the novel compositions metric by ensuring that certain compositions like before and standing up occur individually in many questions, but never together in one question. In the test set, we only retain questions where these ideas are combined.

Sequencing: To test novel compositions in phrases that localize in time, we select six pairs of before-<action> combinations: before-standing up, before-walking through a doorway, before-playing with a phone, before-opening a laptop, before-grasping a doorknob, and before-throwing a broom somewhere. We selected phrases with a variety in frequency: standing up occurs very frequently (in 1704 videos), playing with a phone somewhat frequently (in 849 videos), and throwing a broom somewhere very infrequently (in 29 videos). In the test set, there are 55,119 questions with novel sequencing compositions.

Superlative: To test novel superlative compositions, we select six compositions of the superlative phrase first-<relationship>: first-behind, first-in, first-leaning on, first-carrying, first-on the side of, and first-holding. We chose spatial relationships (behind, in, on the side of) and contact relationships (leaning on, carrying, holding). In the test set, there are 108,003 questions with such novel superlative compositions.

Duration: To test novel duration compositions, we select the same six actions used in the sequencing category: standing up, walking through a doorway, playing with a phone, opening a laptop, grasping a doorknob, and throwing a broom somewhere. The test set includes questions that involve the length of these actions. In the test set, there are 10,050 questions with novel duration compositions.

Object-relationship interaction: Finally, to test for novel object-relationship compositions, we combine a variety of small and large objects with spatial and contact relationships that each occur frequently. The pairs we look at are: table-wiping, dish-wiping, table-beneath, dish-

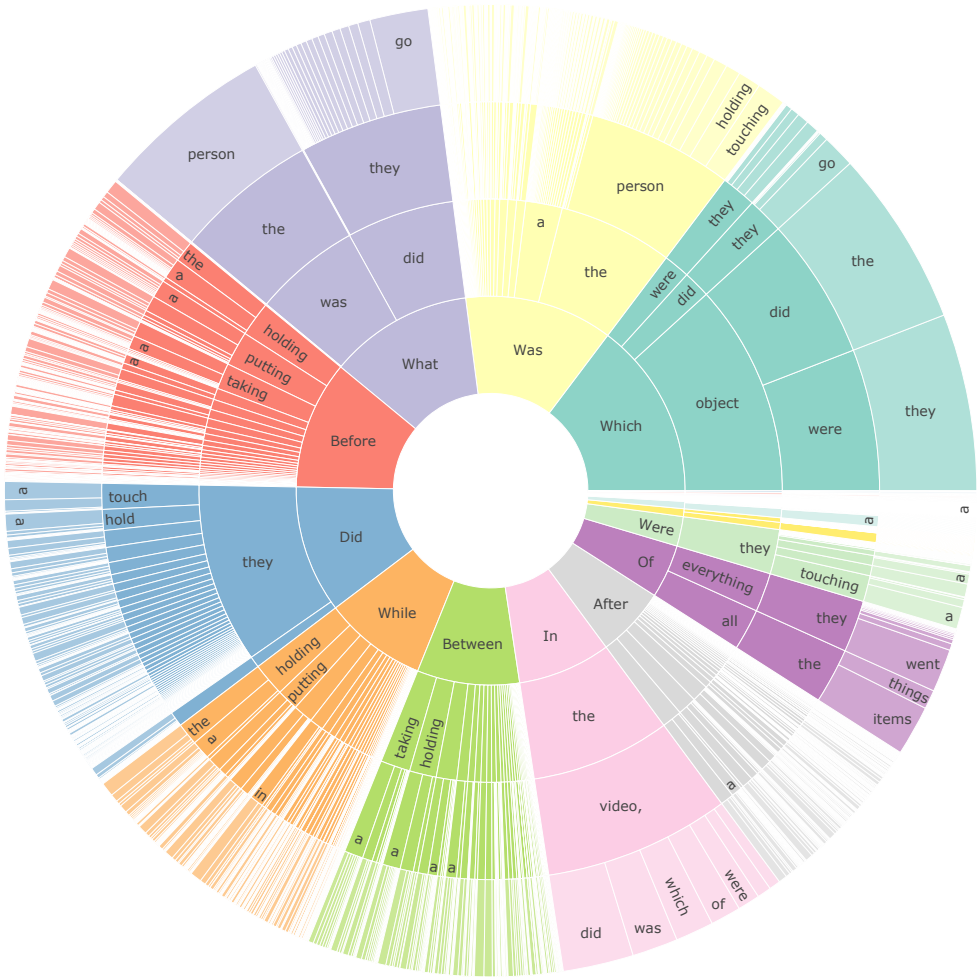


Figure 5. Although the questions in AGQA are generated from just 28 templates, they are linguistically diverse. There are 3.9 million total questions in the balanced benchmark and 2.39 million uniquely worded questions. Above are the first four words of all questions in the balanced benchmark, beginning from the center ring and moving outwards.

beneath, food-in front of, paper-carrying, and chair-leaning on. Any question that directly asks about this object-relationship pair (e.g Q: "What were they carrying?" A: "paper"), or that contains this object-relationship pair in an indirect reference (e.g. "the object they were carrying"), is removed from the training split and kept in the test split. In the test set, there are $24,005$ questions that contain object-relationship novel compositions.

6.7. Human study

We used humans to validate the correctness of AGQA's questions. This section covers the errors annotators found in our benchmark that originate from both our question generation process and those inherited from Action Genome and Charades' human annotation. Some errors enter the question generation process through poor video quality and missing, incorrect, and inconsistent scene graph annotations. We also found challenges in training annotators with the proper tools and term definitions to effectively annotate question correctness.

We run two human validation studies, one in which they verify our presented answer, and one in which they choose the correct answer from a dropdown list. We expected to find a performance drop when annotators selected their own answer from the dropdown list as it is a more difficult task to generate one's own answers. By analyzing both studies

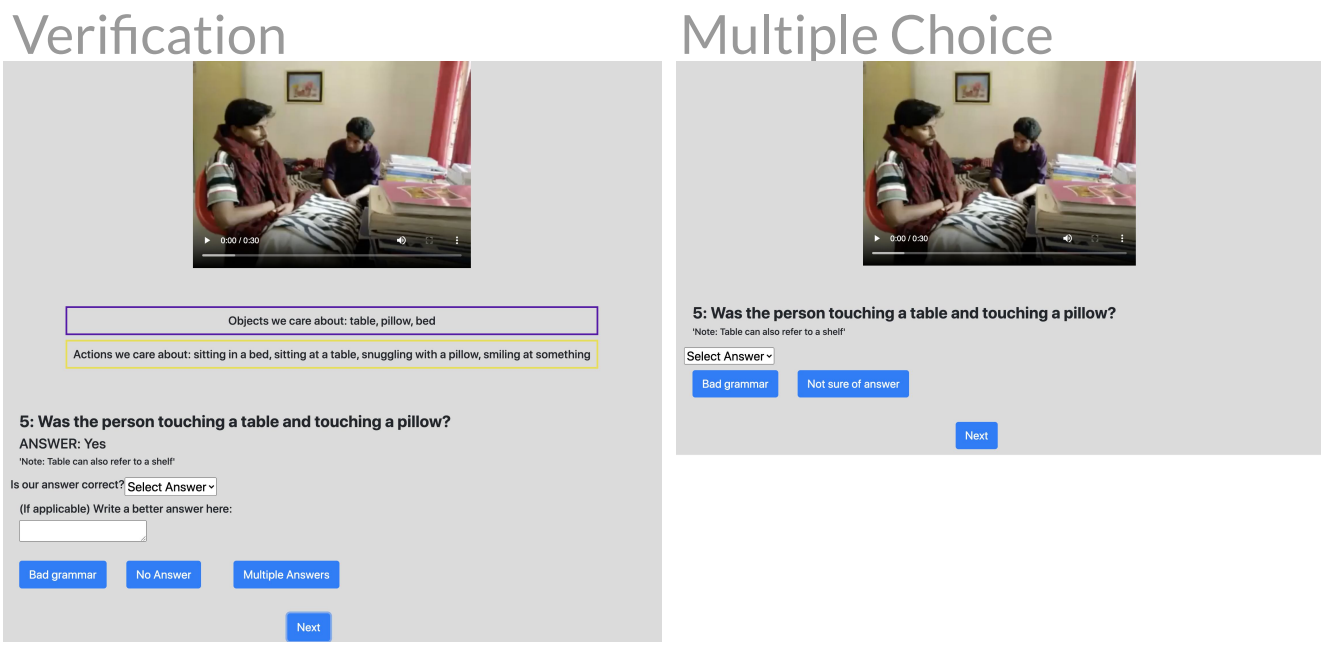


Figure 6. **Left:** Each annotator watches five videos, each associated with a question and an answer. Annotators indicate if that answer is Correct or Incorrect. **Right:** The annotators pick the closest answer from a dropdown menu of all activities occurring in the video.

we can identify what types of questions require higher cognitive effort from annotators and, therefore, lead to larger gaps in performance between the two task formats.

We share these findings in the hope that they synthesize the difficulties in the question-answering task and in inferring visual data from scene graphs. We will conclude with directions for future work on dataset generation to encourage exploration into these problems.

Unclear visual errors in videos: Some errors emerge because the objects, actions, and relationships in the video are visually unclear. This uncertainty arises from subtle movements and difficult to see objects. Charades is an visually diverse dataset because crowdworkers filmed the videos. However, this diversity in objects and quality may lead to uncertainty in annotation.

Missing annotations in scene graphs: Many of the scene graphs are missing action, object, or relationship annotations. In some videos, events occur before the first annotation or after the last annotation, leaving these events at the beginning and end of a video unannotated. Furthermore, some existing objects and relationships were not annotated in Action Genome because they were not a relevant object in a Charades annotation. For example, the person in the video may briefly touch a table, but not as a part of any larger action. Therefore, AGQA will answer the question "Did they touch a table?" with "No," even though that object-relationship pair occurs in the video.

Action Genome often had the most salient relationships annotated, but not all relationships. Many missing contact annotations could be added through entailments (e.g. someone holding an object is also touching it). However, not all were recoverable. For example, when watching a phone, subjects often, but not always, touched it as well. Therefore, we did not add that entailment. Spatial relationships were especially difficult to add through entailments, so we only included questions about spatial relationships for the 70% of videos in which at least 60% of object annotations included a spatial relationship.

AGQA does not include questions in which the relationship is an answer (e.g. Q: "What were they doing to the phone?" A: "watching"), because the scene graphs very often missed other relevant relationships that were not annotated.

Overall, missing annotations caused errors in AGQA's answers in two ways: assuming an existing event did not occur, or assuming there was one answer to a question when there were actually multiple possible answers. We addressed some of these errors through entailments, propagating labels to all annotated frames in an action, creating action priors, and ignoring spatial annotations on videos with sparsity. However, these steps did not fix all errors, and a full overhaul would require large-scale re-annotation efforts.

Incorrect annotations: Sometimes, existing annotations were incorrect. For instance, some objects would be annotated as different items than they were in reality. Similarly, one object in the video was sometimes annotated as different objects at different points of the same video (e.g. an-

notated as a blanket in some frames and as clothes in other frames).

The Charades action annotations were also often incorrect in their start time, end time, and length. Actions that occur in sequence overlapped in their time stamp annotations. For actions of the same family (e.g. taking a pillow and holding a pillow), we could infer the sequence and adjust the annotations so they did not overlap. However, this procedure does not work for for actions of different families, so some overlapping annotations remained. Incorrect time stamps also propagate to Action Genome's annotations. Action Genome uniformly sampled 5 frames from within the action for annotators to annotate. If the action's time stamps were incorrect, these sampled frames may not have been relevant and would have been difficult to annotate.

Incorrect augmentations: Incorrect annotations originated mostly from the Action Genome and Charades datasets. However, some were also added by our entailments strategies. For example, when people began holding a dish in the middle of the video, the annotation taking a dish was often missing, so we automatically added that annotation. However, in the case when the subject walked into frame in the middle of the video, already holding the dish, our entailments inserted incorrect actions.

Inconsistent annotations: Different annotators appear to have brought different priors on terms and annotation styles. Annotators also use synonymous annotations interchangeably, leading to inconsistent labels (e.g. eating something and eating some food). Annotators also used inconsistent definitions on terms such as in front of, behind, above, beneath, closet, leaning on, snuggling, sitting down, and standing up. These inconsistencies lead to inconsistencies in questions. The question "Were they leaning on a closet?" may have different answers dependent on the annotator's definition of leaning on and closet. As described in Section 6.2, we addressed some of the above and beneath inconsistencies by keeping, switching and ignoring them by class, but our mitigation strategies did not solve all inconsistencies.

The annotators were inconsistent along several other fronts as well. Some annotators annotated interactions with the phone that was filming the video, while others ignored those interactions. Some annotators annotated actions performed by animals in the video doing actions like watching out of a window, while others ignored those actions. Some annotators annotated each individual action separately (e.g. "eating some food" each time the person raised food to their mouth), while others annotated groups of actions (e.g. one "eating some food" annotation for the entire process). We did not include questions about the number of times each action occurred because of these inconsistencies, and we merged overlapping identical annotations.

Annotators also held different priors as to the length of actions that indicate a transition in state (e.g. putting a dish somewhere and sitting down). We did not include questions asking about the length of transition verbs to avoid bringing these inconsistencies to our questions.

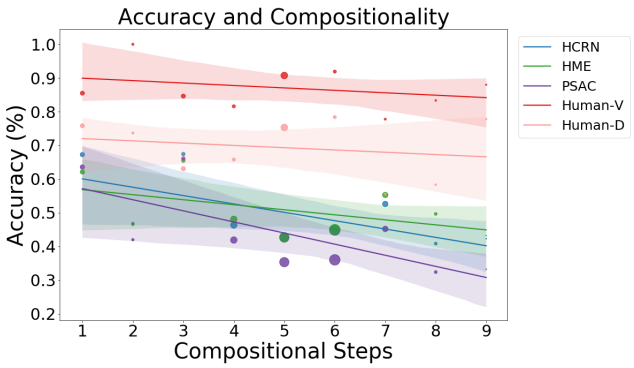


Figure 7. For all three models, we fit a linear regression and find that accuracy is negatively correlated with the number of compositional reasoning steps used to answer the question. Although the R^2 scores are relatively weak for all three models: HCRN (.43), HME (.24), and PSAC (.51), the correlation is weaker for both the human verification task (Human-V) and the dropdown task (Human-D) with R^2 scores of .09 and .04 respectively. The size of the dots correlates with the number of questions with each number of steps, with the model's test set size scaled to 1000x smaller. The shaded area is the 80% confidence interval.

Human and AGQA definition mismatches: Similar inconsistencies in term definitions among the annotators who annotated Charades and Action Genome appear in annotators answering AGQA's questions. In reducing the effect of synonyms and multiple annotations of the same object causing errors, we combined terms with similar semantic meaning (e.g. towel and blanket are all referred to with the term blanket blanket). However, the adjusted term may not best describe the item in the annotator's mind. To minimize the effect these errors had on our reported accuracy, we wrote notes next to the question to specify the constraints we used. However, this shift in definition requires extra cognitive effort from the annotator answering the question.

Annotators also occasionally said the AGQA answer was incorrect, then wrote as a correct answer a term that did not occur in the dataset. Similarly, annotators did not know constraints on the possible object-relationship pairs, so they inferred some pairs that do not exist in AGQA (e.g. watching a pillow).

Explaining these errors to our annotators: To evaluate AGQA, we designed our human evaluation protocol by minimizing the errors due to incorrect definitions and missing annotations. We designed a qualification task that introduced these different errors to annotators and only allowed them to evaluate AGQA's questions once they passed the qualification. The qualification task provided detailed instructions on our interface. The annotators were given sev-

eral examples representing different categories of questions and asked to complete the task. If they did not provide the correct answers in the qualification task, we gave explanations for why their given answers were wrong and did not allow them to proceed until they changed the answer to be correct.

Human evaluation tasks: To validate the correctness of our question-answer generation process and determine the percent of questions in of our dataset that include these errors, we run human validation tasks on Amazon Mechanical Turk and pay $\$15$ USD per hour. As it is infeasible to verify all $192M$ questions in AGQA, we randomly sampled a subset of questions such that there are at least 50 questions per reasoning, semantic, and structural category.

Since the videos are filmed in peoples' homes, they often contain objects and actions outside of the benchmark's vocabulary. Therefore, we indicate which objects and actions are relevant. However, when we asked for free form answers to our questions, annotators gave answers outside of the model's vocabulary. Therefore, we tested human accuracy with a verification task. To provide more insight on which questions require high cognitive effort to answer, we also ran a task in which annotators select the answer from a dropdown menu. We develop our interfaces (see Figure 6) using EasyTurk [2].

For each task, annotators answered one question for each of 5 videos. To improve annotator quality, we ran an qualification task which prevented annotators from proceeding until they placed the correct answer. To ensure annotators answered the questions under the same set of assumptions as AGQA, we added notes next to questions to clarify terms, if necessary. We crowdsourced 3 instances of each question and counted the majority vote.

Verification Task: The verification task showed annotators a video, a question, and a potential answer (Figure 6). They marked the answer as Correct or Incorrect. If they marked the answer as incorrect, we asked them to write a better answer in a textbox. To gather more data, we also asked them to select if the question had bad grammar, multiple answers, or no possible answer. Finally, we added question-answer pairs we knew to be incorrect as a gold standard and to introduce variety. Annotators marked as incorrect 80% of the examples we deliberately made incorrect.

Multiple Choice Task: The multiple choice task showed annotators a video, a question, and a dropdown list of potential answers selected from the events in the video (Figure 6). We also allowed them to select if the question had bad grammar or if they felt unsure of the answer. We judged a annotator's response as correct if their choice from the dropdown menu matched our answer.

Task design effect on results: As the purpose of the human error analysis is to determine which of the questions in AGQA are correct, we used the verification task's results in the main paper. On both tasks human accuracy levels remained consistent as the number of compositional steps increased (Figure 7). However, across nearly every category, performance decreased when people were asked to select the question from a dropdown menu (Table 10). Performance decreased more for open-ended questions. This decrease could originate from the higher cognitive load it takes for people to generate the answer or from AGQA answers that are correct but ambiguous. The activity recognition category is especially difficult within the dropdown task. It has high cognitive load because there are on average 7.4 possible answers in the dropdown menu, and the beginning and endpoints of actions may be ambiguous. Both tasks served to illuminate the source of errors in AGQA that we have described.

Recommendations for future dataset annotation projects: AGQA generates questions referring to specific

Table 10. We run two tasks on human workers, a verification task in which they verify given answers, and a dropdown task in which they select the answer from a dropdown list. Workers perform better on the verification task, especially on open-ended questions.

Question Types			Verification	Dropdown
Reasoning	obj-rel	B	78.95	68.42
		O	90.90	63.64
		All	80.65	67.74
	rel-action	B	90.20	78.43
	obj-act	B	93.75	83.33
	superlative	B	81.81	72.73
		O	80.77	55.77
		All	81.25	63.54
	sequencing	B	94.73	78.94
		O	85.18	59.26
		All	90.77	70.77
	exists	B	79.80	74.03
	duration	B	91.89	70.27
		O	92.31	69.23
		All	92.00	70.00
	activity recognition	O	78.00	54.00
Semantic	object	B	87.39	74.19
		O	90.90	60.52
		All	87.97	72.93
	relationship	B	83.58	75.37
	action	B	90.21	73.91
		O	80.95	57.14
		All	86.45	67.10
Structure	query	O	83.53	58.82
	compare	B	92.53	78.16
	choose	B	83.02	66.04
	logic	B	70.69	70.69
	verify	B	88.26	76.93
	Overall	B	86.65	73.85
		O	83.53	57.93
		All	86.02	71.56

details in the video. This specificity creates challenging questions that inform us about the weaknesses of existing video understanding models. However, our question generation approach relies on the details of the scene graph and a thorough representation that is difficult and expensive to achieve.

We present several recommendations for annotation practices of scene graph representations of videos that would help address the above errors. First, we suggest that annotators cover the entire video in order to avoid small actions occurring before or after annotations. Second, the time stamps of action annotations should be sequenced in terms of global context to avoid the overlapping of actions that actually occur in sequence. Third, annotators should have explicit definitions of ambiguous concepts; e.g. spatial relationships like "above" should be clearly annotated with respect to the camera or with respect to the subject. Finally, an ideal representation should avoid polysemy, even if that object can be referred to with multiple terms. For example, in Action Genome a sandwich is often annotated as both a sandwich and as food. Even though they refer to the same object in the video, they provided different bounding boxes and appeared on non-identical sets of frames. A representation with one annotation per object that has hierarchical levels of semantic specificity would ameliorate this issue.

As future work continues to improve the symbolic representation of videos, benchmarks will be better able to measure detailed video understanding.

6.8. Conclusion

Despite the challenges outlined in the supplementary materials, our pipeline produced a large balanced dataset of video-question answer pairs that requires complex spatio-temporal reasoning. Our dataset is challenging, as the state of the art models barely improved over models using only linguistic features. We also contribute three new metrics that measure a model's ability to generalize to novel compositions, indirect references, and more compositional steps. Current state of the art models struggle to generalize on all of these tasks. Furthermore, although humans perform similarly on both simple and complex questions, models' performance decreased as question complexity increased.

Our benchmark can determine the relative strengths and weaknesses of models on different types of reasoning skills and opens avenues to explore new types of models that can more effectively perform compositional reasoning.

References

[1] 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. 2

[2] Ranjay Krishna. Easyturk: A wrapper for custom amt tasks. https://github.com/ranjaykrishna/easyturk, 2019. 14

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

[4] 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. 2