Measuring Compositional Consistency for Video Question Answering

Mona Gandhi[1]*, Mustafa Omer Gul[2]*, Eva Prakash[2], Madeleine Grunde-McLaughlin[3],
Ranjay Krishna[3], Maneesh Agrawala[2]
Veermata Jijabai Technological Institute[1], Stanford University[2], University of Washington[3]
{mbgandhi_b18}@ce.vjti.ac.in, {momergul, eprakash, maneesh}@stanford.edu,
{mgrunde, ranjaykrishna}@cs.washington.edu

Abstract

Recent video question answering benchmarks indicate that state-of-the-art models struggle to answer compositional questions. However, it remains unclear which types of compositional reasoning cause models to mispredict. Furthermore, it is difficult to discern whether models arrive at answers using compositional reasoning or by leveraging data biases. In this paper, we develop a question decomposition engine that programmatically deconstructs a compositional question into a directed acyclic graph of sub-questions. The graph is designed such that each parent question is a composition of its children. We present AGQA-Decomp, a benchmark containing $2.3M$ question graphs, with an average of 11.49 sub-questions per graph, and $4.55M$ total new sub-questions. Using question graphs, we evaluate three state-of-the-art models with a suite of novel compositional consistency metrics. We find that models either cannot reason correctly through most compositions or are reliant on incorrect reasoning to reach answers, frequently contradicting themselves or achieving high accuracies when failing at intermediate reasoning steps.

1. Introduction

Compositional reasoning is fundamental to how humans represent visual events [20, 26, 32, 38]. For instance, Figure 1 visualizes a video consisting of actions such as taking a picture and holding a bottle; the action holding a bottle involves an actor initially twisting the bottle and then later holding it. This ability to compose interactions and actions is reflected in the compositional nature of language people use to communicate about what they see [5,28]. To measure compositional reasoning of visual events, the computer vision community has proposed multiple video benchmarks using question answering [12, 23, 39]. These benchmarks ask questions such as "Is a phone the first object that the person is touching after taking a picture?", where models need to compose actions (taking a picture) with relationships (touching) and objects (phone) to arrive at the correct answer. Using these benchmarks, researchers have recently concluded that state-of-the-art models [8,22,25] struggle to reason compositionally [12].

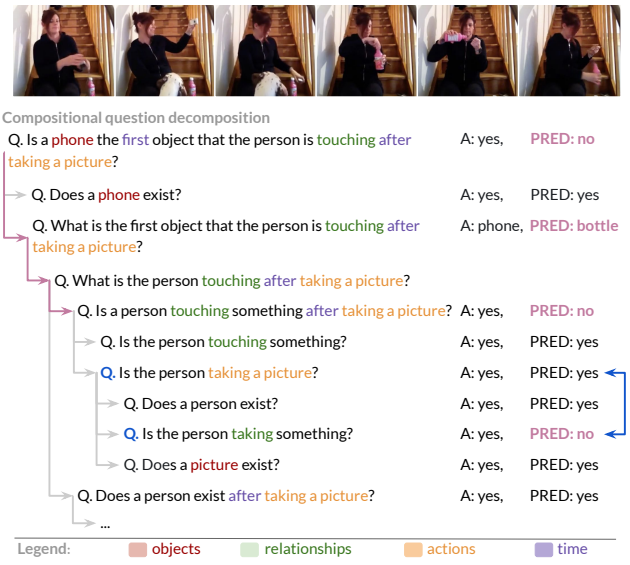

Figure 1. We introduce a question decomposition engine, which produces a DAG of sub-questions from a compositional question about visual events. A sub-question is designed to contain a subset of the original question's reasoning steps. Our engine produces a benchmark with $4.55M$ question answer pairs associated with $9.6K$ videos. We design handcrafted programs and templates for each sub-question as well as composition rules to compose sub-questions together. We analyze existing models using a suite of new compositional consistency metrics using our DAGs. Our DAGs isolate which composition rules cause mispredictions (error path is shown by pink arrows). They also highlight scenarios where models might exhibit self-contradiction (blue arrows).

Unfortunately, existing benchmarks are unable to explain *why* video question answering models struggle with

*Equal contribution

Table 1. We visualize our hand-designed sub-questions, which consist of a subset of the reasoning steps found in the AGQA benchmark [12]. Each sub-question consists of a functional program and a natural language template. Parentheses indicate further categorization.

Sub-question type	Description	Example
`Object exists`	To verify if an object exists	Does a doorway exist?
`Relation exists`	To verify if a relationship exists	Is the person holding something?
`Interaction`	To verify if there is a particular relationship between person and an object	Is the person touching a dish?
`Interaction temporal loc.`	A filter on an interaction type question	Is the person holding a book while smiling at something?
`Exists temporal loc.`	A condition on object/relationship exists question	Does a phone exist after looking in the mirror?
`First/last`	Getting the first/last instance of the given object	What is the first object that the person is above before walking through the doorway?
`Longest shortest action`	Getting the longest/shortest action	What does the person do for the longest amount of time?
`Conjunction`	Get a new exists question by combining two interaction questions with a conjunction	Is the person in front of the mirror and behind the table while looking in the mirror?
`Choose`	Compares between two objects, actions, relationships, or time lengths	Is the doorknob or the dish the first object that the person is holding?
`Equals`	Compares two objects and verifies if they are the same Verifies if the given action is longer/shorter than the other one	Is the doorway the object they are interacting with while holding a dish?

compositional reasoning. In Figure 1, a model incorrectly answers the root question as "no" instead of the correct answer of "yes." However, this information does not explain what caused the model to err: Did the model struggle with words requiring temporal reasoning, such as first or after? Did it fail at detecting the phone or identifying the relationship touching? Or did it struggle to compose the relationship with the object? Even if we assume the model had correctly answered the question, it remains uncertain whether this behavior was due to proper compositional reasoning or a reliance on spurious correlations to "cheat."

Not only do standard evaluation schemes fall short in this regard, but existing approaches for dissecting model behavior also struggle to resolve this uncertainty. Attribution methods, such as GradCAM [35] or LIME [34], can highlight important aspects of the input data, but are agnostic to the structure of compositional reasoning. Approaches that rely on counterfactuals to illuminate model behavior, such as contrast sets [9], focus primarily on model decision boundaries by performing minor, local changes to the input. These local changes, however, cannot capture the full range of compositional reasoning steps required to answer compositional visual questions [12], which assess multiple, often interdependent, reasoning abilities at once.

In this paper, we develop a question decomposition engine that decomposes a compositional question into a directed acyclic graph (DAG) of sub-questions (see Figure 1). A sub-question isolates a subset of the reasoning steps that the original question requires, exposing model performance on subsets of intermediate reasoning steps. This exposure enables us to identify difficult sub-questions and study which compositions cause models to struggle. It also allows us to test whether models are right for the right reasons. For instance, the root question mentioned earlier can not only decompose into intermediate reasoning steps that determine if the "the person was touching something after taking a picture," but also isolate basic perception capabilities, such as determining whether a "phone exists".

Using our engine, we construct the AGQA-Decomp dataset[1], which decomposes the $2.3M$ compositional questions in the updated version[2] of the recent balanced AGQA benchmark [12] to produce $1.62M$ unique sub-questions for $9.6K$ videos for a total of $4.55M$ sub-questions. To generate sub-questions, we hand-design 21 sub-questions, each with a functional program and natural language template (Table 1). To compose the sub-questions within a DAG, we hand-design 13 composition rules (Table 2). Finally, we create a suite of new metrics to evaluate compositional reasoning. One of those metrics — internal consistency — measures whether models are self-consistent when they answer questions within a DAG. To enable this metric, we further hand-design 10 consistency rules between sub-questions (see Table 5 in the Supplementary).

We evaluate three state-of-the-art video question answering models, HCRN [22], HME [8] and PSAC [25] using our DAGs and metrics. Our analyses reveal that for a majority of compositional reasoning steps, models either fail to successfully complete the step or rely on faulty reasoning mechanisms. They frequently contradict themselves and achieve high accuracies even when failing at intermediate steps. Models even struggle when asked to choose between or compare two options, such as objects or relationships. Finally, we find that for HCRN and PSAC, there is no correlation between internal consistency and accuracy across

[1]Project page: https://tinyurl.com/agqa-decomp
[2]AGQA 2.0: https://tinyurl.com/agqavideo

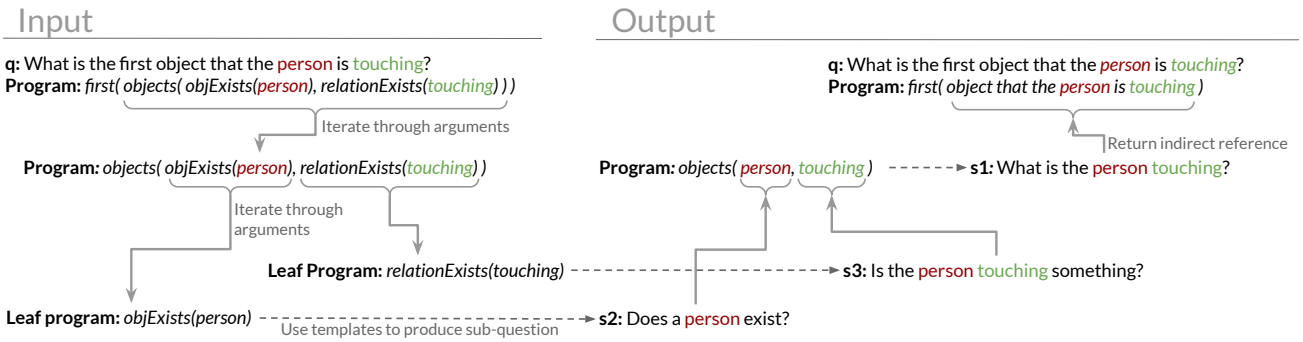

Figure 2. Our question decomposition engine expects a compositional root question as input and outputs a DAG of sub-questions. The root question has an associated functional program which explains the reasoning steps necessary to answer the question. We recursively iterate over the arguments of the function until we reach a leaf function. We design natural language templates for each leaf function, converting them into sub-questions. Once a leaf function is converted to a question, we return an indirect reference of the answer back to its parent. The parent uses composition rules to combine the indirect references from its children to similarly generate questions.

DAGs. For HME, however, there is a weak negative correlation, suggesting that the model is frequently inaccurate and propagates this inaccuracy due to its internal consistency. We believe that our decomposed question DAGs could further enable a host of future research directions: from promoting transparency through consistency to developing interactive model analysis tools.

2. Related Work

We contrast our contributions against recently proposed evaluation measures in machine learning, focusing especially on video question answering. We also contextualize the idea of question decomposition to related work in computer vision and in natural language processing (NLP).

Video question answering. Despite the popularity of video question answering as a benchmark task [10, 12, 15, 23, 39, 46, 47], questions in several prominent benchmarks rely on dialogue and plot summaries instead of a video's visual contents [18, 23, 39, 49], focus on short video clips or only a handful of objects [29, 45], or suffer from biases associated with human generated questions [15, 23, 39, 47]. These limitations reduce benchmarks' effectiveness at reasoning over compositional visual events. Given these limitations, we focus on the recent AGQA benchmark [12] of question answer pairs for compositional visual reasoning.

Evaluating consistency. Our focus on providing an evaluation metric beyond standard task accuracy is in line with recent efforts toward more metamorphic evaluation of machine learning models [2, 9, 24]. While we may be the only method to date proposing a consistency-based metric for video question answering, the role of consistency has been explored for image question answering [2, 11, 14, 31, 33, 36, 37, 48] and for text question answering [9, 43]. Existing metrics measure whether models can consistently answer sets of questions logically entailed by a given question [11, 14, 31, 33] or answer counterfactuals with different answers [9, 43]. To enable these metrics, researchers have collected datasets by asking human annotators to generate perceptual questions associated with reasoning questions [36], used large language models to generate counterfactuals [43], or asked domain experts to compile rules to generate contrast sets [9]. In comparison, we programmatically decompose questions by hand-designing composition rules over programs associated with questions.

Decomposing question answering. Decomposing the question answering task into simpler tasks has appeared within both the computer vision [1, 3] and NLP communities [42]. Most prominently in computer vision, neural module networks and related architectures [1, 4, 13] break down questions into modular programs defining the architecture of the neural network instantiated to answer the question. To design modular architectures, ACMN [3] decomposes questions using dependency parses. The GQA [14] and AGQA [12] benchmarks use programs associated with each question to compute answers from scene graphs [19] and spatio-temporal scene graphs [16]; however, these programs are unused beyond dataset generation.

In NLP, "multi-hop" reasoning questions are decomposed into "single-hop" ones (e.g. decomposing "Which team does the player named 2015 Diamond Head Classic's MVP play for?" into the simpler "Which player was named 2015 Diamond Head Classic's MVP?"). Multi-hop models answer simpler questions and combine their answers to ultimately answer the original multi-hop question [27, 30]. In a similar vein, explanation methods have decomposed language statements into tree-structured sets of premises that entail the original statement (e.g. "eruptions block sunlight" entails "eruptions can kill plants") [7]. While Break-ItDown [42] decomposes questions for HotPotQA [44] into programs to design neural architectures, we decompose questions to design evaluation metrics.

Table 2. We hand-design composition rules to generate questions q using indirect references produced by its sub-questions $\{s_1, s_2, \ldots\}$.

Composition rules	Description	Example
`Interaction`	Verify if an interaction exists	q : Is a person holding a doorway? *s1*: Does a person exist? *s2*: Is a person holding something? *s3*: Does a doorway exist?
`Temporal loc.` `(After, before, while, between)`	Combine two interaction or exists questions using a temporal localizer	q: Is the person touching a doorway before smiling at something? *s1*: Is the person touching a doorway? *s2*: Is a person smiling at something?
`First/last`	Getting the first/last occurrence from a set of object/actions	q: What is the first object that the person is holding? *s1*: What is the person holding?
`Conjunction` `(And, xor)`	Combine two interaction questions using a conjunction	q: Is the person putting some clothes and behind a book before walking through the doorway? *s1*: Is the person putting some clothes before walking through the doorway? *s2*: Is the person behind a book before walking through the doorway?

`Choose`
`(Choose (object/Time)`
`longer/shorter choose)`
	Chooses one of two possible options	q: Is the doorway or the book the first object they were in front of?
		s1: Is the doorway the first object they were in front of?
		s2: Is the book the first object they were in front of?

`Equals`	Compares two objects/actions to verify if they are the same	q: Is a book the first object that the person is carrying? *s1*: Does a book exist? *s2*: What is the first object that the person is carrying?

Compositional reasoning. While multiple definitions of compositionality exist, we 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" [6]. In our case, reasoning about the question "Was the person holding a bottle after touching a phone?" entails being able to answer simpler questions (e.g. "Did the person touch a phone?"), which can be further decomposed into perceptual questions (e.g. , "Does a phone exist?") and spatio-temporal relationship detection (e.g. "Did the person touch something?"). Recent work has argued the importance of compositionality in enabling models to generalize to new domains, categories, and logical rules [21, 40] and has discovered that current models struggle with multi-step reasoning [8, 12]. These studies motivate our contribution.

3. Question decomposition engine

Given a question q as input, our engine outputs a directed acyclic graph (DAG) $(N_q, E_q) \in G_q$ of sub-questions for that question. The nodes N_q are the list of sub-questions for question q while the directed edges identify the composition rule used to compose a question from a node's sub-questions. For example, the decomposition of "What is the first object that the person is touching?" will produce the following list of sub-questions: $\{s1$: "What is the person touching?", $s2$: "Does a person exist?", and $s3$: "Is the person touching something?" $\}$. The edges are: $\{(q, s1, \text{first}), (s1, s2, \text{interaction}), (s1, s3, \text{interaction})\}$, where "first" and "interaction" are composition rules.

To generate the DAG, we first represent the question q as a functional program, which consists of the individual reasoning steps needed to answer q. The program structure defines the structure of the DAG (as shown in Figure 2). We recursively iterate over this program and its arguments to generate the DAG.

While our composition rules and templates are tailored towards AGQA [12], our engine can be generalized to other datasets involving questions paired with functional programs, such as GQA [14], CLEVR [17] or CLEVRER [45]. This will require defining composition rules and templates based on the datasets' function programs.

3.1. Representing questions as programs

We assume all questions have a corresponding functional program, with multiple reasoning steps. For instance, the program for q is `first(objects(objExists(person), relationExists(touching)))`. Intuitively, this particular program searches through all the frames of a given video to find instances where there is a person present: `objExists(person)`. Similarly, it finds the frames where a person is touching something: `relationExists(touching)`. From those frames, it extracts the objects that are being touched by a person: `objects(objExists(person), relationExists(touching))`. Finally, it returns the first object from the list of objects identified: `first(·)`.

Each reasoning step is a function composed of multiple arguments: For example, the function `objects(·)` contains the following arguments: `objExists(·)` and `relationExists(·)`. We utilize the $2.3M$ questions, each generated using 27 unique functions associated with 217 natural language templates, in AGQA.

3.2. Decomposing questions using programs

To decompose q, we topologically iterate over all the arguments of the top-level reasoning function and recursively decompose each argument. For instance, the top level reasoning function for q is `first(·)`. We iterate over its argument `objects(·)` and then recursively iterate over its two arguments: `objExists(·)` and `relationExists(·)`.

Eventually, we will arrive at a "leaf" program with no further functions as arguments (e.g.

`objExists(person)`). To convert the leaf program into a node in the DAG, we design natural language question templates for every program (see Table 1). For instance, `objExists(·)` has the template: "Does an [object] exist?" that creates the subquestion $s2$. We check if we have already added $s2 \in N_q$ while traversing another argument. If $s2 \notin N_q$, then we use the template to create a new node $s2$ = "Does a person exist?" and add it to N_q.

Once we convert a leaf function into $s2$, we parse the template to extract an indirect reference and send it back to its parent function. The parent function, in this case `objects(objExists(person), relationExists(touching))` uses its arguments $s2$ and $s3$, along with a compositionality rule to produce the node $s1$ = "What is the person touching?". We design a set of compositionality rules, listed in Table 2, to ingest the indirect references passed back ($s2 \rightarrow$ "person" and $s3 \rightarrow$ "touching") into its corresponding template: ""What is the [object] [relationship]?". Next, we add the edges between $s1$ and its two arguments to E_q with the composition rule, `interaction`, used to compose the arguments together. This process continues until we return back to the original top-level function `first(·)`.

Our recursive decomposition process makes an average of 11.49 sub-questions for each of the $2.3M$ questions in the balanced AGQA questions, creating $4.55M$ sub-questions.

3.3. AGQA answer generation

Once all the questions are decomposed into DAGs of sub-questions, we programmatically propagate answers from the original AGQA questions to the sub-questions. Some sub-questions are already present in the original unbalanced AGQA dataset; for these, we automatically have the answers. For others, we craft logical consistency rules to generate answers (see Table 5 in the Supplementary).

For example, if the answer to an `Interaction` question is "yes", then all its sub-questions should also be answered "yes". If the answer to "Is the person touching something?" is "yes," for instance, then the answer to "Does a person exist?" is also "yes". If a "choose X or Y" question's answer is "X", then all sub-questions along X's recursive call should be answered "yes," while Y's answer should be "no." If, for example, "Did the person throw the blanket but not hold the blanket?" is answered "yes", then the answer to "Did the person throw the blanket?" is "yes" but "Did the person hold the blanket" is "no". Similar logical rules apply for `Before` and `After` question types.

Our answer generation rules are unable to propagate answers for questions answered "no". For instance, if the answer to "Is the person touching something?" is "no", we can not entail an answer to the question "Does a person exist?". To answer such questions, we run a large-scale annotation task on Amazon Mechanical Turk to identify all objects that appear in a randomly selected subset of videos in AGQA (see Supplementary for details). We use these annotations to propagate "no" answers to the relevant sub-questions.

Finally, we balance the answer distribution to arrive at our final dataset. When generating AGQA's original balanced dataset, the authors used an answer smoothing algorithm to mitigate biases in the training process. Adding our sub-questions to AGQA changes the training answer distributions. To reduce the bias in the new answer distributions, we adopt the same answer smoothing algorithm. This process results in $1.62M$ unique new sub-questions across the dataset, and a total of $4.55M$ sub-questions.

4. Metrics

Using the sub-question types and composition rules we handcrafted, we design novel metrics that measure models' compositional accuracy, test whether models are right for the wrong reasons, and identify whether models are internally consistent. Our metrics are complementary and should be used together to guide error analysis. Formal definitions for the metrics can be found in the Supplementary.

Compositional accuracy (CA): A model reasoning compositionally should be able to answer a given parent question q correctly when it answers its sub-questions correctly. We operationalize this intuition with the **CA** metric, which measures parent question accuracy across compositions where a model answers all immediate sub-questions correctly. Low CA scores for a given category indicate difficulty performing that intermediate reasoning step.

Right for the wrong reasons (RWR): Given that the sub-questions of a given question q represent intermediate reasoning steps, a model reasoning compositionally should answer all sub-questions correctly if it answers q correctly. Failure to do so implies the model is relying on faulty decision mechanisms to reach correct answers. The **RWR** metric aims to determine to what extent such faulty reasoning occurs. To compute this, we measure parent question accuracy across compositions where a model answers at least one sub-question incorrectly. High RWR scores for a given category imply that the model's reasoning is faulty for those intermediate steps. For granularity, we additionally compute parent question accuracies across compositions where a model answers exactly n sub-questions incorrectly, where n is an integer. We denote this variant **RWR-n** and present its results in the Supplementary (Tables 6, 7).

Delta: We derive additional insights by computing the difference between **RWR** and **CA** values. Ideally, **RWR** will be lower than **CA**, leading to negative **Delta** values. A positive **Delta** value implies incorrect reasoning since the model performs better when it errs on a sub-question.

Internal Consistency (IC): A model that reasons compositionally should produce answers that don't contradict each other, regardless of accuracy. Unlike most past work on

Table 3. We report accuracy, compositional accuracy (**CA**), right for the wrong reasons (**RWR**), delta (**RWR-CA**) and internal consistency (**IC**) values. We also present accuracy for the Most-Likely baseline and the rate at which annotators agreed with ground-truth answers in our AMT study (Human). Models particularly struggle at `Interaction Temporal Localization`, `Choose` and `Equals` questions as well as basic question types such as `Object Exists`. N/A indicates there were no valid compositions for a given type.

	Accuracy				CA			RWR			Delta			IC			
Question Type	HCRN	HME	PSAC	Most-Likely	HCRN	HME	PSAC	HCRN	HME	PSAC	HCRN	HME	PSAC	HCRN	HME	PSAC	Human
`Object Exists`	47.03	46.74	45.02	50.00	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	92.00
`Relation Exists`	52.14	51.21	36.44	50.00	73.17	8.99	N/A	16.67	N/A	20.22	-56.50	N/A	N/A	82.87	100.00	50.00	92.00
`Interaction`	46.71	50.57	62.33	50.00	62.50	32.66	N/A	33.31	23.58	48.63	-29.19	-9.08	N/A	84.87	62.66	50.00	88.00
`Interaction Temporal Loc.`	49.53	50.43	45.20	50.00	57.82	57.96	3.91	47.39	50.46	46.92	-10.43	-7.51	43.01	77.47	62.56	61.49	96.00
`Exists Temporal Loc.`	47.82	49.69	53.52	50.00	90.92	22.60	67.68	45.44	1.96	18.69	-45.49	-20.64	-48.99	74.19	76.36	77.05	92.00
`First/Last`	9.28	12.31	8.20	3.79	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	88.00
`Longest/Shortest Action`	3.24	1.67	1.58	3.57	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	76.00
`Conjunction`	49.60	50.07	50.01	50.00	71.64	85.26	85.81	42.19	39.85	39.92	-29.45	-45.42	-45.89	64.00	65.36	54.34	76.00
`Choose`	24.44	35.16	26.03	1.89	51.19	55.24	46.49	47.05	48.28	48.09	-4.14	-6.96	1.59	5.75	0.65	12.18	88.00
`Equals`	50.53	50.08	49.92	50.00	47.71	52.88	49.00	51.67	47.15	50.36	3.96	-5.72	1.35	47.56	51.03	47.61	70.00
Overall	21.27	30.47	21.29	3.31	74.59	49.28	60.97	46.22	25.29	36.68	-28.37	-23.99	-24.28	62.88	58.34	57.95	84.36

measuring consistency [14, 33, 36], we can use our logical consistency rules (see Table 5 in the Supplementary) and their contrapositives to determine whether models are self-consistent without access to ground-truth answers. We note that most compositions considered for the **IC** metric have multiple logical consistency rules associated with them. To compute the **IC** metric for a given composition rule, we first measure the percentage of consistency checks a model satisfies for each of its logical consistency rules. Then we average these percentages to obtain the **IC** score for that composition. With this, we avoid overemphasizing a more common rule. **IC** scores for individual logical consistency rules can be found in the Supplementary (Table 8).

Accuracy: To obtain a baseline understanding of model performance, we additionally compute accuracy per question type. To elevate the role of answers on the long tail of the answer distributions, we compute accuracy per ground-truth answer and then normalize across answers.

5. Experiments

We evaluate three state-of-the-art video question answering models on our DAGs to analyze their compositional visual reasoning capability. We start by analyzing model accuracy on leaf nodes testing basic perception. We then analyze three different groups of compositional reasoning steps: `Choose` and `Equals` questions, `Conjunction` questions, and the `Temporal Localization` categories. In these analyses, the CA metric helps determine which reasoning steps models struggle at, the RWR metric checks whether models achieve high accuracies even when failing at intermediate reasoning steps, and the IC metric determines how often models contradict themselves. We additionally cite exact values for RWR-n scores, IC values for individual consistency rules and accuracies per ground-truth answers to support analysis. Full tables for these values can be found in the Supplementary (Tables 6-9).

Models. We use the three models evaluated in the AGQA paper: HME [8], HCRN [22] and PSAC [25]. HME fuses memory modules for visual and question features [8], HCRN creates a multi-layer hierarchy of a reusable module that integrates motion, question, and visual features at each layer [22] and PSAC integrates visual and language features using positional self-attention and co-attention blocks [25]. Like the AGQA paper, we also consider a model (Most-Likely) that outputs the most common answer for each question type as a baseline relying only on linguistic biases.

Training. We trained models on a version of the AGQA balanced dataset that is augmented with the balanced sub-question DAGs we produced. We stop training when validation accuracy plateaues.

5.1. Human evaluation

To evaluate the quality of the questions and answers our engine generates, we run a human evaluation study. We hire annotators at a rate of $15/hr in accordance with fair work standards on Amazon Mechanical Turk [41]. We present annotators with at least 25 randomly sampled questions per sub-question type and adopt the human evaluation protocol presented in AGQA [12]. Annotators are asked to verify a question and answer pair by watching the video associated with them. The majority vote of 3 annotators per question labeled 84.36% of our answers as correct, implying that about 15.64% of our questions contain errors (see Table 3). These errors originate in scene graph annotation errors and ambiguous relationships. We describe in supplementary materials the sources of human error. To put this number in context, GQA [14], CLEVR [17] and AGQA [12], three recent automated benchmarks, report 89.30%, 92.60%, and 86.02% human accuracy, respectively.

5.2. Performance on Leaf Nodes

Upon inspecting model accuracy (Table 3) on the `Object Exists` and `Relation Exists` categories, we find that each model struggles on basic perceptual questions, casting doubt on good performance on more complex categories. Model accuracy on both categories is either on par with or poorer than the Most-Likely baseline. By investigating model accuracy per-ground truth answer (see

Table 9 in the Supplementary), we find that HME is heavily biased towards "no" answers for Relation Exists, achieving 99.11% and 3.29% accuracy on "no" and "yes" answered questions respectively. PSAC is similarly biased on the Object Exists category, achieving 86.67% and 3.38% accuracy on "no" and "yes" answered questions. HCRN, finally, has near or below-chance performance on both categories, only achieving above 50% accuracy on "No" answered questions of the Relation Exists category with a score of 55.84%.

5.3. Performance on Choose and Equals

Our CA, RWR and IC metrics (Table 4) help demonstrate not only that models struggle at the Choose and Equals categories, but that they also rely on incorrect reasoning for them. Firstly, by looking at the CA scores, we find that even when models answer all child questions correctly, they obtain around or below 50% accuracy for these binary questions. Models particularly struggle at Longer/Shorter Choose compositions. HCRN, HME and PSAC, for instance, obtain 42.02%, 41.90% and 38.51% CA for Longer Choose. Furthermore, models achieve an IC score of at most 12.18% for Choose compositions, providing evidence for incorrect reasoning. Models' reasoning is particularly faulty when the Choose composition requires ordering two events (Table 8), with HCRN, HME and PSAC's predictions being self-consistent only 4.92%, 0.54% and 9.56% of the time for this rule. We can reach a similar conclusion for the Equals composition. HCRN and PSAC have Delta scores of 3.96% and 1.35% respectively, meaning they are better at answering parent questions upon making mistakes at child questions. In contrast, HME obtains a Delta score of −5.72% (Table 4), indicating that errors on intermediate reasoning steps have only a small negative impact on its performance, which shouldn't occur if reasoning compositionally.

5.4. Performance on Conjunctions

Models' inability to reason compositionally largely persists for the logical Conjunction categories. While both HME and PSAC obtain high CA scores (Table 4) for And (95.81% and 88.31%) and Xor (78.91% and 84.32%) compositions, their success stems primarily from their performance when the parent question has "no" as a ground-truth answer. For the CA metric, HME and PSAC predict 41.95% and 37.60% of "yes" answered questions correctly for And compositions and only 1.41% and 14.79% of "yes" answered questions correctly for Xor compositions. Both models obtain approximately 80% RWR-1 performance for And and over 80% RWR-2 performance for Xor compositions (Table 7). Their performance is far above chance when making mistakes on intermediate reasoning steps, indicating that their success on "no" answered questions is not due to an understanding of logical conjunctions. HCRN, however, behaves differently. For Xor, it obtains a poor CA score of 52.33%, which is close to chance. On the other hand, HCRN appears to properly understand the And composition. It achieves a high CA score of 88.49, answering 90.66% of "yes" and 84.52% of "no" answered questions correctly. Its IC score is a similarly high 84.31% (Table 4), where it is internally consistent for 74.37% and 94.24% of consistency checks where the parent is "yes" and "no" respectively (Table 8). While its RWR-1 score of 48.35 (Table 7) casts doubt on whether HCRN has a grounded understanding of what the question asks, its high CA and IC scores nonetheless indicate that it can successfully execute the And reasoning step.

5.5. Performance on Temporal Reasoning

We finally analyze model performances on the Temporal Localization categories, starting with the Exists Temporal Localization question type. We split analysis by the temporal localization composition types: After, Before, While or Between. We first find that HME fails on After, Before and While compositions, obtaining poor CA scores of 30.88%, 31.95% and 24.36% respectively (Table 4). While PSAC and particularly HCRN obtain higher CA scores on these compositions, their success is likely due to faulty reasoning. Both models obtain IC scores less than 50% when answering "yes" to the parent question (Table 8), contradicting themselves over half the time in one common setting. HCRN's above chance RWR-1 scores of 61.24%, 65.16%, 66.01% for these compositions (Table 7) further indicate incorrect reasoning. Model performances on Between compositions, however, are reminiscent of those on And compositions. While HME obtains a high CA score of 94.54%, it achieving an IC score of 42.01% when the parent is "yes" (Table 8) and a high RWR-1 score of 77.83% (Table 7) indicates that this success is due to incorrect reasoning. Meanwhile, HCRN and PSAC achieve high CA scores, do not have RWR values far above chance (Tables 4, 7) and obtain high IC scores of 85.95% and 87.97% respectively. These models can successfully execute the Between reasoning step even if their understandings of the underlying Before and After compositions are suspect. Interaction Temporal Localization, on the other hand, additionally involves an Interaction composition and requires the model to temporally reason about two different relationships or actions. PSAC, given its 3.91% CA score, is incapable of performing this task. HCRN and HME, on the other hand, likely rely on spurious correlations even when they are correct. For instance, while HCRN and HME obtain CA scores of 57.82% and 57.96% respectively (Table 4), they also obtain RWR-2 scores of 55.34% and 93.92% (Table

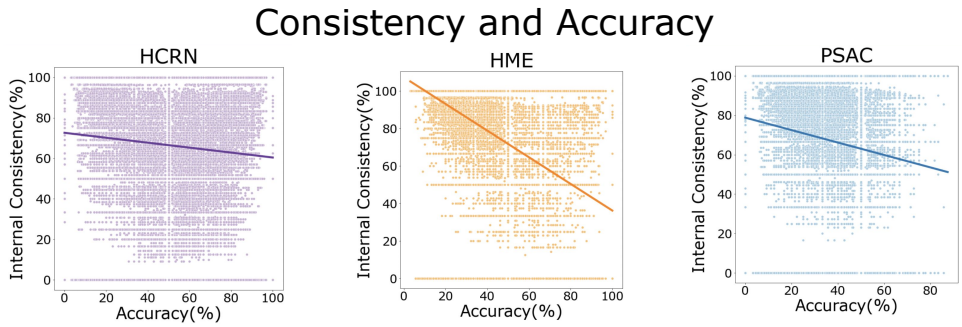

Figure 3. We measure the internal consistency of each DAG using handcrafted consistency rules. HCRN and PSAC do not have a correlation between the internal consistency of a DAG and the accuracy across all its questions while HME has a weak negative correlation (Pearson Correlation Coefficient: -0.086 for HCRN, -0.293 for HME and -0.109 for PSAC).

Table 4. We calculate the compositional accuracy (**CA**), right for the wrong reasons (**RWR**), delta (**RWR-CA**) and internal consistency (**IC**) metrics with respect to composition rules for HCRN, HME and PSAC. We find that models are either unable to reason over a given composition or are right for the wrong reasons, often due to self-contradiction.

	CA			RWR			Delta			IC		
Composition Type	HCRN	HME	PSAC	HCRN	HME	PSAC	HCRN	HME	PSAC	HCRN	HME	PSAC
`Interaction`	58.42	42.09	92.75	40.73	38.85	49.94	-17.70	-3.24	-42.82	87.81	63.63	49.98
`First`	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
`Last`	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
`Equals`	47.71	52.88	49.00	51.67	47.15	50.36	3.96	-5.72	1.35	47.56	51.03	47.62
`And`	88.49	95.81	88.31	34.86	40.79	42.46	-53.63	-55.03	-45.85	84.31	78.36	52.33
`Xor`	52.33	78.91	84.32	49.21	38.76	36.98	-3.12	-40.15	-47.34	43.70	52.36	56.34
`Choose`	52.42	57.02	47.64	47.42	48.58	48.32	-5.00	-8.45	0.68	5.75	0.65	12.18
`Longer Choose`	42.02	41.90	38.51	38.68	41.04	41.00	-3.34	-0.87	2.49	N/A	N/A	N/A
`Shorter Choose`	40.28	50.88	38.86	36.83	41.87	41.44	-3.45	-9.01	2.58	N/A	N/A	N/A
`After`	78.10	30.88	57.34	48.02	22.45	30.00	-30.08	-8.43	-27.34	70.24	70.08	71.17
`Before`	78.49	31.95	58.48	51.93	21.93	28.73	-26.57	-10.02	-29.75	69.24	71.05	71.67
`While`	89.36	24.36	64.53	44.40	9.33	23.88	-44.96	-15.03	-40.66	71.32	66.96	72.33
`Between`	84.80	94.54	89.38	17.37	5.85	12.25	-67.43	-88.69	-77.12	85.95	70.90	87.97
Overall	69.70	51.90	62.29	45.84	27.98	37.82	-23.87	-23.92	-24.47	62.88	58.34	57.95

7), meaning that their performance does not depend on whether they are accurate for intermediate reasoning steps. Models' overall poor performance on `Interaction Temporal Localization` is similar to the performance on `Choose` and `Equals` questions, both of which also require reasoning over two distinct components.

5.6. Correlation between consistency and accuracy

We test whether our IC metric is predictive of model accuracy, as this can aid users at inference time. Specifically, we measure whether IC is correlated with accuracy. To do this, we compute internal consistency on DAGs by measuring the percentage of correct logical consistency checks across all compositions in a DAG and compare against accuracy on the entire DAG. Figure 3 shows that internal consistency either has no correlation with accuracy or a weak negative one, with HCRN, HME and PSAC having correlation coefficients of -0.086, -0.293 and -0.109 respectively. HME's negative correlation can be explained by its consistent bias towards "no" answers (see Table 9 in the Supplementary), which are less frequent in our DAGs as our consistency checks can only propagate "yes" answers. As such, while HME is highly consistent, it is also frequently incorrect, which causes inaccuracies to propagate throughout hierarchies. HCRN and PSAC, on the other hand, are less biased models but are nonetheless often right for the wrong reasons. As they do not reason compositionally, their internal consistency is not predictive of their accuracy.

6. Discussion

In conclusion, we developed a question decomposition engine and generated the dataset AGQA-Decomp hoping to facilitate the analysis of video question answering models beyond average accuracy. Our work is a continuation of a shift in machine learning away from standard accuracy metrics towards more metamorphic evaluation [2, 9, 24]. Our results are bleak: models frequently contradict themselves and are often right for the wrong reasons.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 3

[2] Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. *arXiv preprint arXiv:2103.09591*, 2021. 3, 8

[3] Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, and Liang Lin. Visual question reasoning on general dependency tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7249–7257, 2018. 3

[4] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021. 3

[5] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002. 1

[6] MJ Cresswell. Logics and languages. 1973. 4

[7] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*, 2021. 3

[8] 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/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 1, 2, 4, 6

[9] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020. 2, 3, 8

[10] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions & temporal reasoning. In *International Conference on Learning Representations*, 2020. 3

[11] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020. 3

[12] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4, 6

[13] 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. 3

[14] 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/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3, 4, 6

[15] 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. 3

[16] 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. 3

[17] 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. 4, 6

[18] 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. 3

[19] 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. 3

[20] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008. 1

[21] 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. 4

[22] 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. 1, 2, 6

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 1, 3

[24] Chuanrong Li, Lin Shengshuo, Leo Z Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. Linguistically-informed transformations (lit): A method for automatically generating contrast sets. *arXiv preprint arXiv:2010.08580*, 2020. 3, 8

[25] 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. 1, 2, 6

[26] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally compos-

able human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 812–819, 2014. 1

[27] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. *arXiv preprint arXiv:1906.02916*, 2019. 3

[28] Richard Montague et al. Universal grammar. *1974*, pages 222–46, 1970. 1

[29] 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. 3

[30] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. *arXiv preprint arXiv:2002.09758*, 2020. 3

[31] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019. 3

[32] 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. 1

[33] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, 2019. 3, 6

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 2

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[36] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011, 2020. 3, 6

[37] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 3

[38] 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. 1

[39] 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. 1, 3

[40] 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. 4

[41] 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. 6

[42] Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198, 2020. 3

[43] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. 3

[44] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018. 3

[45] 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. 3, 4

[46] Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. 3

[47] 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. 3

[48] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. Perception matters: Detecting perception failures of vqa models using metamorphic testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16908–16917, 2021. 3

[49] 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. 3