Compositional Temporal Grounding with Structured Variational Cross-Graph Correspondence Learning Supplementary Material

Anonymous CVPR submission

Paper ID 19

1. Overview

In this supplementary material we present:

- The detailed statistics of the proposed datasets (Section 2).
- Implementation details (Section 3).
- Additional experimental results (Section 4).
- Most common words in query sentences (Section 5).
- Most common novel compositions (Section 6).
- Most common novel words (Section 7).
- Additional examples (Section 8).

2. Dataset Statistics

Table 1 summarizes the detailed statistics of our proposed Charades-CG and ActivityNet-CG datasets.

The distribution of the composition types and their corresponding examples are illustrated in Figure 1. Note that adjective-noun phrases are rare in the original Charades-STA dataset, and most of them are some high-frequency phrases, so the proportion of novel adjective-noun compositions is relatively small in our Novel-Composition set.

3. Implementation Details

For all methods, we use the public official implementations to get their compositional temporal grounding results. We train them on the training set and evaluate them on the test-trivial, novel-composition, and novel-word splits respectively. Following [11], we use unified video and language features for more fair comparisons. Concretely, we use I3D features [2] for the video in Charades-CG and C3D features [10] for the videos in ActivityNet-CG. We use pre-trained GloVe [8] word vectors to initialize each word in the language queries.

In our proposed framework, we use the I3D model [2] pretrained on kinetics [6] dataset as our action detector, and use Faster R-CNN with ResNet-101 [1, 5, 9] pre-trained on Visual Genome [7] dataset as our object detector. For an untrimmed video, we divide it into a sequence of segments with a fixed length (*i.e.* 32 frames), and then adopt the off-the-shelf object and action detectors to extract objects and actions for each segment. For each segment, we select the top-3 action classes and top-5 object classes with the highest confidence score as action nodes and object nodes, respectively. The dimension of input video features is 1024 and the dimension of GloVe [8] vectors is 300. We set the dimension of all node (three hierarchies of the two graphs) representations as 384. During training, we set the batch size to 32 and use Adam as optimizer [3], where the learning rate is set to $1e^{-4}$.

4. Additional Experimental Results

We present the compositional temporal grounding performance of the CTRL [4] and SCDM [12] in Table 2.

5. Most Common Words in Query Sentences

Table 3 and Table 4 show the most common nouns, verbs, adjectives, adverbs, and prepositions, respectively.

6. Most Common Novel Compositions

We show the most common novel compositions in Table 5 and Table 6.

7. Most Common Novel Words

Table 7 and Table 8 show the most common novel words.

8. Additional Examples

Figure 2 and Figure 3 show some more examples in the novel-composition and novel-word splits of the Charades-CG dataset. Figure 4 and Figure 5 show some more exam-

Dataset	Split	Videos	Average Video Length	Queries	Average Query Length
Charades-CG	Training	3555	30.58s	8281	5.93
	Novel-Composition	2480	30.70s	3442	6.86
	Novel-Word	588	31.26s	703	7.24
	Test-Trivial	1689	30.82s	3096	5.96
ActivityNet-CG	Training	9659	116.94s	36724	13.33
	Novel-Composition	4202	121.12s	12028	14.78
	Novel-Word	2011	124.35s	3944	14.61
	Test-Trivial	4775	119.60s	15712	11.31

Table 1. Statistics of Charades-CG and Activity-CG.

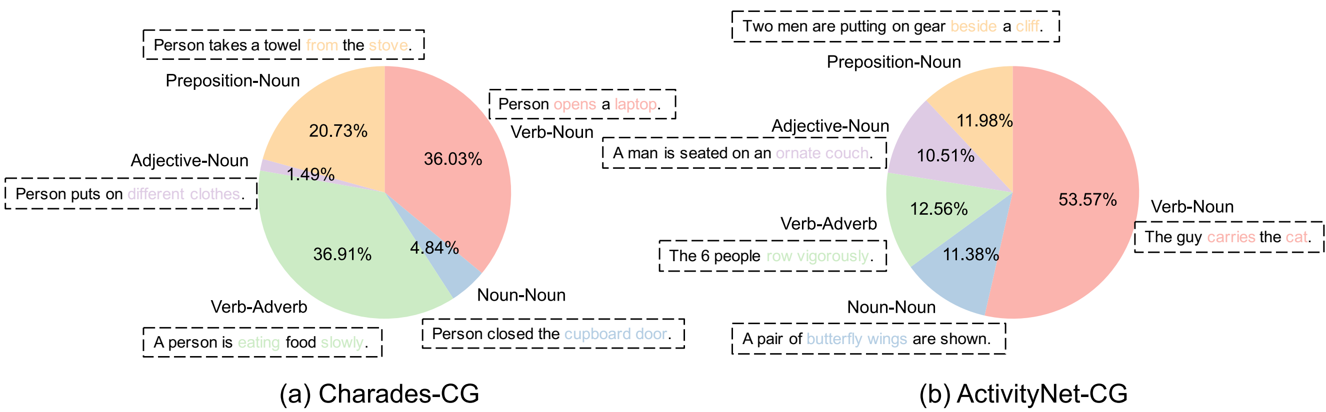

Figure 1. The distribution of the composition types. Texts inside dashed boxes are query examples for each composition type.

Method	Dataset	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
CTRL	Charades-CG	18.53	8.59	22.03	4.62	0.17	11.21	4.22	0.22	10.60
	ActivityNet-CG	13.25	4.49	17.51	5.22	1.55	11.21	5.17	1.59	11.17
SCDM	Charades-CG	46.63	24.17	42.08	27.73	12.25	30.84	26.20	11.69	27.64
	ActivityNet-CG	37.86	22.41	40.09	21.32	9.34	28.52	20.73	8.95	27.46

Table 2. Additional experimental results of CTRL and SCDM on the Charades-CG and ActivityNet-CG datasets.

ples in the novel-composition and novel-word splits of the Charades-CG dataset.

Type	Most Common Words
Noun	person, door, light, glass, book, shoe, bag, table, food, sandwich, box, cabinet, chair, laptop, window, cup, shelf, room, clothes, floor, phone, pillow, water, doorway, closet, picture, bed, refrigerator, blanket, towel
Verb	be, put, open, take, eat, close, sit, hold, turn, run, throw, drink, start, begin, walk, sneeze, look, laugh, smile, pour, stand, cook, undress, watch, awaken, wash, dress, fix, read, play
Adjective	open, other, laundry, second, same, front, small, light, nearby, few, dressed, more, plastic, dirty, first, multiple, undress, large, different, undressed, several, entryway, close, red, next, oven, folded, full, closed, little
Adverb	away, back, inside, next, also, finally, around, again, when, quickly, outside, so, down, aside, in, there, where, repeatedly, suddenly, out, still, nearby, on, twice, slowly, just, very, by, off, later
Preposition	on, in, of, from, off, into, down, at, up, out, to, with, through, onto, by, as, behind, over, for, under, around, towards, away, after, across, inside, toward, against, outside, past

Table 3. Most common words of query sentences in the Charades-CG dataset.

Type	Most Common Words
Noun	man, woman, people, camera, person, girl, ball, men, hand, water, boy, screen, front, group, side, dog, hair, lady, field, table, room, shirt, game, car, video, shot, floor, time, bar, horse
Verb	be, show, see, play, stand, walk, continue, hold, talk, begin, do, sit, put, speak, use, jump, run, take, go, get, throw, move, watch, start, rid, wear, hit, look, make, appear
Adjective	several, other, white, large, more, black, young, small, blue, red, various, little, different, green, close, long, high, same, yellow, many, old, slow, few, first, right, wooden, ready, pink, big, fourth
Adverb	then, back, around, again, how, as, well, next, together, away, when, still, outside, all, down, very, after, finally, forth, also, where, now, up, quickly, over, forward, more, once, just, slowly
Preposition	in, of, on, with, to, up, into, as, down, at, around, off, by, from, over, out, for, behind, onto, before, after, along, through, about, across, inside, towards, under, against, between

Table 4. Most common words of query sentences in the ActivityNet-CG dataset.

Type	Novel Compositions
Verb-Noun	throw pillow, open laptop, close laptop, pour coffee, tidy wardrobe, close window, throw shoe, throw book, throw box, watch car, watch laptop, wash window, carry towel, throw broom, wash table
Adjective-Noun	different clothes, young woman, few grocery, small closet, same window, bottom shelf, several drink, small refridgerator, small desk, bottom cabinet, young guy, different person, more soda, few notebook, near doorway
Preposition-Noun	with towel, in wardrobe, around box, on head, into hallway, towards table, for work, inside cabinet, under desk, through hall, outside door, onto wall, above stove, over top, past door way
Noun-Noun	medicine bottle, cupboard door, closet doorknob, kitchen pantry, laptop bag, work clothes, bathroom shelf, towel rack, wine glass, phone camera, shower curtain, food bag, detergent cabinet, food dish, kitchen doorway
Verb-Adverb	awaken suddenly, come suddenly, eat slowly, sneeze repeatedly, awaken quickly, throw repeatedly, undress partially, dress quickly, read intensely, look back, dress again, smile together, open twice, look over, come out

Table 5. We present 15 common novel compositions of each type in the Charades-CG dataset.

Type	Novel Compositions
Verb-Noun	pull row, advertise event, boil noodle, announce winner, pick hose, wipe boot, see boxer, push rake, extend palm, lower cap, find friend, park bike, remove plastic, leave chair, fill basket,
Adjective-Noun	live music, cold river, large museum, red vase, different pumpkin, golden coin, vacant kitchen, crowded stage, dry land, furry dog, tiny fish, messy bedroom, wooden shelf, strange costume, green plate
Preposition-Noun	towards sea, beside box, behind building, against target, around lady, after stone, below surface, with certificate, through pipe, in cabinet, on logo, into store, along ridge, until stop, without teacher
Noun-Noun	hockey tournament, girl referee, sea turtle, kid playground, foot pedal, princess costume, baby shark, farm building, group selfie, race trail, fire stick, sugar mixture, stone tunnel, art skill, music book
Verb-Adverb	add directly, sit backward, fly away, move vigorously, work carefully hit immediately, leave suddenly, plan carefully, remove quickly, groom cleanly, go speedily, aim accurately, play passionately, kick repeatedly, complete successfully

Table 6. We present 15 common novel compositions of each type in the ActivityNet-CG dataset.

Type	Novel Words
Verb	talk, bend, prepare, cover, kick, prepare, need, stretch, let, struggle, slide, toss, encounter, drop, pack, burn, cause, examine, swing, lift
Noun	hand, stair, hair, dryer, corner, tissue, stack, cave, basket, dinner, arm, reflection, remote, tool, coat, sheet, bucket, wrapper, cap, napkin
Adjective	old, funny, white, bright, fresh, dusty, hot, sick, canvas, stray, dim, rampant, original, visible, confused, own, humorous, favorite, loud, short
Adverb	somewhere, well, slightly, furiously, periodically, constantly, freshly, intently, really, downstairs, randomly, thoughtfully, everywhere, continuously, gently, lastly, simultaneously, somewhat, shortly, often

Table 7. We present 20 common novel words of each type in the Charades-CG dataset.

Type	Novel Words
Verb	mop, encase, utter, moisten, analyze, encase, cling, radiate, cultivate, disconnect, retouch, originate, matter, reconstruct, resist, boost, endorse, identify, surpass, consume
Noun	bull, camel, pumpkin, dart, shield, carpenter, theory, hunting, washboard, killer, priest, ox, rapper, hero, donkey, squid, extreme, physician, vacancy, mansion
Adjective	solitary, vivid, over, religious, acceptable, exotic, structural, glad, foggy, horrified, married, sequential, improper, evident, functional, european, hydraulic, strategic, mechanic, early
Adverb	accurately, improperly, carelessly, confidently, identically, absolutely, remotely, cautiously, regardless, recently, anyway, furthermore, inwards, luxuriously, erratically, vividly, poorly, anyhow, whenever, greatly

Table 8. We present 20 common novel words of each type in the ActivityNet-CG dataset.

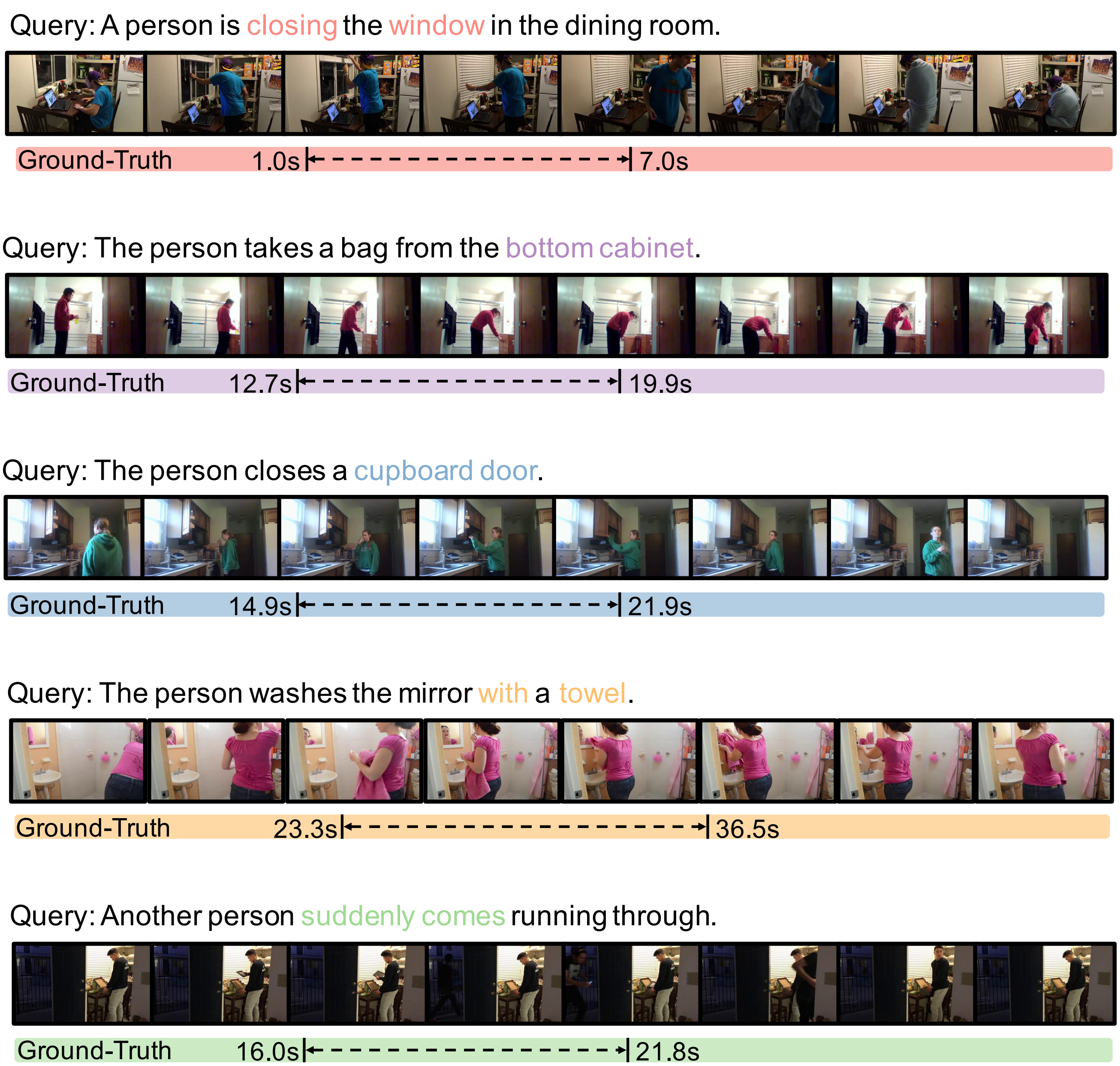

Figure 2. Examples in the novel-composition split of the Charades-CG dataset.

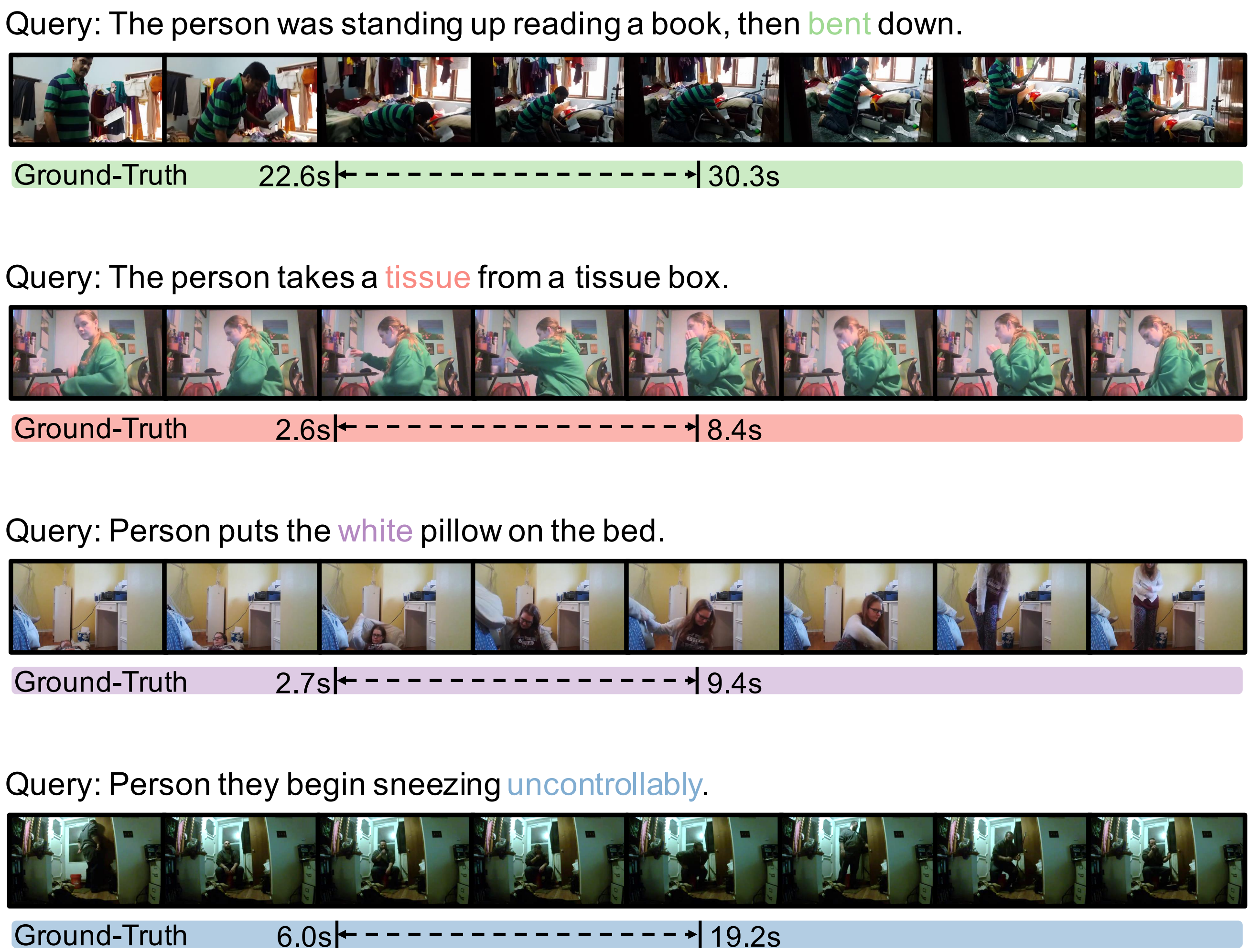

Figure 3. Examples in the novel-word split of the Charades-CG dataset.

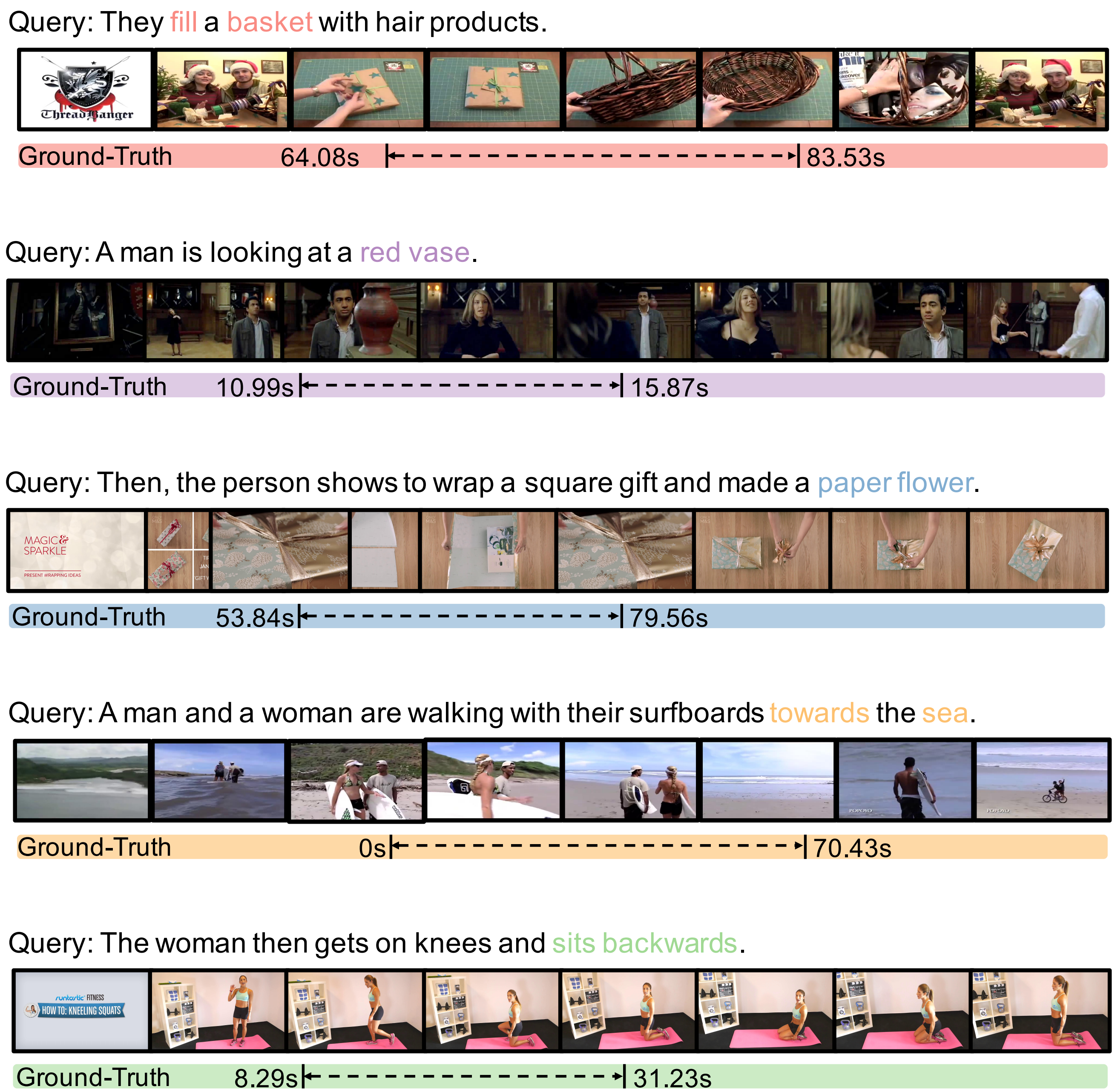

Figure 4. Examples in the novel-composition split of the ActivityNet-CG dataset.

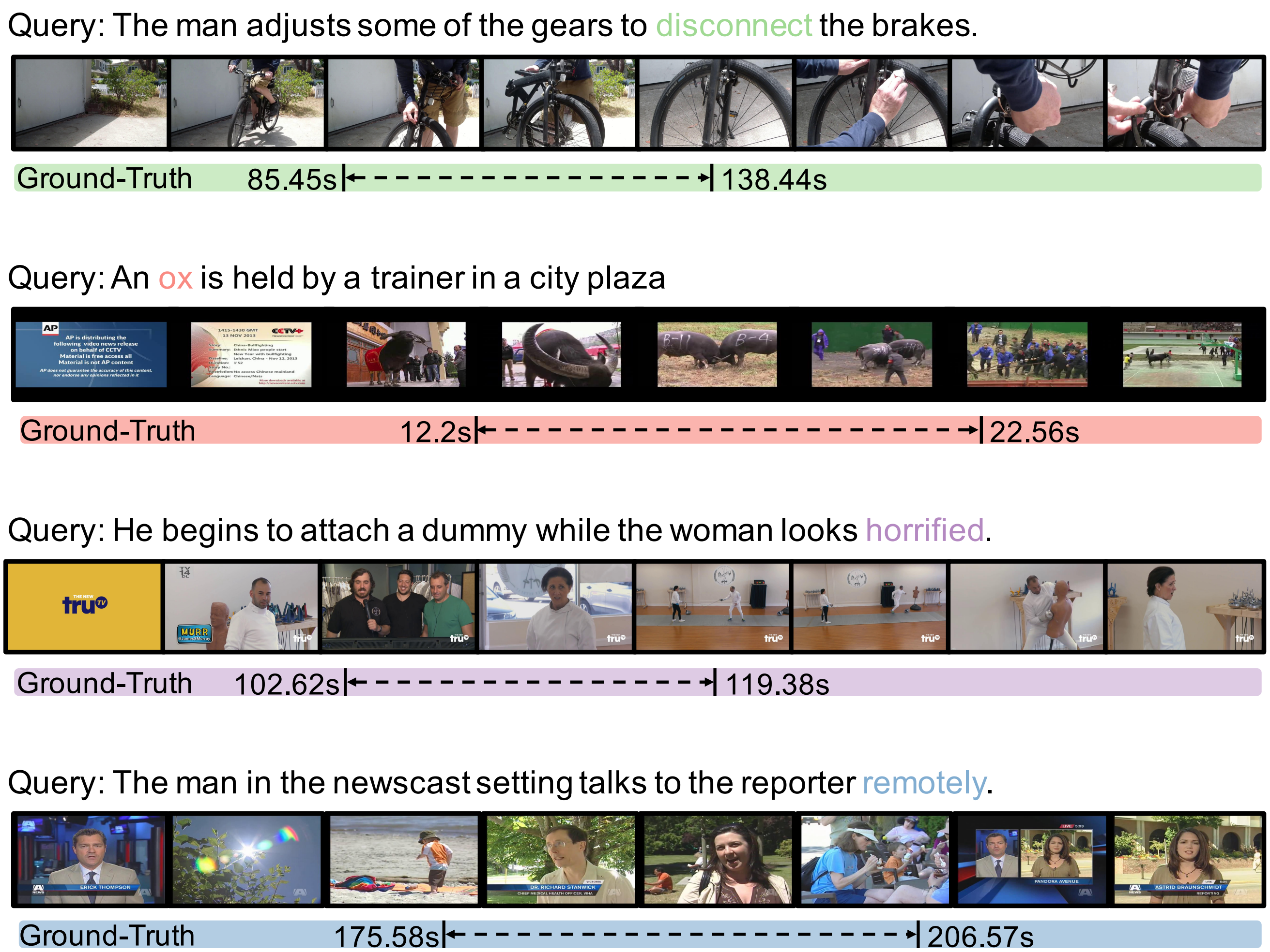

Figure 5. Examples in the novel-word split of the ActivityNet-CG dataset.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[3] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 1

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[11] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 1

[12] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *arXiv preprint arXiv:1910.14303*, 2019. 1