Visual Relation Grounding in Videos - Supplementary Material

1 Implementation Details

Trajectory Generation. In the test phase, to group continuous frames into candidate segments, two adjacent frames will be grouped as long as the distance between them are smaller than a value of $D_{group} = max(1, 10 * r)$, where r denotes the frame sample rate of the video, and the maximal sample distance is $\frac{1200}{120} = 10$ frames in the dataset. Besides, to obtain the bounding boxes on un-sampled frames, we adopt linear interpolation as follows. Given two bounding boxes of a trajectory B_k and B_{k+1} corresponding to the sampled frames F_k and F_{k+1}, we achieve the bounding boxes in between by

$$B_c = \frac{F_{k+1} - F_c}{F_{k+1} - F_k} * B_k + \frac{F_c - F_k}{F_{k+1} - F_k} * B_{k+1},\ F_c \in (F_k, F_{k+1}) \tag{1}$$

where B_c denotes the bounding box in frame F_c. Since the maximal distance between two sampled frames is 10 (about 1/3 seconds), it is reasonable to think that the objects move linearly in such a short time.

Baselines. In both implementation variants of WSSTG [1], we extract 15 trajectory proposals (resulting in 15*14 = 210 pairs) for each video based on the frame-level region proposals. Each trajectory is then evenly divided into 20 sub-segments, with each segment represented by region appearance feature (average across frames) and sequence feature I3D-RGB as in [1]. Then, the trajectories are modeled with LSTM and interacted with the query sentence to get the similarities between them.

2 Additional Results

To evaluate the models' performances on different kinds of relationships, we classify the relationships defined in ImageNet-VidVRD [3] into dynamic and static ones (Refer to Sec. 3) and separately report results on them. As a result, we obtain 2343 video-relation instances for static relationships and 2492 for dynamic ones, both covering the 200 test videos.

As shown in Table 1, the results on static relationships are better than those on dynamic ones. When compare our method with the baseline approaches, we can achieve consistently better results as in the main text. We speculate the reason is that moving objects will result in deformation, motion blur and occlusion which are very challenging for grounding. Fortunately, the online optimization mechanism in our approach exploits the relationships to pinpoint some objects with the condition of their related partners, and thus cope better with the dynamic scenario than the baselines. This speculation is also supported by the

Table 1. Grounding results on dynamic and static relationships.

Methods	Dynamic			Static			Overall		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V_1 [1]	15.19	8.27	3.40	24.05	11.36	4.96	20.27	10.68	3.99
T-Rank V_2 [1]	15.81	5.28	1.89	23.56	7.96	4.18	20.83	7.35	3.16
Co-occur [2]	20.05	21.21	13.81	33.43	30.82	21.4	25.90	25.23	16.48
vRGV (ours)	**32.47**	**32.86**	**22.7**	**37.53**	**36.51**	**26.15**	**36.77**	**36.30**	**24.58**

Table 2. Grounding results under different temporal overlap thresholds.

Methods	tIoU=0.3			tIoU=0.5			tIoU=0.7		
	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R	Acc_S	Acc_O	Acc_R
T-Rank V_1 [1]	36.51	28.67	15.05	20.27	10.68	3.99	6.15	2.67	0.55
T-Rank V_2 [1]	36.99	20.70	12.81	20.83	7.35	3.16	6.19	1.30	0.21
Co-occur [2]	35.30	35.50	23.23	25.90	25.23	16.48	16.81	15.04	8.94
vRGV (ours)	**49.97**	**48.98**	**33.16**	**36.77**	**36.30**	**24.58**	**24.27**	**22.11**	**13.69**

observation on the co-occurrence baseline that the performance drops significantly from 21.4% to 13.81% from static to dynamic scenario, without modeling of the relationships between objects.

We also compare the models on different temporal overlap thresholds in Table 2, from which we can draw similar conclusion as in the main text that our method shows superiority to the compared baselines under different settings. As shown in Fig. 1, our method can reasonably ground the relation in both space and time. Taking (a) for instance, the learned temporal attention regarding the query relation *dog-walk_left-turtle* is small when the dog is walking on the right of the turtle, and then it becomes larger when the dog is walking on the left[1]. Besides, the model can successfully pinpoint the subject *dog* and object *turtle* on the grassland most of the time. Similarly, in (b), our method succeeds in finding the relation *person-stand_above-bicycle* when the person is standing above the bicycle but not sitting on it after he goes up the slope.

We specially analyze some failing cases in Fig. 1. For the relation *elephant-kick-ball* in (g), our method fails to stop the grounding even though the relation is disappeared, and thus results in false positives. This could be due to that the relation is very transient and further the ball is too small to localize, and hence brings in great challenges in spatio-temporal grounding. In (h), our method wrongly grounds the subject and object on the same visual entity when the subject and object belong to the same category (*e.g.*, horse). We speculate the reason is that we directly take the textual representations of the two words which are indistinguishable in feature space as semantic clues to retrieve the related visual subject and object in the spatial attention unit. Thus, the unit is prone

[1] According to the dataset definition, the spatial relationships are in camera view.

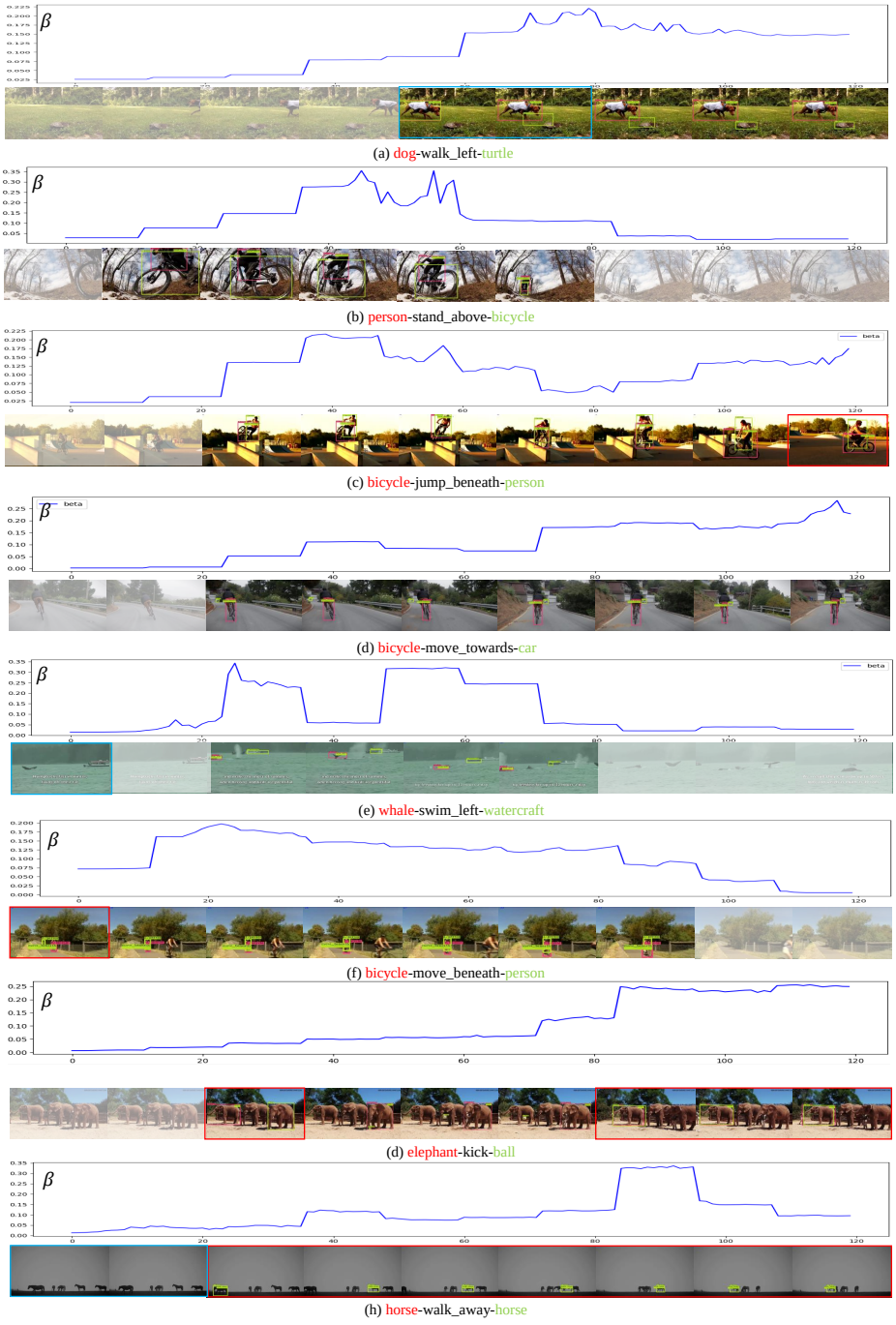
to get a redundant location of the subject as the object. Nevertheless, our model can disambiguate the visual entities of same category in the scenarios where one entity matches the query relation and others do not match. For example, in (c) and (f), there are other people present in the video, but our method can precisely find the person with the bicycle jumps (in (c)) or moves (in (f)) beneath him. Similarly in (d) and (e), there are two cars (d) or two watercrafts (e) present in the videos, and our methods can successfully ground the car/watercraft that satisfy the respective relations.

3 Dataset

There are 35 objects and 132 predicates (relationships) defined in the ImageNet-VidVRD [3] dataset. The **objects** are: *turtle, antelope, bicycle, lion, ball, motorcycle, cattle, airplane, red panda, horse, watercraft, monkey, fox, elephant, bird, sheep, frisbee, giant panda, squirrel, bus, bear, tiger, train, snake, rabbit, whale, sofa, skateboard, dog, domestic cat, person, lizard, hamster, car, zebra.* The predicates can be classified into static relationships and dynamic ones. The **static relationships** include: *above, beneath, left, right, front, behind, taller, larger, next to, inside, hold, bite, lie above, lie beneath, lie left, lie right, lie inside, lie next to, lie with, stand above, stand beneath, stand left, stand right, stand front, stand behind, stand next to, stand inside, sit above, sit left, sit right, sit front, sit behind, sit next to, sit inside, stop above, stop beneath, stop left, stop right, stop front, stop behind, stop next to, stop with.* The **dynamic relationships** include: *swim behind, walk away, fly behind, creep behind, move left, touch, follow, move away, walk with, move next to, creep above, fall off, run with, swim front, walk next to, kick, creep right, watch, swim with, fly away, creep beneath, run past, jump right, fly toward, creep left, run next to, jump front, jump beneath, past, jump toward, walk beneath, run away, run above, walk right, away, move right, fly right, run front, run toward, jump past, jump above, move with, swim beneath, walk past, run right, creep away, move toward, feed, run left, fly front, walk behind, fly above, fly next to, fight, walk above, jump behind, fly with, jump next to, run behind, move behind, swim right, swim next to, move past, pull, walk left, ride, move beneath, toward, jump left, creep toward, fly left, walk toward, chase, creep next to, fly past, move front, run beneath, creep front, creep past, play, move above, faster, walk front, drive, swim left, jump away, jump with.*

References

1. Chen, Z., Ma, L., Luo, W., Wong, K.Y.K.: Weakly-supervised spatio-temporally grounding natural sentence in video. ACL (2019)
2. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6867–6876 (2018)
3. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1300–1308 (2017)



(a) dog-walk_left-turtle

(b) person-stand_above-bicycle

(c) bicycle-jump_beneath-person

(d) bicycle-move_towards-car

(e) whale-swim_left-watercraft

(f) bicycle-move_beneath-person

(d) elephant-kick-ball

(h) horse-walk_away-horse

Fig. 1. Visualization of relation grounding results. The false-positive and false-negative frames are highlighted with red and blue rectangles respectively.