Text-based Temporal Localization of Novel Events (Supplementary Document)

Sudipta Paul[1], Niluthpol Chowdhury Mithun[2], and Amit K Roy-Chowdhury[1]

[1] University of California, Riverside CA USA
[2] SRI International, Princeton NJ USA
{spaul, amitrc}@ece.ucr.edu, niluthpol.mithun@sri.com

1 Significance of the Problem Setting

Figure 1 illustrates the significance of our problem setting. We evaluate the performance of a trained text-based temporal localization model for both seen events/queries and unseen events/queries. For Charades-STA Unseen, we consider SCDM [5], which predicts the overlap score and temporal offset directly based on candidate moment representation. For ActivityNet Captions Unseen dataset, we consider 2D-TAN [6], which also predicts overlap scores based on candidate moment representation directly. We observe that there is a significant difference of performance between seen events/queries and unseen events/queries for both datasets. It demonstrates the requirement of a system that can retain the performance for seen queries and improve the performance for unseen queries.

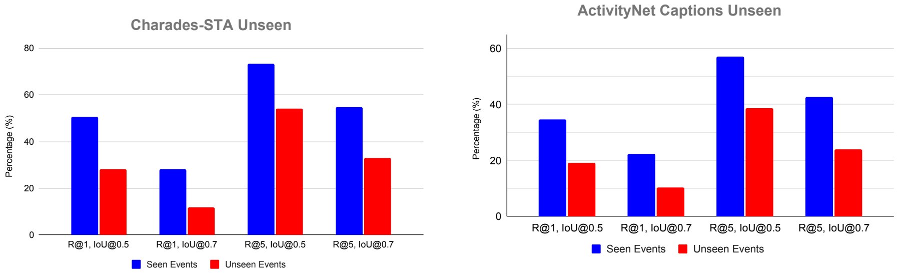


Fig. 1: This figure illustrates the performance of SCDM [5] for Charades-STA Unseen and 2D-TAN [6] for ActivityNet Captions Unseen dataset for seen events and unseen events. For both datasets, performance of the trained model drops significantly for unseen events.

2 Details of the Reorganized Datasets

As existing benchmark temporal moment localization datasets are not designed for the task of temporal localization of moments for unseen queries, we reorganize two of the benchmark dataset namely Charades-STA [1] and ActivityNet Captions [3] to create Charades-STA Unseen and ActivityNet Captions Unseen according to our problem setting. Table 1 reports the number of moment-text pairs for training, unseen testing and seen testing splits of both datasets. Table 2 reports the number of videos in each split of both datasets. Table 3 reports the number verbs and nouns used to create the splits of both datasets. For both datasets, we create splits based on the verbs and nouns present in the text queries. First, we combine all the annotations of the trainset and testset videos of the base dataset. To create the splits, we consider a set of verbs (n_V) and nouns (n_N) present in the combined annotation. Then, we identify videos that contain at least a single query that has a verb or noun not present in the mentioned set. In the selected

Table 1: Tabulated summery of **number of moment-text pairs** in Charades-STA Unseen and ActivityNet Captions Unseen dataset.

Dataset	**Training**	**Unseen Testing**	**Seen Testing**
Charades-STA Unseen	5525	1665	867
ActivityNet Captions Unseen	5669	2553	710

Table 2: Tabulated summery of **number of videos** in the reorganized Charades-STA Unseen and ActivityNet Captions Unseen dataset.

Dataset	**Training**	**Unseen Testing**	**Seen Testing**
Charades-STA Unseen	3366	1271	486
ActivityNet Captions Unseen	3939	1993	513

Table 3: **Number of verbs and nouns** used to create train/test splits of Charades-STA Unseen and ActivityNet Captions Unseen dataset.

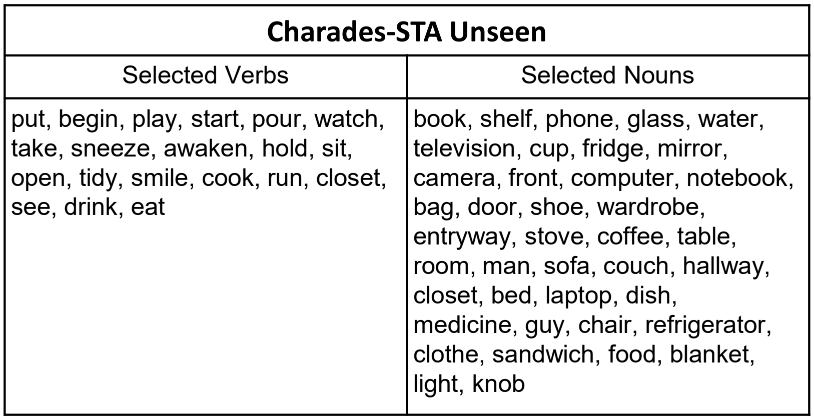
Dataset	**Number of Verbs**	**Number of Nouns**
Charades-STA Unseen	20	40
ActivityNet Captions Unseen	70	250

videos, queries which do not have verbs or nouns from the mentioned set are collected as unseen testset split and, queries which have verbs or nouns from the mentioned set are collected as seen testset split. The training set is created from the rest of the videos, with queries that contain either verb or noun present in the mentioned set. We use spaCy [2] to parse verbs and nouns from text queries.

Charades-STA Unseen. For Charades-STA Unseen, we consider $n_V = 20$ and $n_N = 40$ (excluding 'person' noun). In this way, we have Charades-STA Unseen dataset with 5525, 1665, and 867 training, unseen testing, and seen testing moment-sentence pairs respectively. The number of videos in the training, unseen testing and seen testing splits are 3366, 1271, and 486 respectively. The list of verbs and nouns used are given in Figure 2 for reproducibility.

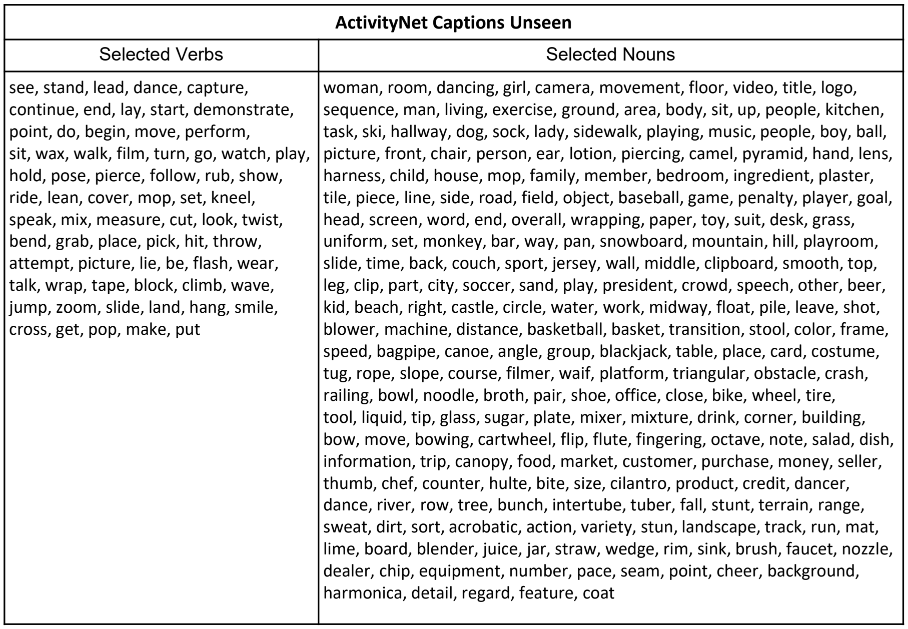
ActivityNet Captions Unseen. For ActivityNet Cations Unseen dataset, we consider $n_V = 70$ and $n_N = 250$. In this way, we have ActivityNet Captions Unseen dataset with 5669, 2553, and 710 training, unseen testing, and seen testing moment-sentence pairs respectively. The number of videos in the training, unseen testing and seen testing splits are 3939, 1993, and 513 respectively. The list of verbs and nouns used are given in Figure 3 for reproducibility.

Fig. 2: List of selected verbs and nouns for Charades-STA Unseen.



Charades-STA Unseen	
Selected Verbs	Selected Nouns
put, begin, play, start, pour, watch, take, sneeze, awaken, hold, sit, open, tidy, smile, cook, run, closet, see, drink, eat	book, shelf, phone, glass, water, television, cup, fridge, mirror, camera, front, computer, notebook, bag, door, shoe, wardrobe, entryway, stove, coffee, table, room, man, sofa, couch, hallway, closet, bed, laptop, dish, medicine, guy, chair, refrigerator, clothe, sandwich, food, blanket, light, knob

Fig. 3: List of selected verbs and nouns for ActivityNet Captions Unseen.



ActivityNet Captions Unseen	
Selected Verbs	Selected Nouns
see, stand, lead, dance, capture, continue, end, lay, start, demonstrate, point, do, begin, move, perform, sit, wax, walk, film, turn, go, watch, play, hold, pose, pierce, follow, rub, show, ride, lean, cover, mop, set, kneel, speak, mix, measure, cut, look, twist, bend, grab, place, pick, hit, throw, attempt, picture, lie, be, flash, wear, talk, wrap, tape, block, climb, wave, jump, zoom, slide, land, hang, smile, cross, get, pop, make, put	woman, room, dancing, girl, camera, movement, floor, video, title, logo, sequence, man, living, exercise, ground, area, body, sit, up, people, kitchen, task, ski, hallway, dog, sock, lady, sidewalk, playing, music, people, boy, ball, picture, front, chair, person, ear, lotion, piercing, camel, pyramid, hand, lens, harness, child, house, mop, family, member, bedroom, ingredient, plaster, tile, piece, line, side, road, field, object, baseball, game, penalty, player, goal, head, screen, word, end, overall, wrapping, paper, toy, suit, desk, grass, uniform, set, monkey, bar, way, pan, snowboard, mountain, hill, playroom, slide, time, back, couch, sport, jersey, wall, middle, clipboard, smooth, top, leg, clip, part, city, soccer, sand, play, president, crowd, speech, other, beer, kid, beach, right, castle, circle, water, work, midway, float, pile, leave, shot, blower, machine, distance, basketball, basket, transition, stool, color, frame, speed, bagpipe, canoe, angle, group, blackjack, table, place, card, costume, tug, rope, slope, course, filmer, waif, platform, triangular, obstacle, crash, railing, bowl, noodle, broth, pair, shoe, office, close, bike, wheel, tire, tool, liquid, tip, glass, sugar, plate, mixer, mixture, drink, corner, building, bow, move, bowing, cartwheel, flip, flute, fingering, octave, note, salad, dish, information, trip, canopy, food, market, customer, purchase, money, seller, thumb, chef, counter, hulte, bite, size, cilantro, product, credit, dancer, dance, river, row, tree, bunch, intertube, tuber, fall, stunt, terrain, range, sweat, dirt, sort, acrobatic, action, variety, stun, landscape, track, run, mat, lime, board, blender, juice, jar, straw, wedge, rim, sink, brush, faucet, nozzle, dealer, chip, equipment, number, pace, seam, point, cheer, background, harmonica, detail, regard, feature, coat

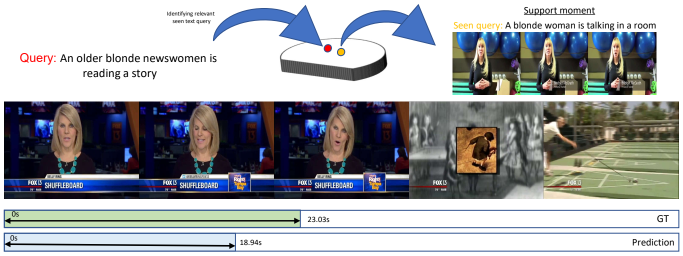

Identifying relevant seen text query
Support moment
Seen query: A blonde woman is talking in a room
Query: An older blonde newswomen is reading a story
FOX 13 SHUFFLEBOARD
0s
23.03s
GT
0s
18.94s
Prediction

Fig. 4: Example illustration from ActivityNet Captions Unseen, where splits are created based on activity annotation. Given the text query ‘An older blonde newswomen is reading a story’ and the corresponding video, our proposed approach retrieve moment corresponding to semantically relevant query ‘A blonde woman is talking in a room’ from the train set, reason on that and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates predicted temporal endpoints of our approach.

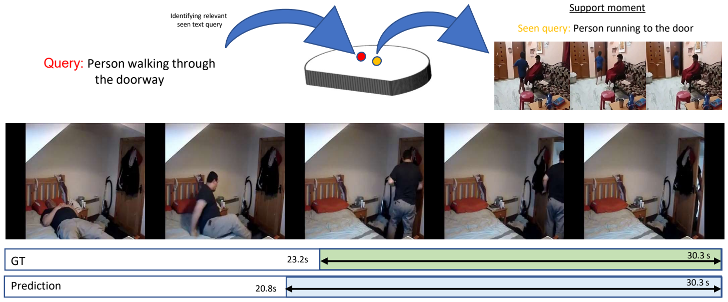

Identifying relevant seen text query
Support moment
Seen query: Person running to the door
Query: Person walking through the doorway
GT
23.2s
30.3 s
Prediction
20.8s
30.3 s

Fig. 5: Given the text query ‘Person walking through the doorway’ and the corresponding video, our proposed approach retrieve moment corresponding to semantically relevant query ‘Person running to the door’ from the train set, reason on that and identifies the correct moment in the video. GT indicates the ground truth timestamps and Prediction indicates predicted temporal endpoints of our approach.

3 Additional Qualitative Results

Figure 4 and 5 illustrates some qualitative results of our proposed approach. Figure 4 shows an example from ActivityNet Captions Unseen dataset. Given the unseen text query ‘An older blonde newswomen is reading a story’ and the corresponding video, our approach retrieves the semantically most relevant query and it’s corresponding moment as the support moment. Then based on reasoning with the support moment, our approach identifies the correct moment in the given video. Figure 5 shows an example from Charades-STA Unseen dataset. It also illustrates that our approach is able to identify correct moments based on relational reasoning for unseen text queries.

4 Efficiency of TLRR

We compare the run-time of our proposed TLRR with conventional temporal localization approaches SCDM and 2D-TAN. It is expected that TLRR would require more inference time due to the extra steps of computation of relevant moments and relational reasoning. We observe from Table 4 that compared to 2D-TAN and SCDM, proposed TLRR takes slightly more time in inference (i.e., $1.76s$ for proposed vs., $1.30s$ for 2D-TAN and $1.23s$ for SCDM).

Table 4: Per batch inference time of TLRR compared to SCDM and 2D-TAN in ActivityNet Captions Unseen dataset.

Method	**Inference Time**
SCDM [5]	1.23 s
2D-TAN [6]	1.30 s
TLRR	1.76 s

5 Performance of TLRR in Original Charades-STA

We conduct experiment on original Charades-STA dataset which is reported in Table 5. Our proposed TLRR is able to show comparable performance on the original temporal localization dataset, even though TLRR is not optimized for seen events (since similar events are available in trainset, all events in testset can be considered as seen events) and have a relatively simple base architecture.

Table 5: This table reports text query based temporal moment localization performance of TLRR on the original Charades-STA dataset.

Method	**R@1, IoU@0.5**	**R@1, IoU@0.7**	**R@5, IoU@0.5**	**R@5, IoU@0.7**
CTRL [1]	23.63	8.89	58.92	29.52
2D-TAN [6]	**39.70**	**23.31**	80.32	**51.26**
TLRR	37.63	21.48	**82.61**	49.27

6 Ablation on Different Reorganization of Dataset

As existing benchmark temporal moment localization dataset splits are not designed for the task of temporal localization of novel events based on unseen text queries, we reorganized the dataset according to our problem setting. In our reorganization of the datasets, excluding queries which contains verb or noun from both seen set and unseen set results in reduced number of moment-sentence pairs. However, the size of the dataset does not have impact on the significance of our proposed problem setup. To demonstrate that we reorganize ActivityNet Captions dataset keeping 400 verbs and 500 nouns in trainset, which results in 10600, 235, and 141 videos in the trainset, unseen testset, and seen testset respectively. Here, the number of videos in the trainset is in the same order as original ActivityNet Captions dataset $(10k)$. Experiment with existing approach (2D-TAN) shows that even with the large and complex training set, there is a significant difference in localization performance for seen events and unseen events which is reported in Table 6. We did not use this reorganization in the original setup as it does not provide a balanced trainset and testset.

Table 6: This table reports 2D-TAN performance difference for seen events and unseen events for reorganized ActivityNet Captions dataset where the trainset consists of $\sim\ 10k$ videos.

Method	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
Seen Event	33.33	21.75	60.00	44.91
Unseen Event	26.90	16.86	54.21	40.96

Moreover, we ensure that the performance degradation in unseen testset is not an overfitting problem due to the reduced training set size. To prove this, we consider the best performing models of existing SOTA approaches trained on original and reorganized ActivityNet Captions dataset. We observe that training accuracy in their respective trainsets are almost similar (as shown in Table 7 for LGI [4] approach). It indicates that the size of trainset is reasonable and doesn't result in overfitting.

Table 7: This table reports the trainset performance of LGI on Original and Reorganized ActivityNet Captions dataset.

Method	R@1, IoU@0.5	R@1, IoU@0.7	mIoU
Original	90.49	83.38	81.60
Reorganized	89.30	80.87	80.40

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
2. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
3. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV) (2017)
4. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10810–10819 (2020)
5. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Advances in Neural Information Processing Systems. pp. 534–544 (2019)
6. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. arXiv preprint arXiv:1912.03590 (2019)