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

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

A. Scene Graph Annotations

A.1. Annotation Pipeline

As mentioned in the main paper, ANetQA is built upon the annotations of ANet-Entities [6], which grounds objects in representative frames with noun phrases (NPs). Nouns and adjectives are extracted from these NPs using the Stanford Parser [4] to form our initial object and attribute vocabularies, respectively. Meanwhile, we handcraft the initial relationship vocabulary on the activity labels of the original ActivityNet [1]. These initial vocabularies are intermittently updated during the annotation process.

We provide a web-based interface shown in Figure 1 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 1, Table 2, and Figure 2, respectively. As our actions are depicted in natural language, we illustrate a word cloud for the most frequent verbs in Figure 3.

*Jun Yu is the corresponding author

A.3. Case Study

In Figure 4, 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 3. 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 4.

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 5. 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 6, 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 7. 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

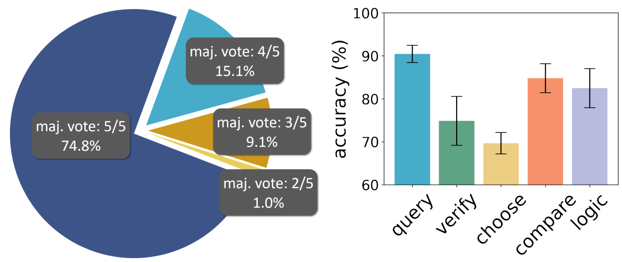
As reported in the main paper, human performance tops out at 84.48% overall accuracy by taking the majority voting over five answers per question. In Figure 8, we provide more detailed analyses of the human evaluation statistics to better understand the behavior of individual annotators. The results in Figure 8a indicate that the deviations among different annotators do exist, and majority voting helps eliminate individual errors. The results in Figure 8b show that different question types lead to diverse accuracies and deviations.

C.2. Val-and-test Consistency

In Table 5, we provide comparisons of the same model on the `val` and `test` split, respectively. The results show that there is no much difference between the performance on the two splits.

C.3. Per-type Accuracy

In Table 6, we report the per-type accuracies of the three models. From the results we can see that the best-performing model All-in-one consistently outperforms the rest models in majority of the question types.



(a) voting distribution (b) average accuracies

Figure 8. 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 [2]	ClipBERT [3]	All-in-one [5]
`val`	41.69	44.34	45.44
`test`	41.15	43.92	44.53

Table 5. Comparative results of the three models on the `val` and `test` 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 6. Per-type accuracy of the three models on the `test` set.

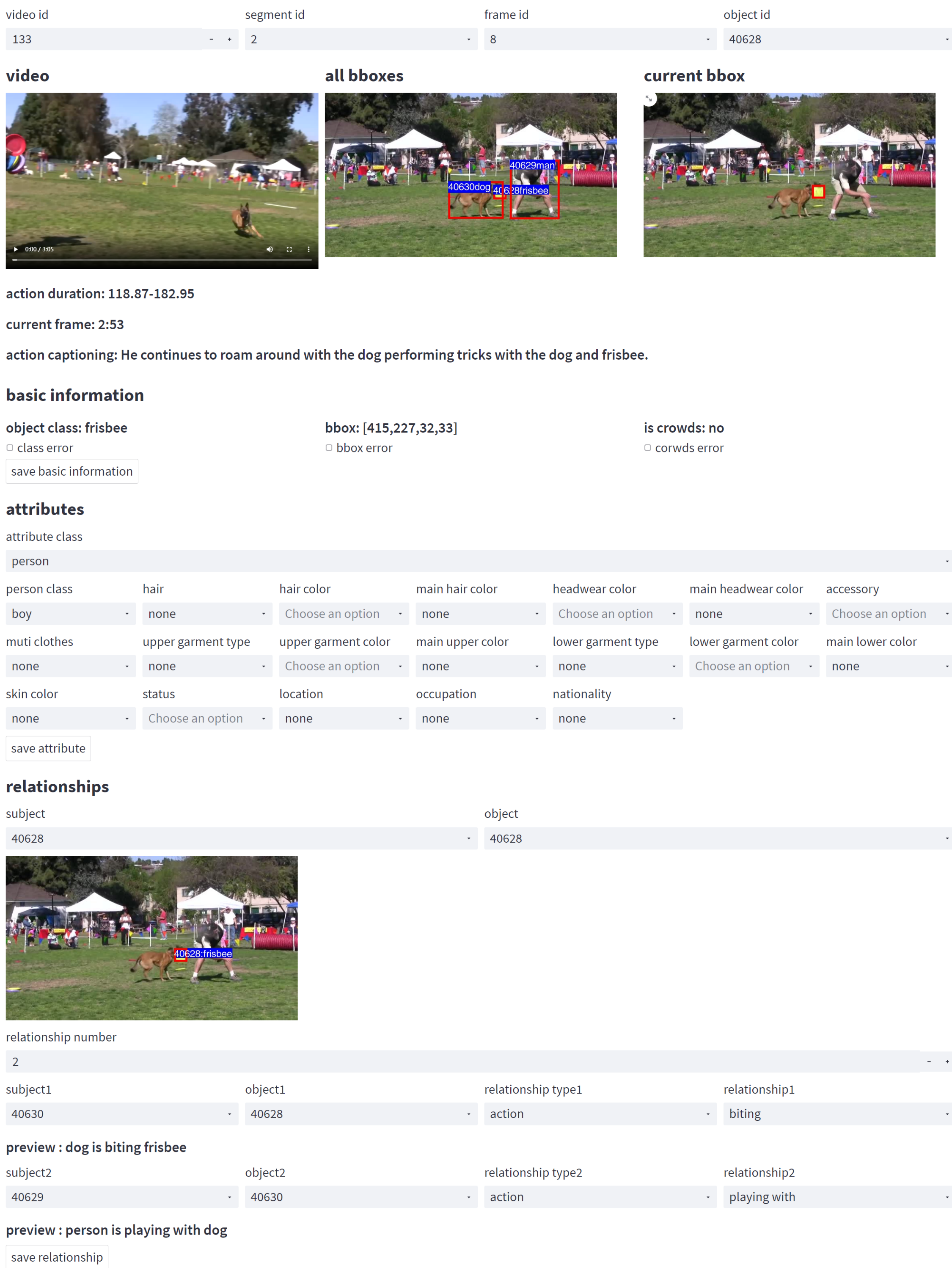


Figure 1. A web-based interface for video scene graph annotation by crowdsourcing. Annotators are asked to watch the video first and then select attributes and relationships from corresponding vocabularies. When no suitable item is available, they can add new items freely. These new items will be manually 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 1. 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 2. 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.

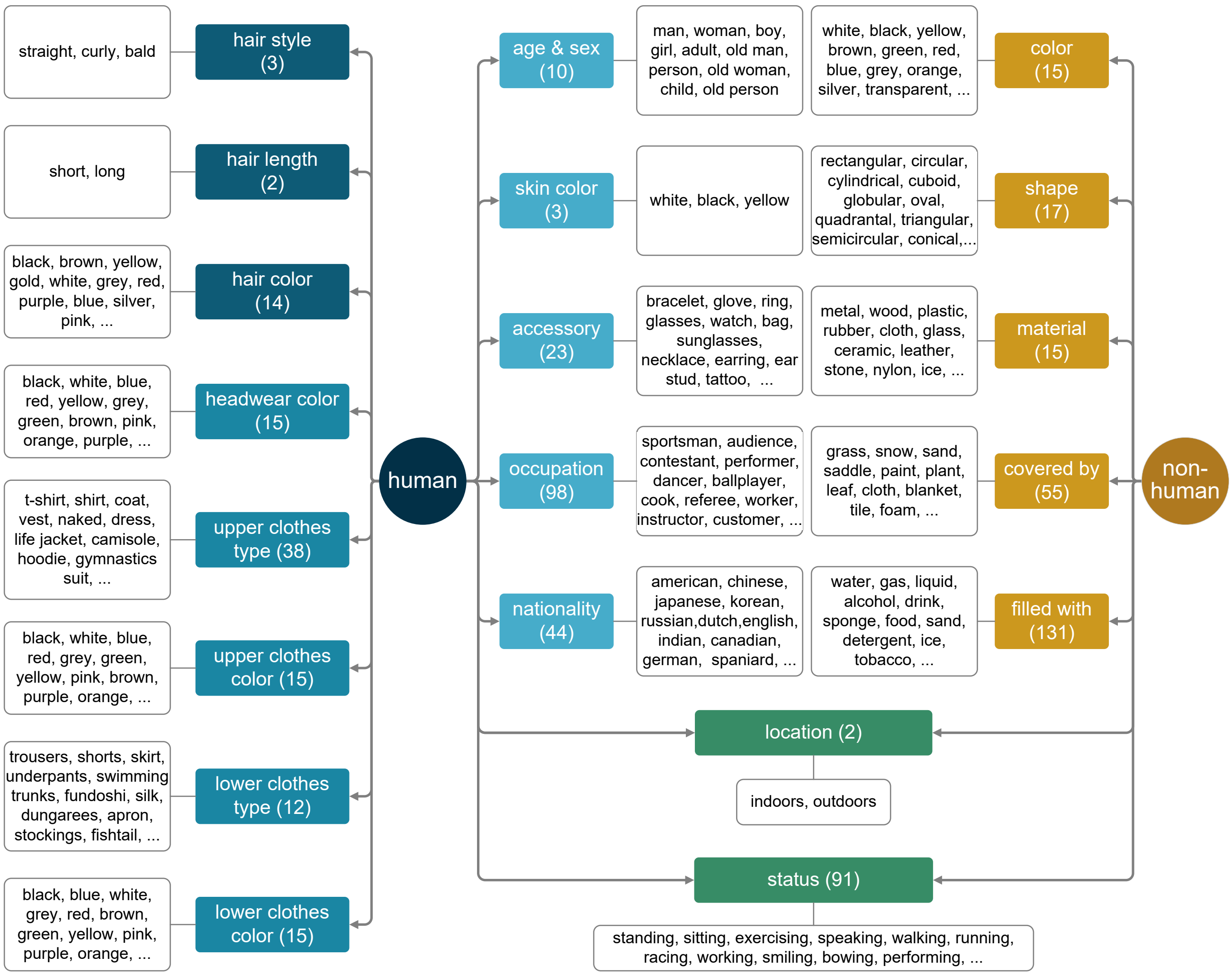


Figure 2. A hierarchy of attributes in our benchmark. The hierarchy consists of three levels. On the **top** level, objects are classified into the *human* and *non-human* groups. On the **middle** level, up to 20 representative attribute types are designed for each top groups (*e.g.*, "*hair style*" 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 3. A word cloud for frequent *verbs* in action descriptions. We merge the words with the same etymon for better visualization.

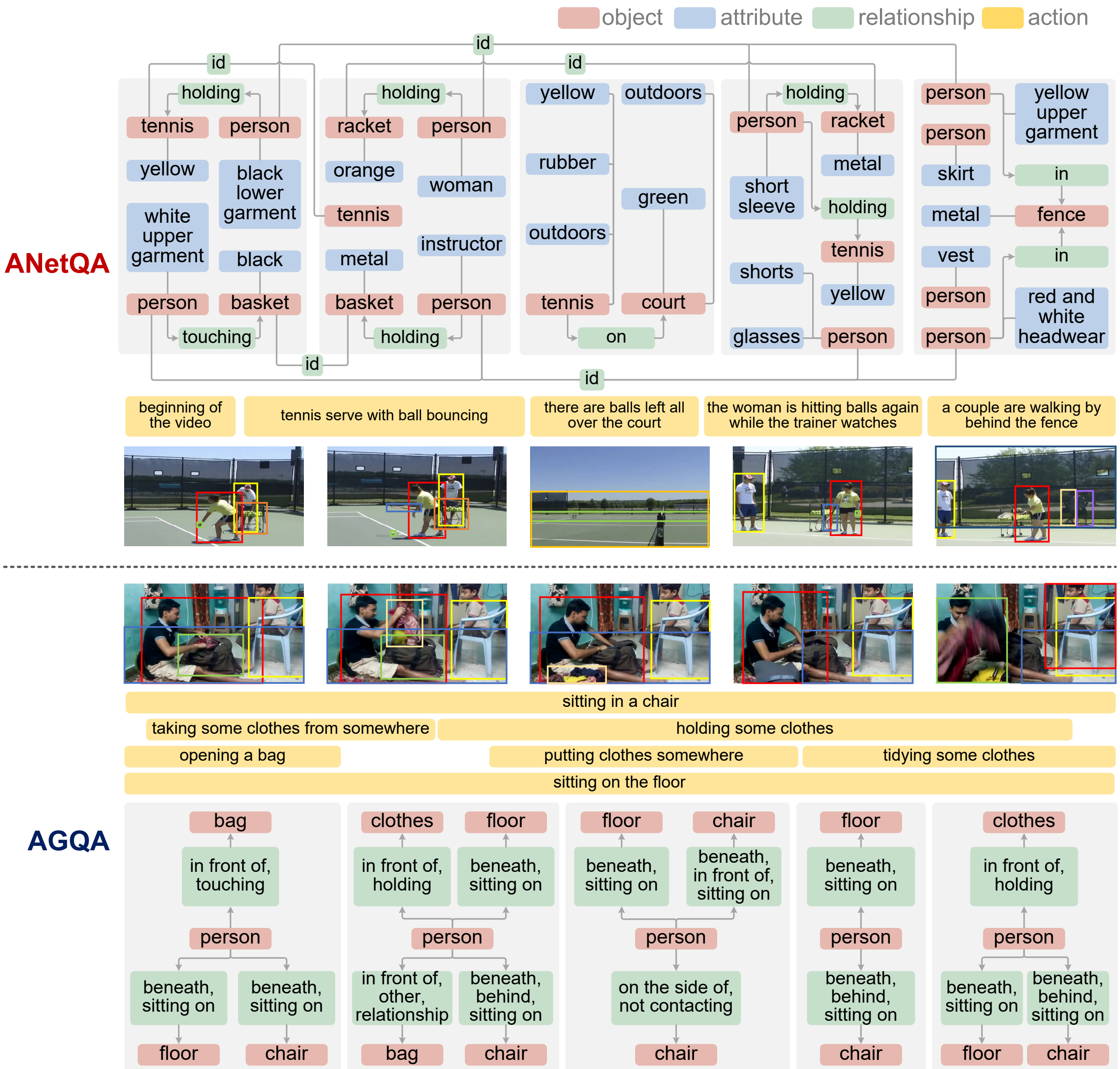


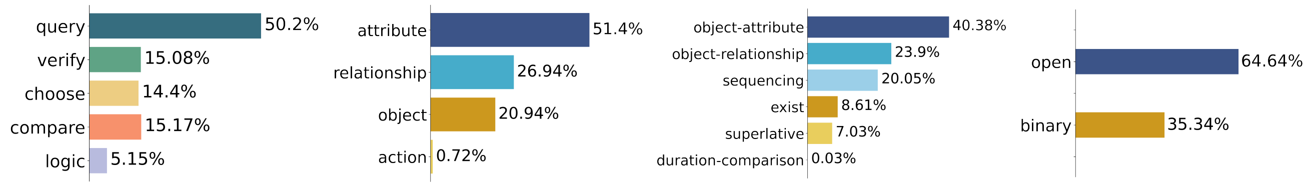
Figure 4. A comparison of the example scene graphs of our ANetQA and AGQA. The visualized results suggest: (i) our scene graph is more informative than that in AGQA as our untrimmed video contains richer semantics with multiple switched scenarios; (ii) our scene graph is much more fine-grained than that 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 3. 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 2. 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 4. 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 5. **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.

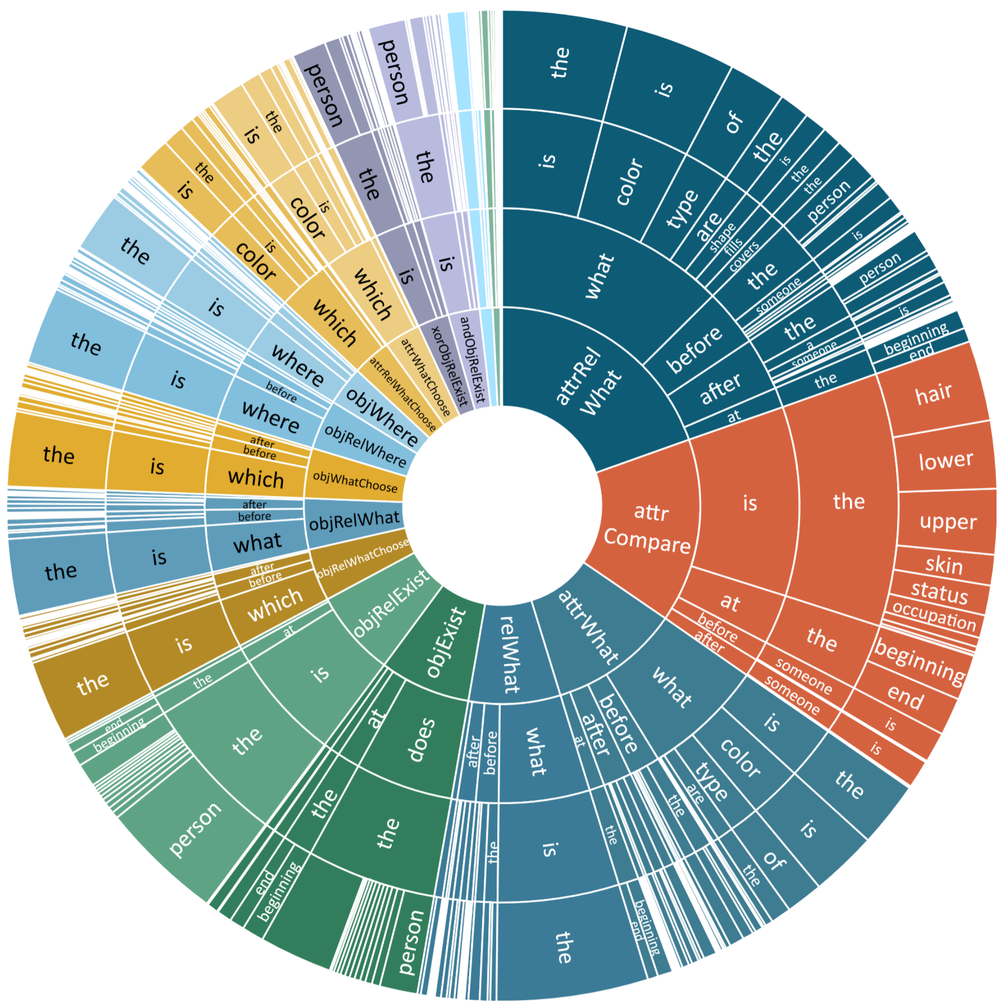
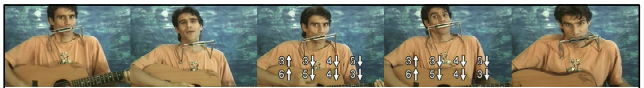
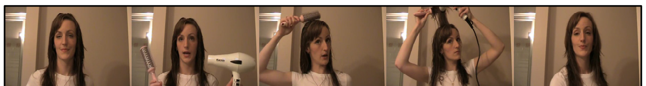


Figure 6. **Question distribution by their first three words** on the balanced benchmark. The innermost ring refers to the 21 question types. The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. For the 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).

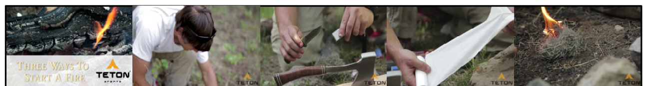


Q1: At the end of the video, what shape is the silver harmonica that the person wearing the t-shirt is playing?	A: cuboid
Q2: Is someone playing the harmonica at the end of the video?	A: yes
Q3: Which color is the upper garment of the person who is performing, yellow or silver?	A: yellow
Q4: Is duration 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 7. 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.

References

[1] 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] 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

[3] 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

[4] 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. 1

[5] 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] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pages 6578–6587, 2019. 1