ANetQA: A Large-scale Benchmark for Fine-grained Compositional Reasoning over Untrimmed Videos

Zhou Yu[1] Lixiang Zheng[1] Zhou Zhao[2] Fei Wu[2] Jianping Fan[1,3] Kui Ren[4] Jun Yu[1]*

[1] School of Computer Science, Hangzhou Dianzi University, China.

[2]Colledge of Computer Science and Technology, Zhejiang University, China

[3]AI Lab at Lenovo Research, China

[4]School of Cyber Science and Technology, Zhejiang University, China

{yuz, lxzheng, yujun}@hdu.edu.cn, {zhaozhou, wufei, kuiren}@zju.edu.cn, jfan1@lenovo.com

Abstract

Building benchmarks to systemically analyze different capabilities of video question answering (VideoQA) models is challenging yet crucial. Existing benchmarks often use non-compositional simple questions and suffer from language biases, making it difficult to diagnose model weaknesses incisively. A recent benchmark AGQA [9] poses a promising paradigm to generate QA pairs automatically from pre-annotated scene graphs, enabling it to measure diverse reasoning abilities with granular control. However, its questions have limitations in reasoning about the ***fine-grained*** *semantics in videos as such information is absent in its scene graphs. To this end, we present ANetQA, a large-scale benchmark that supports fine-grained compositional reasoning over the challenging untrimmed videos from ActivityNet [4]. Similar to AGQA, the QA pairs in ANetQA are automatically generated from annotated video scene graphs. The fine-grained properties of ANetQA are reflected in the following: (i) untrimmed videos with fine-grained semantics; (ii) spatio-temporal scene graphs with fine-grained taxonomies; and (iii) diverse questions generated from fine-grained templates. ANetQA attains* ***1.4 billion*** *unbalanced and* ***13.4 million*** *balanced QA pairs, which is an order of magnitude larger than AGQA with a similar number of videos. Comprehensive experiments are performed for state-of-the-art methods. The best model achieves 44.5% accuracy while human performance tops out at 84.5%, leaving sufficient room for improvement.*

1. Introduction

Recent advances in deep learning have enabled machines to tackle complicated video-language tasks that involve both video and language clues, *e.g.*, video-text retrieval, video captioning, video temporal grounding, and video question answering. Among these tasks, video question answering (VideoQA) is one of the most challenging tasks as it verifies multiple skills simultaneously. Taking the question "*What is the black object that the person is wearing before various fish are seen swimming through the reef?*" in Figure 1 as an example, it requires a synergistic understanding of both the video and question, together with spatio-temporal reasoning to predict an accurate answer.

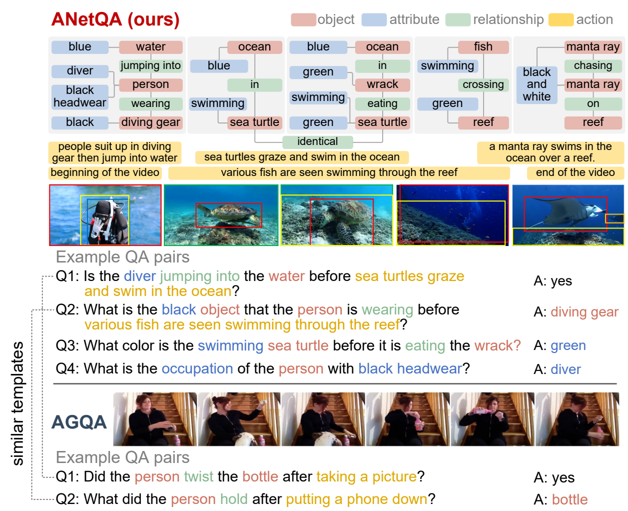


Figure 1. Comparisons of ANetQA and AGQA [9]. The QA pairs in both benchmarks are automatically generated from spatio-temporal scene graphs by using handcrafted question templates. Benefiting from the untrimmed long videos and fine-grained scene graphs, our questions require more fine-grained reasoning abilities than those in AGQA when similar templates are applied. Moreover, the newly introduced attribute annotations allow us to design many fine-grained question templates that are not supported in AGQA (*e.g.*, "*what color*" and "*what is the occupation*").

*Jun Yu is the corresponding author

To comprehensively evaluate the capabilities of existing VideoQA models, several prominent benchmarks have been established [12, 22, 31, 35, 40, 44, 45]. Despite their usefulness, they also have distinct shortcomings. Some benchmarks use simulated environments to synthesize video contents [31, 44], which provides controllable diagnostics over different reasoning skills. However, the synthetic videos lack visual diversity and the learned models on the benchmarks cannot generalize to real-world scenarios directly. Some real-world benchmarks generate QA pairs from off-the-shelf video captions [40, 50] or human annotations [12, 22, 35, 45], which suffer from simple question expressions and biased answer distributions. These weaknesses may be exploited by models to make educated guesses to obtain the correct answers without seeing video contents [25, 42].

One recent VideoQA benchmark AGQA poses a promising paradigm to address the above limitations [9]. AGQA is built upon the real-world videos from Charades [34]. In contrast to previous benchmarks, AGQA adopts a two-stage paradigm instead. For each video, a spatio-temporal scene graph over representative frames is first annotated by humans, which consists of spatially-grounded object-relationship triplets and temporally-grounded actions. After that, different types of questions are generated on top of the scene graph using corresponding question templates, enabling it to measure various reasoning abilities with granular control. Despite the comprehensiveness of AGQA, we argue that its foundation—the spatio-temporal scene graph—has limitations in representing the *fine-grained* semantics of videos. Specifically, their scene graphs encode objects and relationships from limited taxonomies, which are *not* fine-grained enough for generating questions that require reasoning about the detailed video semantics.

To this end, we introduce ANetQA[1], a new benchmark that supports fine-grained compositional reasoning over complex web videos from ActivityNet [4]. Similar to the strategy of AGQA, the QA pairs in ANetQA are automatically generated from pre-annotated scene graphs. As shown in Figure 1, we claim that ANetQA is more fine-grained than AGQA in terms of the following:

(i) The benchmark is built upon untrimmed long videos with fine-grained semantics. Each video may involves multiple indoor or outdoor scenarios, containing complicated interactions between persons and objects.

(ii) The spatio-temporal scene graph consists of fine-grained objects (*e.g.*, "*manta ray*", "*diving gear*"), relationships (*e.g.*, "*jumping into*", "*chasing*"), attributes (*e.g.*, "*swimming*", "*black and white*"), and actions in natural language (*e.g.*, "*a manta ray swims in the ocean over a reef*").

(iii) Benefiting from the fine-grained scene graphs, we are able to design diverse question templates that requires fine-grained compositional reasoning (*e.g.*, "*what color ...*" and "*what is the occupation ...*").

Benefiting from the above fine-grained characteristics, ANetQA obtains 1.4B unbalanced and 13.4M balanced QA pairs. To the best of our knowledge, ANetQA is the largest VideoQA benchmark in terms of the number of questions. Compared with the previous largest benchmark AGQA, ANetQA is an order of magnitude larger than it with a similar number of videos. We conduct comprehensive experiments and intensive analyses on ANetQA for the state-of-the-art VideoQA models, including HCRN [20], ClipBERT [21], and All-in-One [37]. The best model delivers 44.5% accuracy while human performance tops out at 84.5%, showing sufficient room for future improvement. The benchmark is available at here[2].

2. Related Work

We briefly review the field of VideoQA in terms of methods and benchmarks. Since ANetQA is built upon ActivityNet [4], we introduce ActivityNet and its derived benchmarks in particular.

VideoQA approaches. The research of visual question answering lies mainly in the image domain. A number of image question answering (ImageQA) methods have been developed to push state-of-the-art performance on public benchmarks successively [6, 14, 46, 47]. As a natural extension of the ImageQA task, VideoQA is more challenging as it requires effective temporal representation modeling and spatio-temporal reasoning. Existing studies explore end-to-end neural networks in conjunction with hierarchical representations [40, 51], memory networks [7, 32, 35], and graph networks [10, 26, 39]. Motivated by the encouraging success of Transformers [36] in various NLP [16, 33], CV [3, 27], and multimodal tasks [1, 2, 28], Transformer-based approaches have become the mainstream of recent VideoQA research. Early approaches only exploit the Transformer architecture and train models from scratch [15, 24]. More recently, pretrained Transformer models on large-scale datasets have shown effectiveness when fine-tuned on VideoQA tasks. Some approaches incorporate the pretrained language Transformers [17, 43] or multimodal Transformers on image-text pairs [21] to improve VideoQA performance. Some other studies perform video-language pretraining directly on massive video-text pairs, which learn better multimodal representations and achieve state-of-the-art performance on various VideoQA benchmarks [5, 37, 41, 49].

VideoQA benchmarks. The rapid progress in VideoQA is inextricably related to the established benchmarks. Existing

[1] Note that there is a VideoQA benchmark ActivityNet-QA [45] whose QA pairs are fully annotated by humans. To avoid confusion, we name our benchmark ANetQA.

[2] https://milvlg.github.io/anetqa

	video			question		grounding taxonomy			
	type	#videos	avg. len.	#QA pairs	#templates	#objects	#relations	#attributes	#actions
CLEVRER [44]	synth.	20K	5s	305K	5	1	2	13	3
TVQA+ [23]	real	4.2K	7.2s	29.4K	-	**2,527**	-	-	open
HowtoVQA69M [41]	real	**69M**	12.1s	69M	-	-	-	-	open
AGQA [9]	real	9.6K	30s	192M/3.9M	28	36	44	-	157
ANetQA	real	11.5K	**180s**	**1.4B/13.4M**	**119**	2,072	**86**	**618**	open

Table 1. **Comparisons of ANetQA and other representative large-scale VideoQA benchmarks**. Benefiting from the fine-grained video and grounding annotations, ANetQA attains massive fine-grained questions and is an order of magnitude larger than the current largest benchmarks [9, 41] in terms of the number of QA pairs. "open" indicates the grounded actions are depicted in natural language.

VideoQA benchmarks can be categorized into two groups based on whether their videos are synthesized by simulation [31, 44] or collected from the real world [9, 18, 22, 29, 35, 38, 40, 41, 45, 48, 50]. The synthesized benchmarks can easily obtain massive QA pairs without human annotations. Their synthetic nature also enables granular control over reasoning abilities and language biases. However, the synthesized videos are often short and lack visual diversity, making it difficult to generalize the learned models to real-world scenarios.

Establishing VideoQA benchmarks on real-world videos requires human annotations inevitably. Early benchmarks rely on the associated video captions to generate QA pairs automatically [29, 40, 50, 53]. Although these captions are annotated by humans, they are often too general to cover all the fine-grained semantics in videos. This makes these benchmarks be dominated by simple questions that lack detailed information. To obtain fine-grained and diverse questions, some recent benchmarks have been established by asking annotators to design questions of specific reasoning abilities, *e.g.*, object localization [23], relationship recognition [45], and causality analysis [38]. Nevertheless, prohibitive annotation costs restrict the sizes of these benchmarks and free-form question expressions lead to severe language biases. One recent benchmark AGQA introduces a new paradigm to automatically generate QA pairs upon video scene graphs [9]. Through the composition of scene graph elements, AGQA is orders of magnitude larger than its counterparts. Similar to AGQA, our ANetQA is also built upon spatio-temporal scene graphs. In contrast to AGQA, ANetQA shows its fine-grained characteristics in terms of the videos, annotated scene graphs, and generated questions. Detailed comparisons of ANetQA and other representative large-scale VideoQA benchmarks are shown in Table 1.

ActivityNet and its derivatives. ActivityNet (*abbr.* ANet) is one of the most important video recognition benchmarks [4]. It consists of 20K untrimmed videos from 200 activity classes, including both indoor and outdoor scenarios. The benchmark is challenging as its videos contain rich semantics. Therefore, some derived benchmarks are built upon ANet to provide fine-grained annotations [19, 52]. ANet-Captions [19] annotates each video with multiple temporally-grounded captions. ANet-Entities [52] provides spatially-grounded bounding boxes for the noun phrases mentioned in the captions. We establish our ANetQA based on the annotations of these two benchmarks .

3. The ANetQA Benchmark

ANetQA is a large-scale VideoQA benchmark to measure a variety of spatio-temporal reasoning abilities at a fine-grained level. In this section, we first provide an overview of the construction process of our benchmark and then introduce the key stages in detail.

3.1. Overview

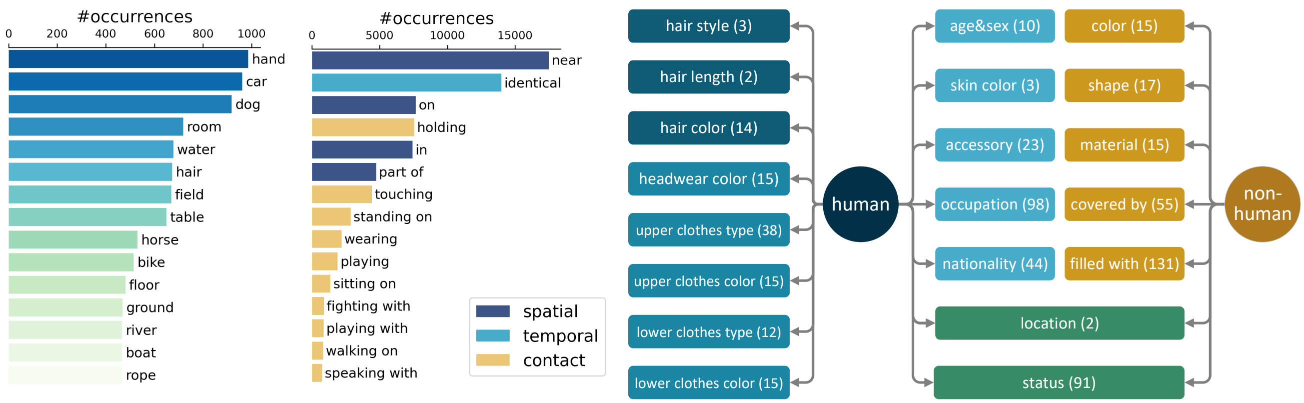
The videos in ANetQA are derived from ActivityNet [4]. As mentioned above, we leverage the auxiliary annotations on ActivityNet [19, 52] to reduce the annotation costs during the construction of our benchmark. These result in 11,525 videos in total, which are comprised of 9,155 and 2,370 videos in the `train` and `val` splits of ActivityNet, respectively. We keep the `train` split unchanged and further divide the `val` split evenly into a new `val` split of 1,185 videos and a `test` split of 1,185 videos.

Next, we annotate each video with a spatio-temporal scene graph via crowdsourcing. Each video has been annotated with temporal-grounded captions [19] and spatially-grounded objects from a few representative frames [52], For each frame, we first clean the mislabeled objects and complement the omitted objects, and then annotate each object with fine-grained relationships and attributes. The accomplished scene graph annotations consist of 118K objects, 83K relationships, 1M attributes, and 16K natural language actions across 43K representative video frames.

Finally, we handcraft a variety of templates to generate linguistically diverse QA pairs with both grammatical and logical guarantees. By composing the elements in the scene graphs and then filling them into proper template slots, we obtain 1.4B unbalanced and 13.4M balanced QA pairs.

3.2. Fine-grained Video Scene Graph Annotation

Representative frames. Annotating a scene graph over all video frames is impractical. Similar to [9], each of our



(a) object distribution (b) relationship distribution (c) attribute hierarchy

Figure 2. **Statistics of the annotated video scene graphs**. We visualize the distributions of the top-15 (a) object occurrences and (b) relationship occurrences. The attributes form a hierarchical taxonomy shown in (c), where the values in the parentheses indicate the number of bottom-level attributes to be annotated. More details are provided in the supplementary material.

scene graph is annotated over a few representative frames in a video. Concretely, we use the selected frames from ANet-Entities [52] as the initialization, which cover the key semantics of all the action segments in ANet-Captions [19]. After that, we manually check and filter out those frames that hamper further annotation, *i.e.*, the frames do not contain any meaningful objects or contain too many objects from the same class. Finally, we obtain 43K frames for further annotation, which indicates an average number of 3.69 frames per video[3].

Objects. ANet-Entities also provides object-level annotations for all the selected frames. Each object is annotated with a bounding box and a noun phrase (*e.g.*, "*a young woman*", "*a black jacket*"). To better organize the object annotations, we first extract nouns from the noun phrases and convert them into a set of object labels. After that, we merge the synonymous object labels (*e.g.*, "*mountain*" and "*hill*", "*saxophone*" and "*sax*"). Finally, we ask annotators to go through all the selected frames to refine the annotations, including object augmentation, label correction, and bounding box calibration. By doing the above, we obtain a total number of 118K objects of 2,072 classes over the selected frames. The top most frequent classes are shown in Figure 2a. We exclude the most frequent class "*person*" for better visualization.

Relationships. Beyond recognizing objects, predicting pairwise relationships between two objects is also important for scene understanding. Referring to the taxonomy in AGQA, we design a set of 86 relationships containing 81 contact relationships (*e.g.*, "*holding*", "*riding*", "*wearing*"), 4 spatial relationships ("*near*", "*on*", "*in*", "*part of*")[4], and 1 temporal relationship ("*identical*"). Our contact relationship categories are broader than AGQA (81 *vs.* 16), because: (i) our videos contain both indoor and outdoor scenarios while AGQA only contains indoor ones; (ii) our relationships contain interactions between two arbitrary objects (*i.e.*, human-object, human-human, and object-object interactions) while AGQA only contains human-object interactions. For each paired objects in one frame, annotators are asked to label *at most* one spatial relationship and one contact relationship, respectively. The "*identical*" temporal relationship indicates the objects in different frames refer to the same instance, which is used to provide indirect references of objects during question generation. Unlike other manually annotated relationships, this relationship is automatically obtained from the annotated attributes, which will be described below. The relationship occurrences follow a long-tail distribution and we illustrate the top most frequent classes in Figure 2b.

Attributes. To distinguish the fine-grained discrepancies between two objects, especially when they share the same object label, we need attribute annotations. Different from the single-label object taxonomy, the attribute taxonomy has a *multi-label* nature in that each object has multiple attributes. Moreover, the attributes for different objects are different. To address the challenges above, we handcraft a *hierarchical* attribute taxonomy by taking the characteristics of our annotated objects into consideration. As shown in Figure 2c, our attribute taxonomy includes three levels. At the top level, we categorize all the object classes into the *human* and *non-human* groups. For each group at the middle level, we design a set of representative attribute types (*e.g.*, "*hair style*" and "*skin color*" for the *human*

[3]The number of sampled frames in our ANetQA is much lower than that of AGQA (3.69 *vs.* 24.4 on average). The motivation derives from our observation that the scene graph elements barely change within an action segment. With a limited annotation budget, we favor the annotation *density* in one frame rather than the annotation *scale* across many frames.

[4]As the viewpoints of our videos are varied, we exclude two spatial relationships ("*in front of*" and "*behind*") in AGQA to avoid ambiguity.

group, "*shape*" and "*material*" for the *non-human* group). A few attribute types like "*location*" and "*status*" are shared across the two groups. At the bottom level, we provide a set of attribute labels for each attribute type (*e.g.*, "*long hair*" and "*short hair*" for the *hair length* attribute type). For each object, annotators are asked to label the bottom-level attributes thoroughly. Due to space limitations, we only show the numbers of attributes at the bottom level in the figure. We have annotated 1M attributes over 118K objects, with an average number of 8.6 attributes per object.

As a by-product, the annotated attributes can facilitate the annotation process of the "*identical*" relationship. Specifically, if two objects in different frames have the same object label, we calculate their overlapping ratio of the annotated attributes. The pairs that surpass a confidence threshold are manually checked to ensure correctness.

To the best of our knowledge, our benchmark is the *first* attempt to provide large-scale and hierarchical attribute annotations for grounded objects in real-world videos.

Actions. In contrast to the objects, attributes, and relationships above, the action segments over specific time intervals of the video often contain much richer semantics. Using a simple label may lose the essential semantics of the action. Therefore, we use a natural language caption to describe each action segment in detail, which has been provided in ANet-Captions [19]. However, some long captions are syntactically complex and are hard to be used for question generation. To this end, we set the maximum length of a caption to 10 and filter out those captions exceeding this threshold. This results in 16K temporally-grounded captions with an average length of 8.1 words.

3.3. Compositional QA Generation

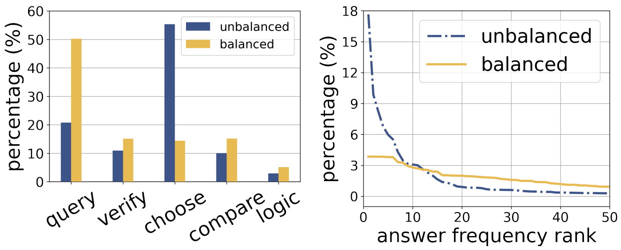
On top of the annotated spatio-temporal scene graphs, we aim to generate massive questions for diverse reasoning abilities. As shown in Table 2, we design a set of 21 question types to cover diverse reasoning skills in varying degrees of complexities. Each question type is categorized into one of the five structures (query, verify, choose, compare, and logic), which refers to the intention of the question. To fulfill the functionality of different question types, we handcraft at least one template for each question type, resulting in 119 grammatical and logical question templates. Similar to AGQA, we design a functional program for each template that traverses and composes the elements in the scene graphs, and fills them into proper template slots to produce compositional QA pairs automatically.

Compared to the question types in AGQA, our major improvements lie in that we introduce 6 extra types with respect to attributes (*i.e.*, the types starting with 'attr' in Table 2). The annotated rich attributes enable us to design up to 101 question templates (*e.g.*, "*what color is ...*", "*what is the shape of ...*"), resulting in 612.6M unbalanced and 6.9M balanced QA pairs. Furthermore, the attribute annotations are also used to describe objects in almost all the rest templates (*e.g.*, "*what is the relationship between the* `[attribute][object]` *and* `[attribute][object]`*?*"). The introduction of attributes not only provides a more precise description of the referred object but also increases the reasoning steps of the generated questions. It is worth noting that although we can describe an object in great detail (*e.g.*, "*a walking young woman wearing green t-shirt and sunglasses*"), this would lead to a risk of combinational explosion and affect the readability of the questions. Therefore, we set the maximum number of attributes used in each question to two.

type	structure	#templ.	#unbal.	#bal.
attrRelWhat†	query	30	169.5M	2.63M
attrWhat†	query	15	70.4M	1.43M
relWhat	query	1	33.1M	1.01M
objRelWhere	query	2	2.5M	0.55M
objRelWhat	query	2	7.1M	0.56M
objWhere	query	1	2.9M	0.43M
objWhat	query	1	0.5M	0.14M
objExist	verify	1	51.7M	1.00M
objRelExist	verify	1	98.3M	0.94M
actExist	verify	1	0.4M	0.08M
objRelWhatChoose	choose	2	347.0M	0.57M
objWhatChoose	choose	1	180.5M	0.55M
attrRelWhatChoose†	choose	36	149.5M	0.42M
attrWhatChoose†	choose	18	85.1M	0.40M
attrCompare†	compare	1	138.0M	2.02M
attrSame†	compare	1	0.09M	0.01M
actTime	compare	1	0.01M	0.01M
actLongerVerify	compare	1	0.01M	0.01M
actShorterVerify	compare	1	0.01M	0.01M
andObjRelExist	logic	1	20.2M	0.35M
xorObjRelExist	logic	1	20.2M	0.35M
overall	-	**119**	**1.4B**	**13.4M**

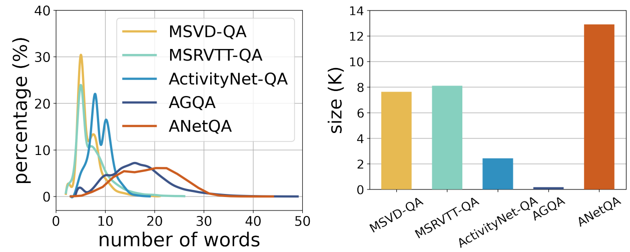
Table 2. **Statistics of the generated questions**. Each question type belongs to a certain structure and contains at least one template. More details are provided in the supplementary material. †: new question types that are not supported in AGQA.

Using the above question templates, we obtain 1.4 billion QA pairs. These QA pairs are *unbalanced* and have strong language biases that models can exploit. We conduct composite balancing strategies on both the questions and answers. Following the question structure distribution in balanced AGQA, our question balancing strategy adjusts the percentages of the query/verify/choose/compare/logic questions to 50%/15%/15%/15%/5%, as shown in Figure 10. While maintaining these percentages above, we conduct answer balancing within each question template to make sure that its answers are uniformly distributed (unbiased). In Figure 3b, we visualize the global answer distributions of the unbalanced and balanced sets in terms of the top-50 most frequent open answers (*i.e.*, the answers to the *query*



(a) question balancing (b) answer balancing

Figure 3. **Distributions before and after balancing.** (a) The question balancing is performed on question structures to adjust the percentages of the query/verify/choose/compare/logic questions to 50%/15%/15%/15%/5%. (b) The answer balancing is conducted on each question template to make its answers follow a uniform distribution. Its effect to the global answer distribution can be observed from the change in the distributions of the top 50 most frequent open answers.



(a) question lengths (b) answer vocabulary sizes

Figure 4. **Question lengths and answer vocabulary sizes.** We compare the (a) question lengths and (b) answer vocabulary sizes of our ANetQA and some typical VideoQA benchmarks like MSVD-QA [40], MSRVTT-QA [40], ActivityNet-QA [45], and AGQA [9]. Compared to the counterparts, our questions are longer and answer vocabulary size is larger, showing the fine granularity, diversity, and difficulty of our benchmark.

structure questions). The obtained results demonstrate the effectiveness of our balancing strategies.

Our final ANetQA benchmark contains 13.4M balanced QA pairs, which consists of 10.4M `train`, 1.5M `val`, and 1.5M `test` QA pairs[5]. We compare the question and answer length distributions of ANetQA to existing VideoQA benchmarks. The results in Figure 4a show that the ANetQA questions have a wider range of lengths and are longer on average than those of all the counterparts, showing the diversity and fine granularity of our questions, respectively. Moreover, according to these challenging questions, our answer vocabulary size is much larger than that of the counterparts (see Figure 4b), which further increases the difficulty of our benchmark.

4. Experiments

This section contains comprehensive experiments and intensive analyses of ANetQA. We conduct evaluations on several state-of-the-art models and diagnose their capabilities to deal with different question structures, semantic classes, reasoning skills, and answer types, respectively. All the models are trained on the `train` split, validated on the `val` split, and evaluated on the `test` split. Furthermore, we also conduct a human evaluation to see the performance gap between the top-performing models and humans. Finally, we investigate the effects of different auxiliary annotations to model performance.

4.1. Experimental Setup

Compared models. We choose three state-of-the-art models for comparison, namely HCRN [20], ClipBERT [21], and All-in-one [37]. HCRN introduces a reusable conditional relation network (CRN) module and stacks multiple CRNs in depth to integrate the motion, question, and appearance features at different levels [20]. We use its default settings to extract 128 appearance features and 8 motion features, respectively.

Different from HCRN, ClipBERT and All-in-one are two Transformer-based models that incorporate vision-language pretraining (VLP) on a large-scale corpus. ClipBERT is pretrained on massive image-text pairs, which enables end-to-end learning by employing a sparse sampling mechanism. We adopt its official pretrained model weights as initial and then finetune the model on ANetQA using the (4×2) sampling strategy, which means 4 segments are sampled (with 2 sampled frames in each segment) at each training step. During model testing, we sample 16 frames uniformly for each video, as recommended in [21]. All-in-one is a current top-performing VideoQA model, which is the first attempt to perform end-to-end video-language pretraining using raw video and textual signals as inputs [37]. It is pretrained directly on a large-scale video-text corpus. We finetune its base model All-in-one-B on ANetQA by randomly sampling 3 frames for each video at each training step. At inference time, we also extract 3 frames uniformly and feed them to the learned model to predict the answer.

Human evaluation. We conduct an intensive human evaluation to quantify the errors and ambiguities induced during the construction of ANetQA. As the labeling costs is unaffordable to provide a thorough evaluation over all the QA pairs, we follow [9, 11] to randomly sample 4,000 QA pairs from the `test` set with the following two rules: (i) each video contains at least one sample, and (ii) each question type contains at least 50 samples. Each sample is assigned to five random annotators from a diverse group

[5]For more efficient evaluation, we additionally provide a `test-dev` split by random sampling 0.3M QA pairs from the `test` split. Note that both the `test` and `test-dev` splits are conducted on the same video set and the evaluation for both splits are performed online.

taxonomy		type prior	HCRN [20] w/	HCRN [20] w/o	ClipBERT [21] w/	ClipBERT [21] w/o	All-in-one [37] w/	All-in-one [37] w/o	human
question structures	query	1.04	21.30	19.24	23.93	16.87	**25.10**	18.40	92.92
	compare	49.70	**55.66**	50.01	55.62	50.06	54.41	50.06	81.34
	choose	29.13	63.97	67.37	69.51	66.17	**70.39**	67.00	71.84
	verify	50.00	68.56	50.02	**72.57**	50.00	72.35	50.00	86.69
	logic	50.00	78.70	76.82	80.06	74.33	**80.58**	74.20	86.06
question semantics	object	17.74	55.99	49.55	58.69	48.22	**59.81**	48.99	84.26
	relationship	22.61	39.65	33.28	40.19	30.89	**40.78**	32.64	90.79
	attribute	14.60	35.80	34.05	39.71	32.81	**40.14**	33.39	82.17
	action	47.83	72.50	50.29	**74.96**	50.99	74.39	51.14	82.33
reasoning skills	object-relationship	10.48	35.17	32.38	37.66	30.03	**38.42**	31.32	86.47
	object-attribute	17.44	40.95	37.02	43.72	35.45	**44.33**	36.39	84.75
	duration-comparison	50.00	49.90	49.38	49.98	50.10	**51.65**	54.34	76.73
	exist	50.00	71.20	56.97	**74.51**	56.31	74.49	56.28	86.52
	sequencing	10.21	31.70	31.36	34.19	28.76	**35.27**	30.10	87.50
	superlative	30.32	47.46	39.78	49.55	38.83	**50.14**	39.60	90.14
answer types	binary	49.96	64.36	53.91	**66.19**	53.55	65.65	53.54	83.72
	open	6.49	29.95	29.00	33.17	26.86	**34.33**	28.25	84.82
overall		17.66	41.15	37.11	43.92	35.55	**44.53**	36.48	84.48

Table 3. **A comprehensive comparison of three VideoQA methods on ANetQA**. All results are evaluated on the `test` set. Apart from the overall accuracy, we follow [9] to report the per-type accuracies under different taxonomies. For each method, the variant trained with vision clues (w/) outperforms its blind counterpart without vision clues (w/o), implying that the language biases are well controlled.

to answer the question and the majority vote over their predictions is regarded as the final human answer.

The human performance reach at 84.48% on the sampled `test` set. We take a closer look into these 15.52% inconsistent human predictions and find that they are constituted by 0.75% annotation errors, 1.95% answer ambiguities, and 12.82% human errors. These results imply that both of our scene graphs and generated QA pairs are of high quality. Furthermore, our benchmark contains difficult questions that even educated humans can not answer correctly. More analyses are provided in the supplementary material.

4.2. Main Results

We provide an intensive comparison of the state-of-the-art methods on ANetQA In Table 3. Besides the overall accuracy, we follow [9] to report the per-type accuracies under different taxonomies, *i.e.*, question structures, question semantics, reasoning skills, and answer types. More detailed descriptions of the taxonomies and corresponding question templates are provided in the supplementary material. For each type, we provide a simple baseline, *type prior*, that uses the most frequent answer as the prediction.

From the results, we have the general observations as follows: (i) The All-in-one model pretrained on large video-text corpus achieves the overall best performance while using the least number of sampled frames. This suggests good video representations play a central role in VideoQA performance; (ii) the best performing model is still far from the human level, showing the difficulty of our benchmark and sufficient room for further improvements; and (iii) for each method, the variant trained with vision clues (w/) steadily outperforms its *blind* counterpart without any vision clues (w/o), indicating that the language biases are well controlled by our balancing strategies.

The observations above are quite different from those on AGQA, where on their benchmark all models are on par with their corresponding blind counterparts. This can be explained that ANetQA has more unbalanced QA samples than AGQA, thus providing more room to perform thorough balancing strategies. Moreover, given the same model HCRN, its accuracy (especially the *open* answer type) on ANetQA is much lower than that on AGQA, verifying the fine-grained nature of our scene graphs elements.

Question structures and answer types. The *query* type questions are the most challenging ones as they have open answers. Among the rest four types which have limited answer choices[6], the *compare* type questions report the lowest accuracy as they require more reasoning steps.

Question semantics. The attribute-oriented questions are the most difficult ones, as they require a more fine-grained understanding of video contents than the rest questions.

Reasoning skills. Similar to AGQA, each of our question is

[6]The *compare*, *verify*, and *logic* type questions have binary answers. The *choose* type question conducts a comparison between [A] and [B], and the answer refers to one of the four choices: [A], [B], both, or none.

associated with one or more reasoning abilities necessary to answer the question. The questions requiring the *sequencing* skills deliver the lowest accuracy as they require the temporal grounding ability. In contrast to the coarse action labels used in AGQA, our actions are depicted in natural language, which are more difficult to understand.

4.3. Effects of Auxiliary Annotations

All the comparative studies above only use the basic annotations (*i.e.*, the QA pairs) for model training. As all the QA pairs are automatically generated from scene graph annotations, it is natural to investigate whether and how auxiliary annotations facilitate model performance. We introduce two auxiliary annotations *scene graph statistics* and *oracle frames* to see their impacts on model performance, respectively. The results are provided in Table 4.

Scene graph statistics. The annotated scene graph of a given video contains all the necessary information to answer any questions on the video. Therefore, it is meaningful to investigate the impact of this information on model performance. The fine-grained characteristics of our scene graphs make it nontrivial to encode each scene graph into a feature bank like [13]. Alternatively, we introduce a simple statistical-based strategy to approximately represent the scene graph to a given video by extracting the top-K high-frequency (HF) words from *all* the questions on this video. The extracted HF words can be seamlessly used in any off-the-shelf model by concatenating them with the question words. We adopt HCRN [20] as the reference model and extract the top-40 HF words from different vocabularies (*i.e.*, objects, relationships, attributes, and their combinations). These HF words are concatenated with the question words in both the training and testing phases.

From the results in the upper part of Table 4, we can see that adding HF objects or relationships solely do not bring further improvement over the reference model. This can be explained by the fact that relationships are strongly coupled with objects, using either of them solely can not provide sufficient scene graph information for the model to understand. Moreover, the model with HF attributes results in a distinct performance gain compared to the counterpart with HF objects. This observation verifies that our questions requires the abilities of fine-grained understanding and reasoning. Finally, exploiting all three types of HF information results in the best performance due to their complementary nature.

Oracle frames. As each question in ANetQA is generated from the scene graph elements in specific video frames, we denote these frames as the oracle frames for the question and investigate whether they can facilitate model performance. For each question, we inject the corresponding oracle frames into its sampled frames to ensure the necessary visual information to answer this question is provided. We use All-in-one [37] as the reference model since it uses few sampled frames and thus has a high probability of not covering the oracle frames. We have experimented with the oracle frames in the training, testing, and both phases, respectively. The results in the lower part of Table 4 show that injecting oracle frames in the training and testing phases bring 0.87 and 0.46 point improvements over the reference model in terms of overall accuracy, respectively. Moreover, when oracle frames are applied to both the training and testing phases, the model performance is further improved due to their synergistic effects.

	binary	open	overall
(a) *scene graph statistics*			
HRCN [20] (reference)	64.36	29.95	41.15
+ high-freq. objects (O)	65.81	29.29	41.18
+ high-freq. relationships (R)	63.84	29.21	40.48
+ high-freq. attributes (A)	67.67	32.21	43.75
+ high-freq. O+R+A	**68.15**	**34.50**	**45.45**
(b) *oracle frames*			
All-in-one [37] (reference)	65.65	34.33	44.53
+ training phase injection	66.54	35.18	45.40
+ testing phase injection	66.04	34.83	44.99
+ both phases injections	**66.88**	**36.02**	**46.07**

Table 4. **Effects of different auxiliary annotations.** (a) The scene graph statistics of a given video are represented as a set of high-frequency words extracted from all the questions of that video. (b) The oracle frames contain necessary visual information to answer a given question, which are injected in different phases.

5. Conclusion and Future Work

In this paper, we present ANetQA, a challenging VideoQA benchmark that examines fine-grained compositional reasoning over untrimmed real-world videos. Benefiting from the fine-grained video scene graphs annotated by humans, ANetQA attains 13.4M balanced QA pairs, which is an order of magnitude larger than all previous VideoQA benchmarks. We provide comprehensive experiments and intensive analyses for state-of-the-art VideoQA methods, and the best-performing model showing that a fine-grained video understanding plays a vital role in our benchmark. Moreover, there remains a significant gap between the best model and humans, indicating the challenge of our benchmark while providing room for future improvements.

We will persistently improve our benchmark. *e.g.*, further reducing the language biases and answer ambiguities, and introducing more question types with diverse reasoning skills like scene-text understanding and causality inference. We hope that our ANetQA will serve as a cornerstone to facilitate future research in the video-language learning.

Acknowledgment.

This work was supported in part by the NSFC (61836002, 62125201), in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang (GK229909299001-001), in part by the NSFC (62072147, 62020106007), and in part by the Zhejiang Provincial Natural Science Foundation of China (LR22F020001, DT23F020007).

A. Scene Graph Annotations

A.1. Annotation Pipeline

As mentioned in the main paper, ANetQA is built upon the annotations of ANet-Entities [52], which grounds objects in representative frames with noun phrases (NPs). Nouns and adjectives are extracted from these NPs using the Stanford Parser [30] to form our initial object and attribute vocabularies, respectively. Meanwhile, we handcraft the initial relationship vocabulary on the activity labels of the original ActivityNet [4]. These initial vocabularies are intermittently updated during the annotation process.

We provide a web-based interface shown in Figure 5 for crowdsourcing. In total, more than 50 human annotators have participated in the annotation process for over 4 months. Each annotator is asked to watch the video first and then select attributes, and relationships from the corresponding vocabularies. When no suitable option is available, they are allowed to add a new option. These new options will be manually checked and the valid ones will be added to the vocabularies intermittently. Meanwhile, the mislabeled objects and inaccurate object bounding boxes are fixed and omitted key objects are complemented during the annotation process. To control the annotation costs, we set the maximum number of augmented objects to three.

A.2. Scene Graph Taxonomies

Our completed scene graph annotations include taxonomies of 2,072 object classes, 86 relationship classes, and 618 attributes classes. The detail taxonomies for objects, relationships, and attributes are shown in Table 5, Table 6, and Figure 6, respectively. As our actions are depicted in natural language, we illustrate a word cloud for the most frequent verbs in Figure 7.

A.3. Case Study

In Figure 8, we provide comparative examples of the annotated scene graphs from ANetQA and AGQA, respectively. From the visualized results we can see that: (i) our scene graph is more informative than that in AGQA as our untrimmed video contains richer semantics with multiple switched scenarios; (ii) our scene graph is much more fine-grained than that in AGQA due to the objects, relationships, actions, especially the newly introduced attributes; (iii) our scene graph contains varied relationships between human-object, human-human, and object-object pairs, while the scene graph of AGQA only contains human-object relationships; and (iv) our scene graph uses the "*identical*" relationship to annotate the same instance across different frames, which effectively avoids the generation of ambitious questions. In contrast, the scene graph of AGQA is centered on *one* person, which cannot always be satisfied in real-world videos. As shown at the bottom, the annotated "*person*" refers to the man in the first four frames and shifts to the boy in the last frame.

B. Compositional QA Generation

B.1. Taxonomies, Templates, and Programs

We show the question taxonomies and templates for our benchmark in Table 7. Similar to AGQA, each question type is categorized into different in terms of different perspectives (*i.e.*, structure, semantics, reasoning skill, and answer type). Each question type corresponds to at least one question template with a maximum number of reasoning steps. Compared with AGQA, ANetQA has more diverse question templates (119 *vs.* 28) , showing the diversity, fine granularity, and difficulty of our benchmark. The functional program for each template is shown in Table 8.

B.2. Question Distributions

ANetQA contains 13.4M balanced QA pairs in total. We display the distributions of these QA pairs in terms of different taxonomies in Figure 9. The results show that: (i) the question structure distribution meets the expectation of our balancing strategy; (ii) the attribute-related questions account for a large percentage in terms of question semantics and reasoning skills, respectively; and (iii) the proportion of the *open* type answers is roughly twice that of the *binary* type answers. In Figure 10, we illustrate the question distribution by the first three words. The results show that our questions are both semantically and linguistically diverse.

B.3. Example QA pairs

We provide some example QA pairs from the `train` and `val` splits in Figure 11. Each example contains five QA pairs on the same video with different question structures (*i.e.*, query, verify, choose, compare, and logic). The examples verify that our questions are diverse, fine-grained, and challenging at the same time.

C. Experiments

C.1. Human Evaluation

As reported in the main paper, human performance tops out at 84.48% overall accuracy by taking the majority voting over five answers per question. In Figure 12, we provide more detailed analyses of the human evaluation statistics to better understand the behavior of individual annotators. The results in Figure 12a indicate that the deviations among different annotators do exist, and majority voting helps eliminate individual errors. The results in Figure 12b show that different question types lead to diverse accuracies and deviations. The average accuracy per individual annotator is 81.5%.

C.2. Per-Split & Per-Type Accuracies

In Table 9, we provide comparisons of the same model on the `val` and `test` split, respectively. The results show that: (i) the results on the `test` split is slightly lower that the `val` split; and (ii) there is no much difference between the performance on the `test` and `test-dev` splits.

In Table 10, we report the per-type accuracies of the three models. From the results we can see that the best-performing model All-in-one [37] consistently outperforms the other two models in majority of the question types.

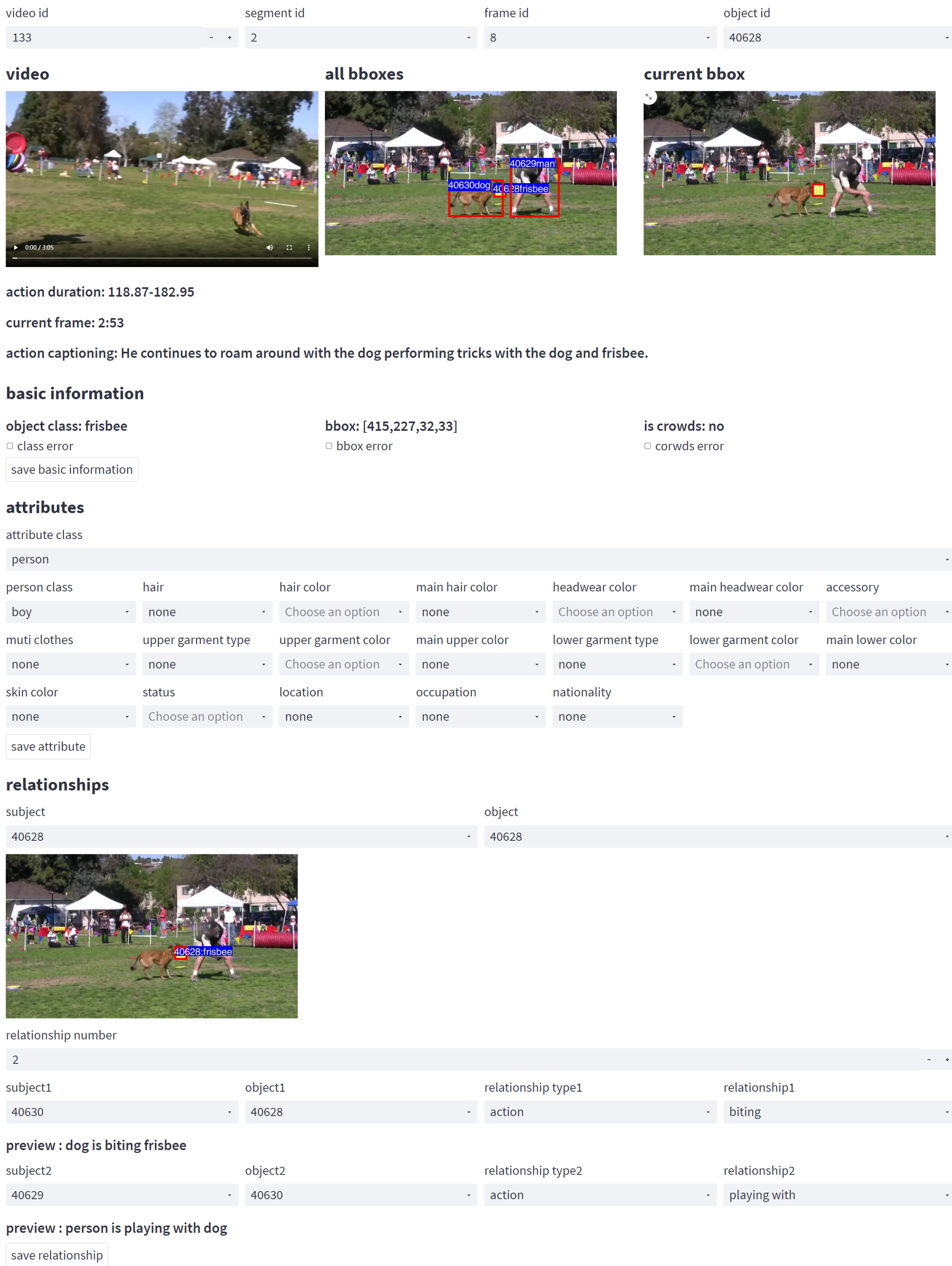


Figure 5. A web-based interface for video scene graph annotation by crowdsourcing. Annotators are asked to watch the video first and then select attributes and relationships from corresponding vocabularies. When no suitable item is available, they can add new items freely. These new items will be manually checked and the valid ones will be appended to the vocabularies intermittently.

hand	car	dog	room	water	hair	field	table
horse	bike	floor	ground	river	boat	rope	board
bar	wall	shoe	hill	arm	bowl	shirt	face
tree	gym	pool	stage	drum	barbell	cup	skateboard
track	clothes	mat	leg	snow	paper	sink	stick
street	brush	tire	tool	court	beach	ingredient	head
chair	glass	grass	knife	machine	roof	foot	cat
wood	plate	pole	bottle	road	house	ocean	food
beam	mower	bull	hoop	frisbee	yard	guitar	box
window	wave	kitchen	towel	sea	pot	football	ski
slope	tube	bucket	nail	bowling ball	fence	leaf	dart
pumpkin	eye	canoe	pasta	building	tile	drink	rock
lawn	camel	surfboard	lake	slide	rubik's cube	ice	pinata
pan	contact len	kayak	counter	hat	violin	bow	pit
raft	arena	fish	swing	cake	potato	cigarette	volleyball
park	arrow	saxophone	baton	motorbike	croquet	racket	cookie
dodgeball	carpet	bread	sandwich	short sleeves	vacuum	hockey	hammer
bag	shovel	area	elliptical machine	javelin	curling	kite	shot
mirror	tennis	piano	lemon	mouth	door	sidewalk	accordion
line	icecream	shop	shuffleboard	table tennis	lane	stair	body
microphone	finger	paint	net	harmonica	helmet	liquid	water polo
discus	product	egg	bathroom	platform	fire	gun	studio
suit	alcohol	back	paddle	sand	glove	mop	hole
sofa	stilt	stand	pin	beer	flute	dish	rag
smoke	scissors	tattoo	sky	tomato	razor	vest	basketball

Table 5. A list of top-200 object classes in terms of occurrences in our benchmark. Sorted by row first.

spatial	near	in	on	part of		
temporal	identical					
contact	pulling	holding	touching	fighting with	wearing	hitting
	playing	standing on	playing with	sweeping	wiping	sitting on
	spitting	stirring	eating	jumping into	taking picture of	driving
	riding	leading	throwing	climbing	leaning on	covering
	lying on	kneeling on	walking on	raising	biting	hugging
	cutting	running on	jumping on	squating on	trimming	scraping
	carrying	pushing	brushing	pointing at	dancing with	chasing
	surfing on	polishing	washing	drinking from	stamping	fishing
	speaking with	pouring	drinking	crossing	dragging	repairing
	smoking	sliding on	bowing to	drawing on	hanging on	drawn on
	making	flying from	drawing	feeding	poured into	flowing from
	kissing	twisting	writing on	burning	lighting	pouring into
	spraying	commanding	blowing	heating	pointing	painting on
	painting	painted on	wirting on			

Table 6. A list of all the 86 relationships in our benchmark, including 4 spatial, 1 temporal, and 81 contact relationships. Sorted by row first in terms of occurrences.

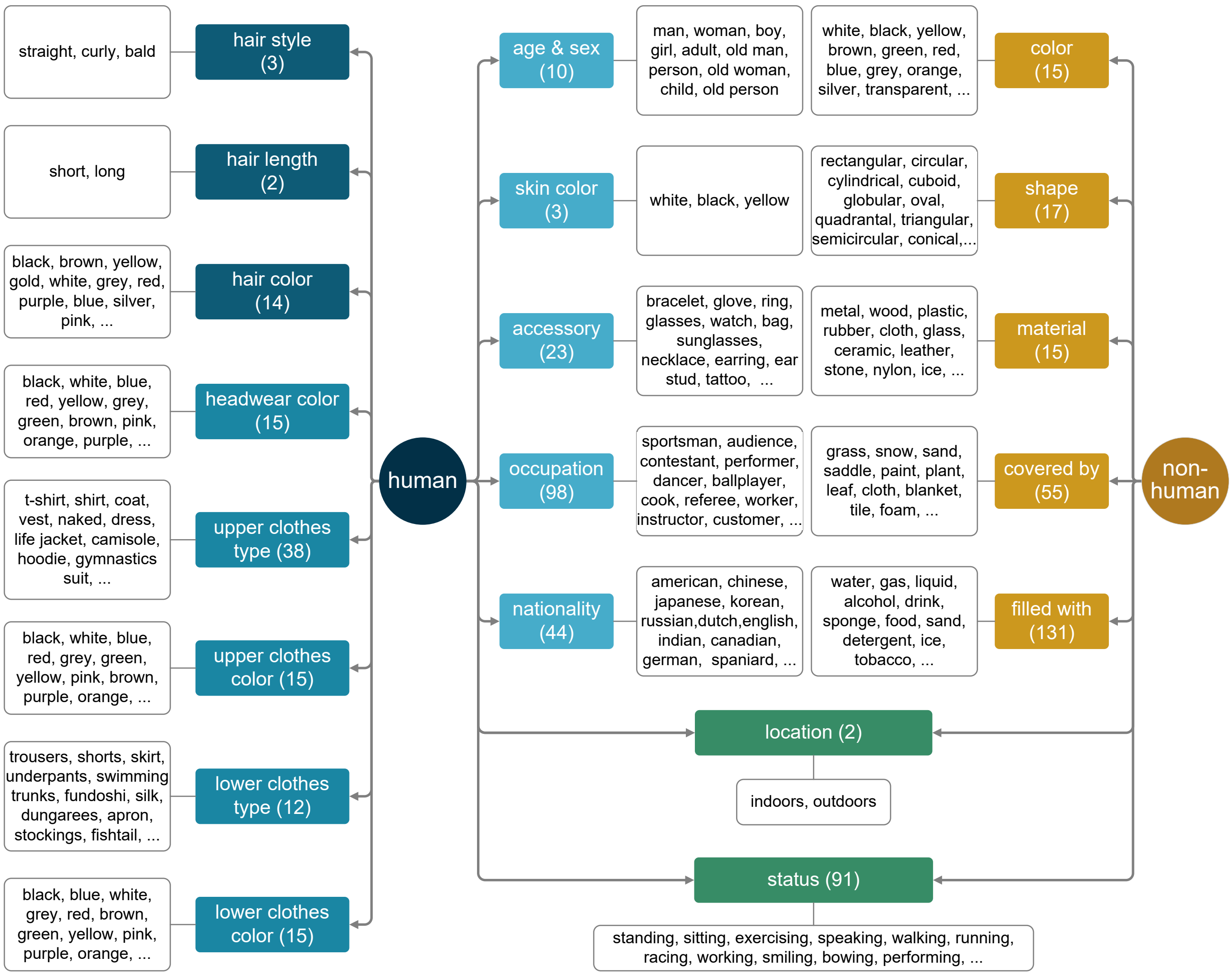


Figure 6. A hierarchy of attributes in our benchmark. The hierarchy consists of three levels. On the **top** level, objects are classified into the *human* and *non-human* groups. On the **middle** level, up to 20 representative attribute types are designed for each top groups (*e.g.*, "*hair style*" and "*skin color*" for the "*human*" group, "shape" and "material" for the "*non-human*" group). A few attributes like "*location*" and "*status*" are shared across the two groups. On the **bottom** level, a total number of 618 attribute labels are provided for all the middle-level attribute types (*e.g.*, "*long hair*" and "*short hair*" for the "*hair length*" attribute type). For each object, annotators are asked to label the bottom-level attributes as thoroughly as possible. Due to space limitations, we show a maximum number of 10 bottom-level attributes for each mid-level attribute type.

Figure 7. A word cloud for frequent *verbs* in action descriptions. We merge the words with the same etymon for better visualization.

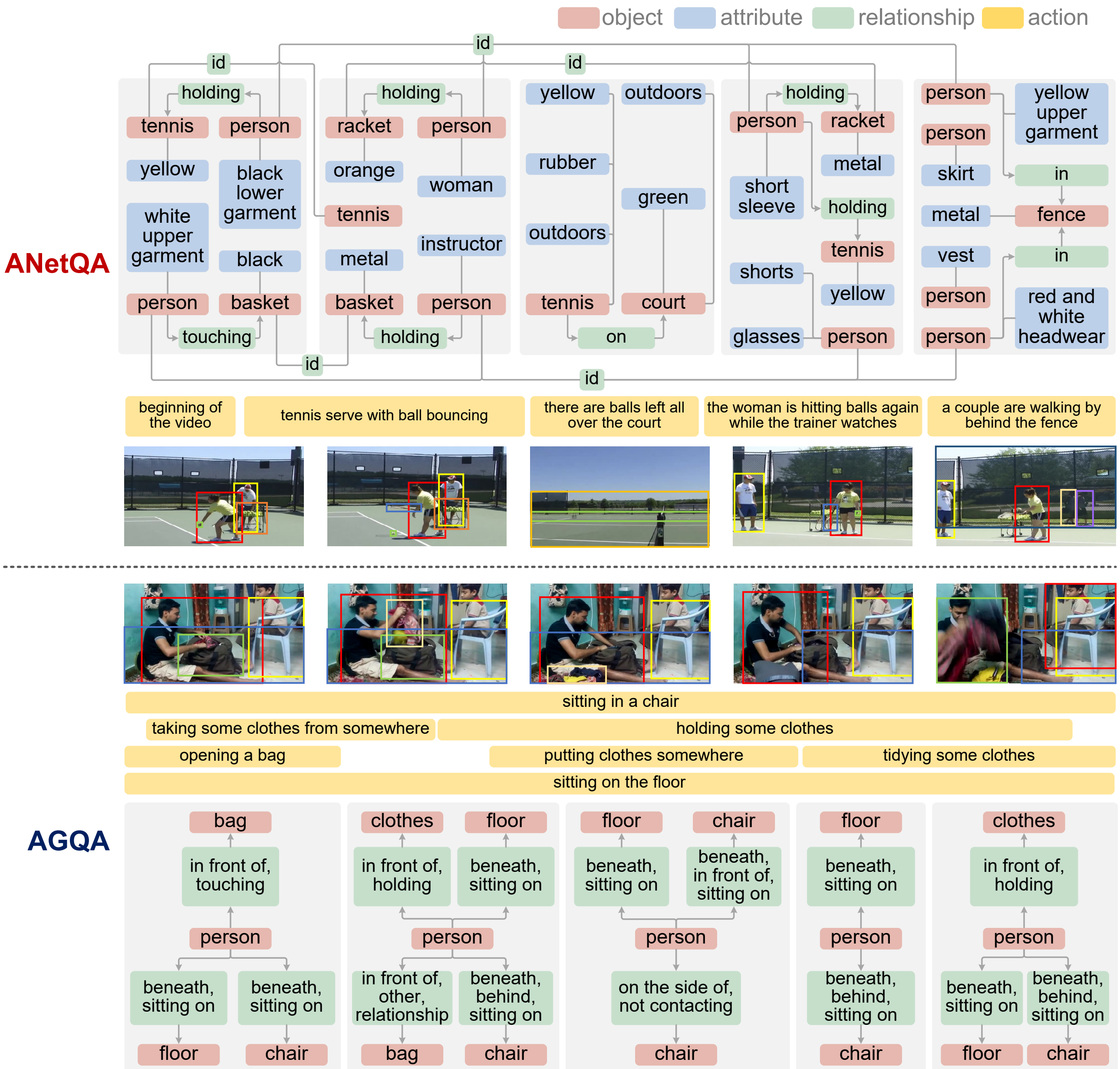


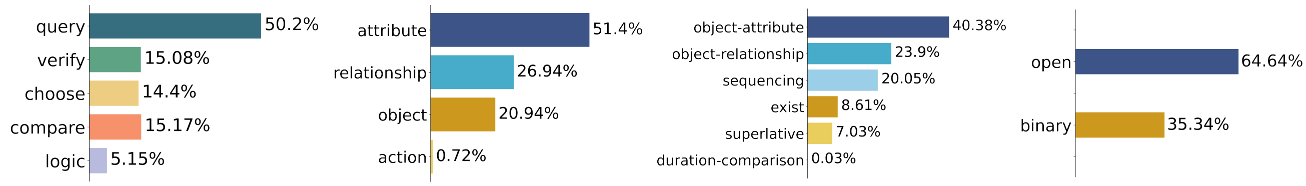
Figure 8. A comparison of the example scene graphs of our ANetQA and AGQA. The visualized results suggest: (i) our scene graph is more informative than that in AGQA as our untrimmed video contains richer semantics with multiple switched scenarios; (ii) our scene graph is much more fine-grained than that in AGQA due to the objects, relationships, actions, especially the newly introduced attributes; (iii) our scene graph contains varied relationships between human-object, human-human, and object-object pairs, while the scene graph of AGQA only contains human-object relationships; and (iv) our scene graph uses the "*identical*" relationship to annotate the same instance across different frames, which effectively avoids the generation of ambitious questions. In contrast, the scene graph of AGQA is centered on *one* person, which cannot always be satisfied in real-world videos. Specifically, the annotated "*person*" refers to the man in the first four frames and shifts to the boy in the last frame.

type	question structures	question semantics	reasoning skill	answer types	reasoning steps	#templ.	question template
attrRelWhat	query	attribute	obj-attr,obj-rel	open	5	15 15	what [attr-type] is the [attr1] [obj1] [rel] [attr2] [obj2]? what [attr-type] is the [attr1] [obj1] that [attr2] [obj2] is [rel]?
attrWhat	query	attribute	obj-attr	open	3	15	what [attr-type] is the [attr] [obj]?
relWhat	query	relationship	obj-attr,obj-rel	open	5	1	what is the relationship between [attr1] [obj1] and [attr2] [obj2]?
objRelWhere	query	relationship	obj-attr,obj-rel	open	5	1 1	where is the [attr1] [obj1] [rel] [attr2] [obj2]? where is the [attr1] [obj1] that [attr2] [obj2] is [rel]?
objRelWhat	query	object	obj-attr, obj-rel	open	5	1 1	what is the [attr1] object [rel] [attr2] [obj2]? what is the [attr1] object that [attr2] [obj2] is [rel]?
objWhere	query	relationship	obj-attr,obj-rel	open	3	1	where is the [attr] [obj]?
objWhat	query	object	obj-attr	open	3	1	what is [attr] object?
objExist	verify	object	exists,obj-attr	binary	3	1	does [attr] [obj] appear?
objRelExist	verify	relationship	exists,obj-attr,obj-rel	binary	5	1	is [attr1] [obj1] [rel] [attr2] [obj2]?
actExist	verify	action	exist	binary	2	1	is someone [act]?
objRelWhatChoose	choose	object	obj-attr,obj-rel	open	5	1 1	which is [attr1] object [rel] [attr2] [obj2], [obj-A] or [obj-B]? which is [attr1] object that [attr2] [obj2] is [rel], [obj-A] or [obj-B]?
objWhatChoose	choose	object	obj-attr	open	3	1	which is [attr] object, [obj-A] or [obj-B]?
attrRelWhatChoose	choose	attribute	obj-attr,obj-rel	open	5	18 18	which [attr-type] is the [attr1] [obj1] [rel] [attr2] [obj2], [attr-A] or [attr-B]? which [attr-type] is the [attr1] [obj1] that [attr2] [obj2] is [rel], [attr-A] or [attr-B]?
attrWhatChoose	choose	attribute	obj-attr	open	3	18	which [attr-type] is the [attr] [obj], [attr-A] or [attr-B]?
attrCompare	compare	attribute	obj-attr	binary	5	1	is the [attr-type] of the [attr] [obj] the same as that of the [attr] [obj]?
attrSame	compare	attribute	obj-attr	open	5	1	what is the same attributes of [attr1] [obj1] and [attr2] [obj2]?
actTime	compare	action	suquencing	binary	5	1	is someone [act] before or after [act]?
actLongerVerify	compare	action	duration-comparison	binary	5	1	is the duration of someone [act1] for longer than the duration of [act2]?
actShorterVerify	compare	action	duration-comparison	binary	5	1	is the duration of someone [act1] for shorter than the duration of [act2]?
andObjRelExist	logic	relationship	exists,obj-attr,obj-rel	binary	8	1	is [attr1] [obj1] [rel] [attr2] [obj2] and [attr3] [obj3]?
xorObjRelExist	logic	relationship	exists,obj-attr,obj-rel	binary	8	1	is [attr1] [obj1] [rel] [attr2] [obj2] but not [attr3] [obj3]?

Table 7. Question taxonomy and templates. ANetQA contains 21 types of questions generated from 119 templates. Each question type is categorized into different taxonomies (*i.e.*, structure, semantics, reasoning skill, and answer type), and refers to a maximum number of reasoning steps. Note that the reasoning skills of *sequencing* and superlative are optionally used in all the question types by inserting a clause starting with "before/after [act]" or "in the beginning/end of the video". [attr-type] refers to a set of templates that ask different middle-level attribute types shown in Figure 6. Note that some attribute types may slightly deviate from the corresponding template (*e.g.*, "*what is the occupation of ...*" or "*what are the accessories of ...*"). Due to space limitations, we do not expand all the templates and only show the most commonly-used one for those question types with multiple templates.

template	functional program
what [attr-type] is the [attr1] [obj1] [rel] [attr2] [obj2]?	select:[obj2]→filter:[attr2]→relate:[obj1],[rel] →filter:[attr1]→query:⟨[attr-type]⟩
what [attr-type] is the [attr2] [obj2] that [attr1] [obj1] is [rel]?	
what [attr-type] is the [attr] [obj]?	select:[obj]→filter:[attr]→query:⟨[attr-type]⟩
what is the relationship between [attr1] [obj1] and [attr2] [obj2]?	select:[obj1]→filter:[attr1]→select: [obj2] →filter:[attr2]→query:⟨relationship⟩
where is the [attr1] [obj1] [rel] [attr2] [obj2]?	select:[obj2]→filter:[attr2]→relate:[obj1],[rel] →filter:[attr1]→query:⟨spatial-relationship⟩
where is the [attr1] [obj1] that [attr2] [obj2] is [rel]?	
what is the [attr1] object [rel] [attr2] [obj2]?	select:[obj2]→filter:[attr2]→relate:_,[rel] →filter:[attr1]→query:⟨object⟩
what is the [attr1] object that [attr2] [obj2] is [rel]?	
where is the [attr] [obj]?	select:[obj]→filter:[attr]→query:⟨spatial-relationship⟩
what is [attr] object?	select:_→filter:[attr]→query:⟨object⟩
does [attr] [obj] appear?	select:[obj]→filter:[attr]→exist
is [attr1] [obj1] [rel] [attr2] [obj2]?	select:[obj1]→filter:[attr1]→relate:[obj2],[rel] →filter:[attr2]→exist
is someone [act]?	select:[act]→exist
which is [attr1] object [rel] [attr2] [obj2], [obj-A] or [obj-B]?	select:[obj2]→filter:[attr2]→relate:_,[rel] →filter:[attr1]→choose:[obj-A] | [obj-B]
which is [attr1] object that [attr2] [obj2] is [rel], [obj-A] or [obj-B]?	
which is [attr] object, [obj-A] or [obj-B]?	select:_→filter:[attr]→choose:[obj-A] | [obj-B]
which [attr-type] is the [attr1] [obj1] [rel] [attr2] [obj2], [attr-A] or [attr-B]?	select:[obj2]→ filter:[attr2]→relate:[obj1],[rel] →filter[attr1]→choose:[attr-A] | [attr-B]
which [attr-type] is the [attr1] [obj1] that [attr2] [obj2] is [rel], [attr-A] or [attr-B]?	
which [attr-type] is the [attr] [obj], [attr-A] or [attr-B]?	select:[obj]→filter:[attr]→choose:[attr-A] | [attr-B]
is the [attr-type] of the [attr1] [obj1] the same as that of the [attr2] [obj2]?	select:[obj1]→filter:[attr1]→select:[obj2] →filter[attr2]→compare:⟨[attr-type]⟩
what is the same attributes of [attr1] [obj1] and [attr2] [obj2]?	select:[obj1]→filter:[attr1]→select:[obj2] →filter[attr2]→compare:⟨attribute⟩
is someone [act1] before or after [act2]?	select:[act1]→localize:[act1]→select:[act2] →localize:[act2]→compare:⟨time⟩
is the duration of someone [act1] for longer than the duration of [act2]?	
is the duration of someone [act1] for shorter than the duration of [act2]?	
is [attr1] [obj1] [rel] [attr2] [obj2] and [attr3] [obj3]?	select:[obj1]→filter:[attr1]→relate:[obj2],[rel] →filter:[attr2]→and→relate:[obj3],[rel] →filter:[attr3]→exist
is [attr1] [obj1] [rel] [attr2] [obj2] but not [attr3] [obj3]?	select:[obj1]→filter:[attr1]→relate:[obj2],[rel] →filter:[attr2]→xor→relate:[obj3],[rel] →filter:[attr3]→exist

Table 8. Functional programs and their corresponding question templates. Each program consists of a sequence of predefined primary functions. The relate function can support the association of either subject or object. The symbol '_' means traversing all objects to meet the following constraint.



(a) question structures (b) question semantics (c) reasoning skills (d) answer types

Figure 9. **Question distributions in terms of different taxonomies** on the balanced version. (a) The question structure distribution meets the expectation of our balancing strategy; (b) and (c) The attribute-related questions account for a large percentage in terms of question semantics and reasoning skills, respectively. (d) The proportion of the *open* type answers is roughly twice that of the *binary* type answers.

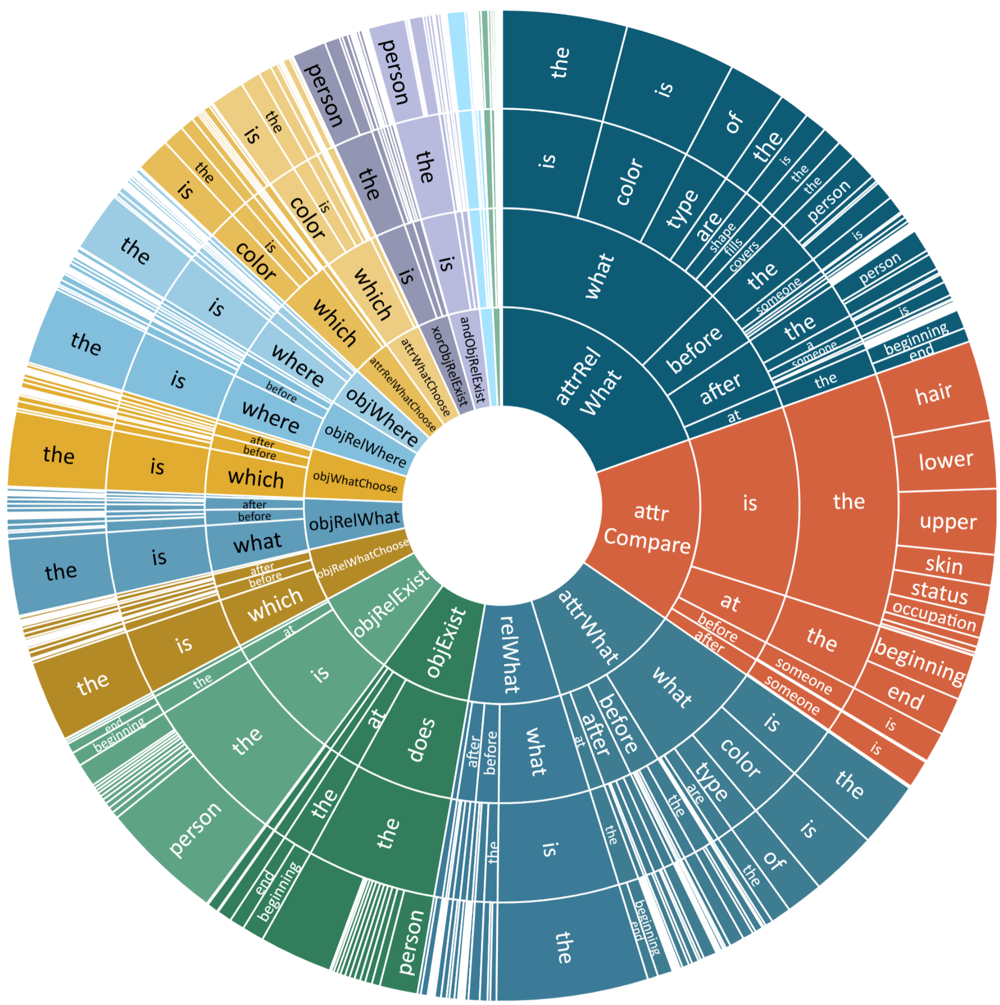
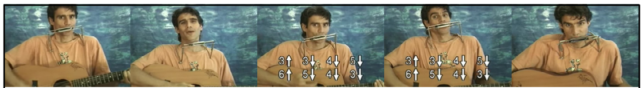
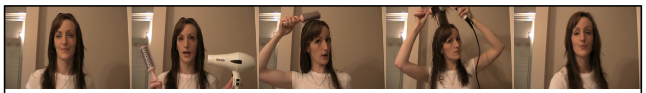


Figure 10. **Question distribution by their first three words** on the balanced benchmark. The innermost ring refers to the 21 question types. The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. For the questions with the same structure (query, compare, verify, choose, and logic), we use the background color from the same color scheme (blue, orange, green, yellow, and purple).

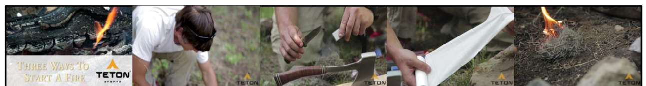


Q1: At the end of the video, what shape is the silver harmonica that the person wearing the t-shirt is playing?	A: cuboid
Q2: Is someone playing the harmonica at the end of the video?	A: yes
Q3: Which color is the upper garment of the person who is performing, yellow or silver?	A: yellow
Q4: Is duration of someone playing the harmonica for longer than the duration of a man plays guitar and harmonica at the same time?	A: yes
Q5: Is the person wearing the yellow upper garment playing the yellow object and the white flute at the beginning of the video?	A: no

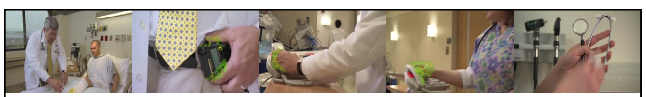
Q1: Where is the target before someone is doing archery?	A: on the field
Q2: Is the long-haired person holding the black arrow?	A: yes
Q3: Before someone is doing archery, Which is the metal object that the person with curly hair is holding, the arrow or the scythe?	A: arrow
Q4: Before someone is doing archery, what is the same attribute of bow and black arrow?	A: material
Q5: Is the person in the vest holding the bow and the metal arrow?	A: yes



Q1: After the lady brushes her hair ,what is the relationship between the hairdryer and the person with long hair?	A: the person is holding the hairdryer
Q2: Does the straight-haired person with the watch appear in the video?	A: no
Q3: Which color is the upper garment of the person who is standing, black or grey?	A: both false
Q4: Is someone blow-drying hair before or after a lady stands in a bathroom talking?	A: after
Q5: After the lady brushes her hair, is the person with straight hair holding the silver comb but not the black brace?	A: yes



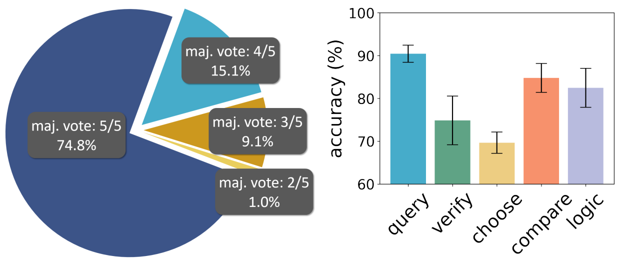
Q1: What color is the upper garment of the brown-haired person in the t-shirt after someone is starting a campfire?	A:white
Q2: Does the curly-haired person wearing the red upper garment appear in the video?	A: no
Q3: Which color is the fire, brown or gold?	A: gold
Q4: Is the duration of someone starting a campfire for shorter than the duration of a camper describes how to make a fire?	A: yes
Q5: Is the person with brown hair holding the knife and the silver object?	A: yes



Q1: What is the lighting green object before someone is washing hands?	A: sparkle
Q2: Is the person with the bracelet holding the phone indoors?	A: no
Q3: Which is the occupation of the person with the glasses and the necktie touching the leg, the doctor or the nail artist?	A: doctor
Q4: Is the hair color of the person who is sitting the same as that of the doctor?	A: no
Q5: Is the person holding the rectangular object and the stethoscope?	A: yes

Q1: What is the orange object filled with powder?	A:baking soda
Q2: Is the person in the white upper garment holding the white toothbrush?	A: yes
Q3: Which is the pink object that the person in the t-shirt is holding, the rag or the tarp?	A: sink
Q4: Is the material of the sink the same as that of the faucet indoors?	A: yes
Q5: Is the standing person holding the brown paint but not the pink rag?	A: no

Figure 11. Example QA pairs from the `train` and `val` splits. Each example contains five QA pairs on the same video with different question structures, *i.e.*, query, verify, choose, compare, and logic.



(a) voting distribution (b) average accuracies

Figure 12. Given the predictions from five individual annotators, we illustrate (a) the distribution of the majority votes and (b) average accuracies with standard deviations in terms of different question structures and the overall type.

	HCRN	ClipBERT	All-in-one
`val`	41.69	44.34	45.44
`test`	41.15	43.92	44.53
`test-dev`	41.18	44.00	44.57

Table 9. Comparative results of the three models which are trained on the `train` split and then evaluated on the `val`, `test`, and `test-dev` splits of ANetQA, respectively.

type	HCRN	ClipBERT	All-in-one
attrRelWhat	24.06	29.03	**29.42**
attrWhat	21.95	26.58	**28.75**
relWhat	16.35	14.59	**16.94**
objRelWhere	15.78	16.81	**16.21**
objRelWhat	19.60	19.36	**22.23**
objWhere	**16.34**	14.25	15.39
objWhat	39.10	39.39	**40.11**
objExist	68.54	72.76	**73.20**
objRelExist	68.00	**71.85**	70.92
actExist	75.34	**78.04**	77.85
objRelWhatChoose	67.09	67.96	**69.13**
objWhatChoose	71.51	77.63	**77.93**
attrRelWhatChoose	56.14	64.60	**65.74**
attrWhatChoose	57.92	65.90	**66.89**
attrCompare	**55.66**	55.60	54.42
attrSame	56.25	**82.14**	58.93
actTime	67.24	**70.44**	56.16
actLongerVerify	50.00	50.00	**52.48**
actShorterVerify	49.79	49.79	**50.83**
andObjRelExist	70.89	70.38	**73.97**
xorObjRelExist	86.50	**89.74**	87.18

Table 10. Per-type accuracy of the three models on the `test` split.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020. 2

[2] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *ACM MM*, pages 797–806, 2021. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[4] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 2, 3, 9

[5] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 2

[7] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585, 2018. 2

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[9] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *CVPR*, pages 11287–11297, 2021. 1, 2, 3, 6, 7

[10] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028, 2020. 2

[11] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 6

[12] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 2

[13] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. 8

[14] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, pages 10267–10276, 2020. 2

[15] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Multi-interaction network with object relation for video question answering. In *ACM MM*, pages 1193–1201, 2019. 2

[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 2

[17] Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *EMNLP*, 2020. 2

[18] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. In *IJCAI*, pages 2016–2022, 2017. 3

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *CVPR*, pages 706–715, 2017. 3, 4, 5

[20] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9972–9981, 2020. 2, 6, 7, 8

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 2, 6, 7

[22] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 9972–9981, 2018. 2, 3

[23] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *ACL*, pages 8211–8225, 2020. 3

[24] 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 *AAAI*, pages 8658–8665, 2019. 2

[25] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *CVPR*, pages 9572–9581, 2019. 2

[26] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *ICCV*, pages 1698–1707, 2021. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 2

[29] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, pages 6884–6893, 2017. 3

[30] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky.

The stanford corenlp natural language processing toolkit. In *ACL*, pages 55–60, 2014. 9

[31] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *CVPR*, pages 2867–2875, 2017. 2, 3

[32] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *ICCV*, pages 677–685, 2017. 2

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. 2

[34] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 2

[35] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016. 2, 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017. 2

[37] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 2, 6, 7, 8, 10

[38] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 3

[39] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, 2022. 2

[40] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, pages 1645–1653, 2017. 2, 3, 6

[41] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 2, 3

[42] Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. What gives the answer away? question answering bias analysis on video qa datasets. *arXiv preprint arXiv:2007.03626*, 2020. 2

[43] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. In *WACV*, pages 1556–1565, 2020. 2

[44] 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 *ICLR*, 2020. 2, 3

[45] 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 *AAAI*, pages 9127–9134, 2019. 2, 3, 6

[46] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019. 2

[47] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018. 2

[48] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, pages 8807–8817, 2019. 3

[49] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 34:23634–23651, 2021. 2

[50] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *AAAI*, 2017. 2, 3

[51] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *ACM MM*, pages 1050–1058, 2017. 2

[52] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pages 6578–6587, 2019. 3, 4, 9

[53] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *IJCV*, 124(3):409–421, 2017. 3