GEB+: A Benchmark for Generic Event Boundary Captioning, Grounding and Retrieval Supplementary Material

Yuxuan Wang[1], Difei Gao[1], Licheng Yu[2], Weixian Lei[1], Matt Feiszli[2], and Mike Zheng Shou[1]

[1] Show Lab, National University of Singapore
[2] Meta AI

Overview

In the supplementary material, we provide more details of annotations (Sec. 1) and more implementation details of the baselines (Sec. 2). Moreover, we conduct more experiments of Boundary Captioning and Grounding for more visual difference representation methods as well as the ablation study for our *TPD Modeling* method (Sec. 3). Finally, we release some common failure cases in our prediction of Boundary Captioning and further discussion on the benchmark (Sec. 4).

1 More Details of Annotations

1.1 Boundary Definition

Specifying the level of details. A great number of our video sources from *Kinetic-400* contain more than one actor or object with different levels of status changes, and different annotators could have high-variance opinions on the boundary positions. According to [6], to reduce the variance among annotators, the highest priority is to specify the level of the spatial and temporal details we take into consideration. For the level of spatial details, we only focus on the event changes that are performed by dominant subjects. Specifically, in the *Example 2* of Fig. 2 in main body, the two girls are repeating the same event, the status of which is unchanged (no event boundary). Instead, the boy performs different events before and after the marked timestamp. For the level of temporal details, we only consider the "one-level-deeper" granularity as in [6]. By specifying this, we ensure that most of the boundaries are in the same granularity, rendering it possible for annotators to basically reach an agreement on the boundary location without predefined classes.

Embracing the Ambiguity. Knowing the specified level of details, however, different annotators could still have some disagreements on the dominant subjects and the "one-step-deeper" granularity events. Following [6], we embrace this varsity when annotating. For each video, we take all the annotations as correct. Then we supervise the consistency among different annotations towards the same video by calculating the F1 score in Sec. 1.2.

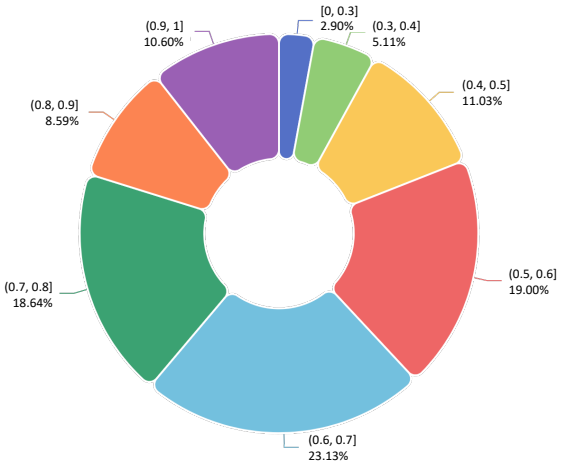

Fig. 1. Distribution of consistency F1 scores in all annotations. We first compute the F1 scores with different thresholds from 0.2s to 1s, and then average the scores in all thresholds as the final score

1.2 Quality Assurance

Criteria for Rejecting a Video. To ensure the quality of videos, we designed a rejection criteria for annotators to filter the video sources. Each video is simultaneously allocated to at least 5 annotators, and each annotator could independently decide whether to annotate or reject the video. Following [6], the criteria is designed based on the understandability and the boundary number of the video. Specifically, a video is expected to be rejected in four cases: *(1)* Not understandable due to blurry or overspeeding. *(2)* Contains no boundary or too many boundaries. *(3)* Includes shot changes like zooming, panning or cutting. *(4)* Violating content. The statistics on the number of annotations in all selected videos is shown in Tab. 1. We could see that a majority of videos are accepted by at least 5 annotators, indicating the consistency of annotators' opinions on our annotated videos.

Table 1. Annotation number per video

#Annotations	1	2	3	4	5
#Videos	605	536	582	928	9783
Per. (%)	4.87	4.31	4.68	7.46	78.68

Table 2. Timestamp v.s. Time Range

Boundary	Timestamp	Time Range
Num.	172103	4578
Per. (%)	97.41	2.59

Evaluation of Annotators' Consistency. Following [6], we compute F1 score to evaluate the consistency of the annotations towards the same videos. When computing, we take the timestamps of each annotation as the "prediction" and all other annotations in the same video as the "ground truth". Then for each threshold varying from 0.2s to 1s, we compute the precision and recall for the "prediction" to obtain its F1 score. Finally, we average the F1 scores under all

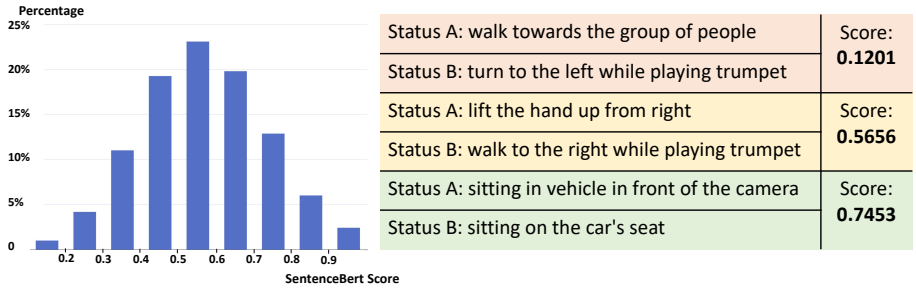

Fig. 2. *Left.* Distribution of Sentence-BERT scores for the sampled captions in the same videos. *Right.* Intuitive examples of different level of Sentence-BERT scores selected from captions in the same videos in Kinetic-GEB+.

thresholds as the final result of the evaluation. The distribution of the average F1 score is shown in Fig. 1, where over 92% percent of annotations are scored higher than 0.4, suggesting that our annotators have very high consistency in determining event boundary positions. They also tend to focus on the same subject on the agreed boundary, providing captions for the same event change with little bias.

1.3 Statistics on the Similarity between Captions

To further investigate the similarity between the captions annotated in the same videos. We first randomly selected 1,000 videos with over 103K captions from our dataset, then computed the Sentence-BERT similarity of status parts' captions in these videos. The results with three examples for different level of scores are shown in Fig. 2. We see that nearly 80% of the caption pairs are less than 0.7 score (only a few words are shared), indicating our captions are unique and fine-grained.

1.4 Statistics based on the Video Categories in Kinetic-400

Since we take the "one-step-deeper" events in videos as [6], the video-level categories in Kinetic-400 could not determine the pattern of events. However, the category provides a higher-level background for our events, thus we conduct further statistics towards it.

Boundary Number and Interval Duration. Firstly we investigate the distribution of boundary numbers in each category of videos. Given a Kinetic-400 category, we compute the average number of boundaries per video in the category. From the result in Fig. 3, we see that the boundary numbers slightly vary with the category and most categories have 2 to 3 boundaries per video. We also illustrate the interval durations versus categories in the right of Fig. 3. In most categories of videos, we could see the average duration of boundary intervals is around 2s.

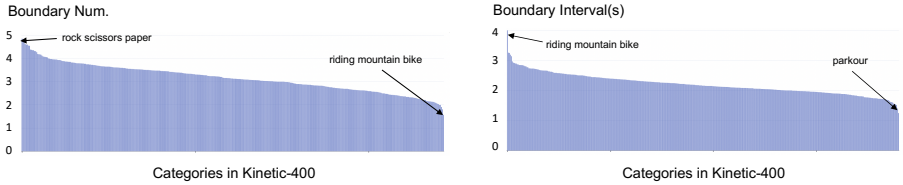

Fig. 3. *Left.* Average number of boundaries in videos in each *Kinetic-400* category. *Right.* Average duration of boundary intervals in videos in each *Kinetic-400* category.

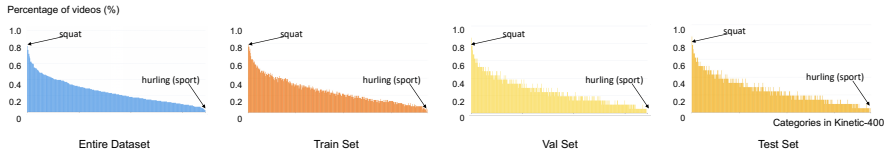

Fig. 4. Percentage distributions of the videos in each *Kinetic-400* category in the entire dataset and the train/val/test splits. The categories are sorted by their video numbers in the entire dataset

Distributions in Splits. Furthermore, we conduct statistics on the video numbers of each category in our train/val/test splits. The percentage distribution is shown in Fig. 4, where the categories are sorted by their video numbers in the entire dataset. We see that the categories' distribution in the three splits are consistent with the distribution in the entire dataset.

1.5 Details of the Adjustment for Downstream Tasks

In the raw annotation of Kinetic-GEB+, each video is allocated to more than 5 annotators. Due to the variance of annotators' opinions, the boundary locations in different annotations towards the same video are not the same. When preparing the data for downstream tasks, we select one annotator whose labeled boundaries have highest F1 score (computed in Sec. 1.2) for each video. Then, we use these boundaries' timestamps as the anchors to merge other annotators' captions, preserving the diversity of different opinions. Thus, one video corresponds to multiple boundaries, and each boundary could be with multiple captions. Finally, we collected 40k anchors/boundaries and from the total 176,681 boundaries in 12,434 videos, where 80% anchors/boundaries have more than 3 captions and around 10% of anchors have only 1 unique caption.

As mentioned in main body, the videos in our Kinetic-GEB+ could sometimes contain repeated events or actions, which could disturb the Boundary Grounding task. We found that the difference among some pairs of boundaries within a video is too subtle even for humans to distinguish. Therefore, we need to find these "equal" pairs of boundaries and mark them as "equal" boundaries to each other for the Boundary Grounding task. Specifically, when querying with the caption

of one boundary in the pair, the timestamps of other "equal" boundaries are also correct answers. When queried by a boundary caption, the machine is supposed to answer the locations of that boundary as well as all its "equal" boundaries. An example is shown in Fig. 7, where the man changes his status from sitting to standing twice in the video, thus these two status changes are marked as an "equal" pair.

To find and mark these "equal" pairs, we employ Sentence-BERT [5] to compute the similarity score between the annotated captions of every two different boundaries inside a video. Firstly, we take the filtered annotations. Then for each video, we combine every two of its boundaries to form all the possible pairs. After that, we separate each pair of captions into *subject*, *status before* and *status after* items, and then compute the similarity score for each item using Sentence-BERT. The range of similarity scores is from 0 to 1. In order to distinguish these "equal" pairs, we need to set a maximum threshold for similarity scores. First we find that the item pairs scoring less than 0.9 usually have significant differences that are easy for humans to recognize. Hence, we collect the pairs of all the three items that score higher than 0.9, and then we annotate manually to classify if each pair is an "equal" pair. After that, we simulate the decision accuracy of different candidate thresholds varying from 0.9 to 1.0 and finally choose 0.93 as the threshold, where the corresponding accuracy is 95.5% (i.e. the 2-sigma probability in normal distribution). Finally, we found and marked 4,426 "equal" pairs consisting of 4,295 boundaries.

1.6 More Examples of Kinetic-GEB+

Here we illustrate more examples from our Kinetic-GEB+ in Fig.5. The *Example 1* to *4* are all based on Change of Action. The *Example 5* is based on Change of Subject, since the man was at first appearing in the scene and then disappears after the boundary. In *Example 6*, the color of the stage suddenly changes from blue to pink, causing the boundary based on Change of Color. In *Example 7*, the woman was first interacting with the trophy and then retreats her hands to stop interacting after the boundary. This boundary is thus due to the Change of Object being interacted with. Finally in *Example 8*, the boundary is based on Multiple types of status changes. The man in the scene changes his action and simultaneously stops interacting with the iron ball at the boundary.

2 More Details of Implementation

2.1 Schemes for frame sampling

In all our experiment groups, if not specified, we employ the two following schemes of frame sampling when extracting visual information for boundary timestamps:

Scheme 1. In most cases, when using the ground truth boundaries, we set two sampling ranges before and after each boundary timestamp. For the range

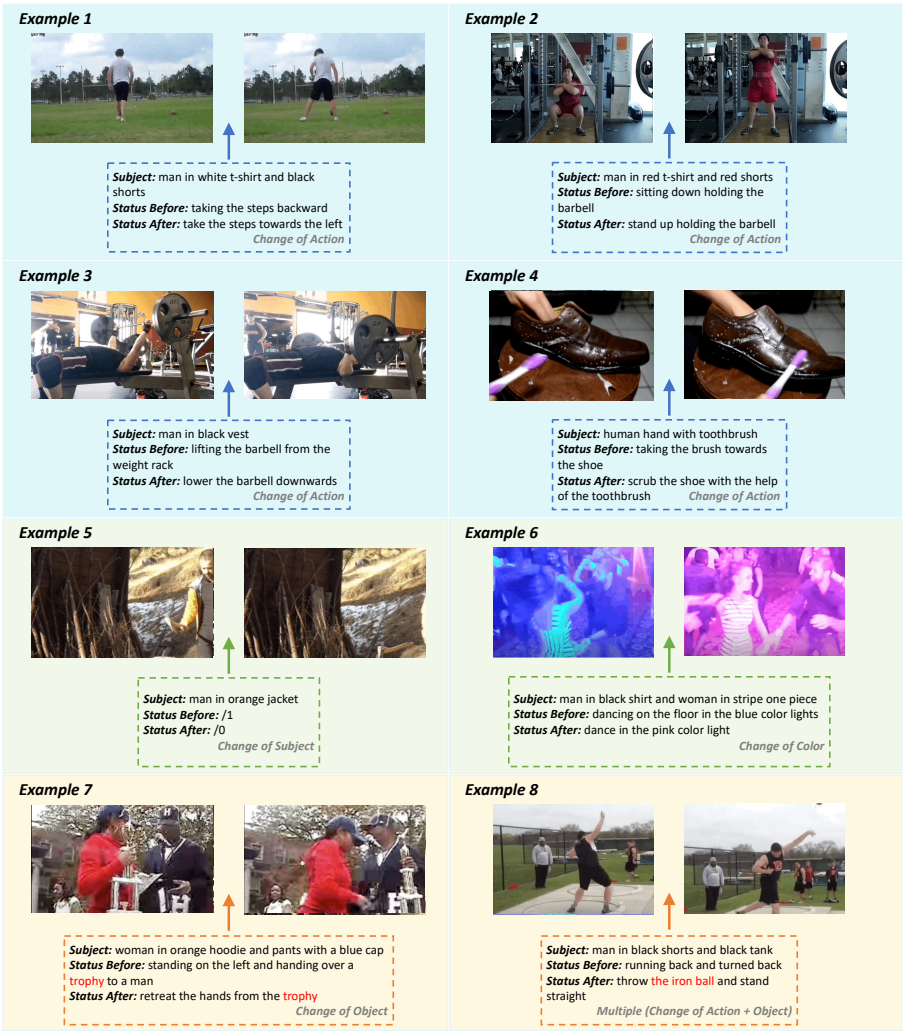

Fig. 5. More samples from *Kinetic-GEB+* dataset

before the boundary, we set the preceding boundary as the start and the current boundary as the end. Similarly, the range after is between the current boundary and the succeeding boundary. Notably, the predecessor of the first boundary in videos is set to 0, and the successor of the last boundary in videos is set to the end of videos. Finally, we sample 10 frames in each range and 1 frame at the timestamp of the current boundary. This scheme is also employed when using the proposal timestamps generated by GEBD baseline [6], like in the testing period of Boundary Grounding specified with "GEBD" suffix.

Scheme 2. Sometimes there is no predefined boundary or proposal, and thus the locations of the preceding and succeeding timestamps are unknown. Therefore, we replace the predecessor and successor with the timestamps 1s before and after the current timestamp. Then we sample 10 frames in each range and 1 frame at the candidate timestamp for further extraction. For example, in the testing period of Boundary Grounding (in the groups without "GEBD" suffix), this scheme is employed by sampling a timestamp candidate every 0.1s for all videos.

2.2 Further Details in Training

For each backbone utilized in our experiments, we trained for 50 epochs. For all the BERT based models, we used AdamW optimizer with a linearly decreasing learning rate starting from $5e^{-5}$. Notably, in Boundary Grounding we modify the original contrastive loss in FROZEN [1] by adding an additional intra loss. Given a batch of embeddings, the intra loss is computed in the same way yet only among the caption and context embeddings from the same videos. Besides, as mentioned in previous sections, we design a batch-random sequential sampler for Boundary Grounding. It ensures more boundaries in the same video to be collected in the same batch, since the boundaries are sequentially sorted by their videos in the dataset. This intra loss and new sampler encourage the model to learn the differences among the boundaries in the same videos, which conforms to the goal of Video Grounding that is selecting the best match among all timestamps in a video.

2.3 Post-processing and Evaluation

In Boundary Captioning, we separate and evaluate the *Subject*, *Status Before* and *Status After* items of the generated captions. We found that the conventional BLEU [4] metric is not suitable for our task and its scores are often inconsistent with humans' impression, since it only considers the simple repetition of word grams. Samples of predicted captions in a video are illustrated in Fig. 6. We see that the first two generated captions are relatively great, while the caption generated from the last boundary is not satisfying. For Boundary Grounding, we conduct a post-processing after the models generating the matching scores of all candidates. First we apply the LoG filter [3] to find the local maximas following [6]. Then we select the top-K maximas as final prediction following the statistics of the ground truth timestamp numbers for all queries. After that,

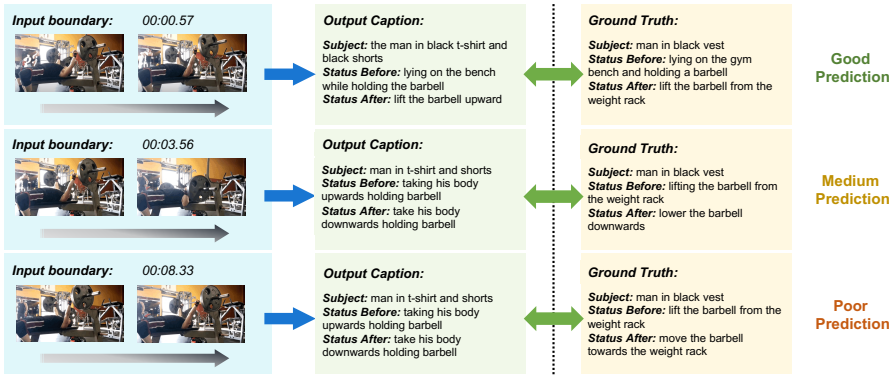

Fig. 6. Samples of Prediction in Boundary Capitioning

we evaluate the finalized prediction by calculating the F1 score under different thresholds, where the computation is the same as in Sec. 1.2. Samples of predictions are shown in Fig. 7. Notably, the boundaries at 00:00.93 and 00:06.11 are a pair of equal boundaries, thus we mark both of their timestamps as the ground truths for their caption queries. For Boundary Caption-Video Retrieval, several samples of predicted ranking are illustrated in Fig. 8. For the first three samples in the figure, the prediction result is relatively satisfying and the ground truth video is within the top-5 of the ranking. However, given the caption of the last sample in the figure, the machine could not clearly recognize the target video from the corpus, and the ground truth video is ranked to #42.

3 More Exploration on Experiments

3.1 Boundary Captioning

Here we delve deeper on the design of fusion mechanism for status changes. In Tab. 3, we further compare subtraction operation (TPD method) with another 2 operations: Simple concatenation (denoted as *Concat*) and *Multimodal Tucker Fusion* [2]. As shown, our TPD outperforms the other fusion methods. Furthermore, to investigate the contribution made by different parts in our **TPD Modeling** method, we conduct an ablation study. In Tab. 3, we see that *part a* contributes more than *part b* and *part c*, while the combination of the three parts enables the model to have the best performance.

3.2 Boundary Grounding

In order to investigate the contribution of different captioning parts in Boundary Grounding task, we evaluate some variants of *FROZEN-revised-GEBD* which

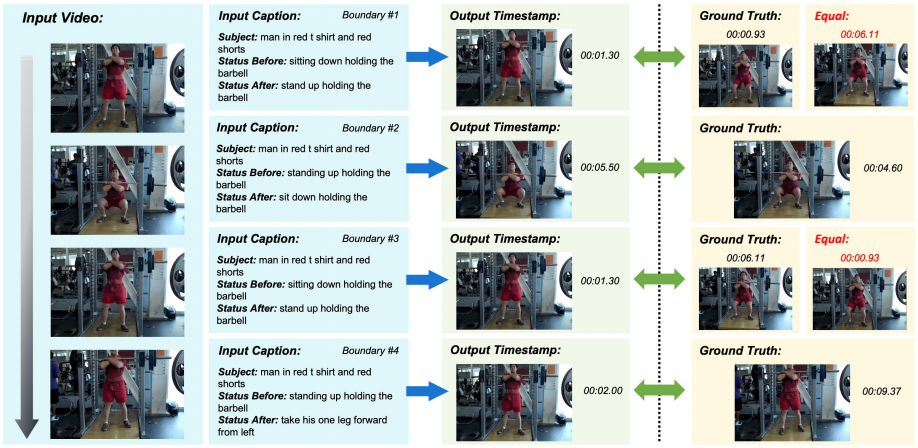

Fig. 7. Samples of Prediction in Boundary Grounding

Table 3. Results of more exploration on Boundary Captioning, including the comparison among different fusion methods and the ablation study on our **TPD Modeling** method

Method	CIDEr				SPICE				ROUGE_L			
	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.
Concat	68.16	86.35	62.99	55.15	18.92	20.13	19.08	17.57	26.71	39.25	20.96	19.93
Tucker	67.25	85.15	63.08	53.51	18.71	20.39	18.97	16.78	26.91	39.28	21.42	20.02
TPD (part a)	72.45	89.7	70.45	57.20	19.39	**20.64**	19.87	17.67	27.74	39.49	22.86	20.87
TPD (part b)	70.78	**90.04**	66.59	55.70	19.22	20.46	19.67	17.54	27.27	39.6	21.93	20.27
TPD (part c)	69.01	86.9	66.11	54.03	19.11	20.34	19.47	17.53	27.31	**39.69**	21.98	20.27
TPD	**74.71**	85.33	**75.98**	**62.82**	**19.52**	20.10	**20.66**	**17.81**	**28.15**	39.16	**23.70**	**21.60**

take no caption (i.e. random guess based on boundary proposals from GEBD) or only subject parts as input for grounding. The F1 scores of *no caption/only subject/full caption* under 0.1s are *3.09/3.25/4.20* respectively. The performance doesn't improve much with only the subject. The reason is that boundaries in the same video are often caused by the same subject, requiring the model to understand captions depicting detailed status changes to ground the video.

4 More Discussions

4.1 Common Failures in Boundary Captioning

In the task of Boundary Captioning, we find some failure cases happened in our prediction. Here we present two types in Fig. 9: (1) the model misses the target subject due to another subject without event change is visually salient. (2) the

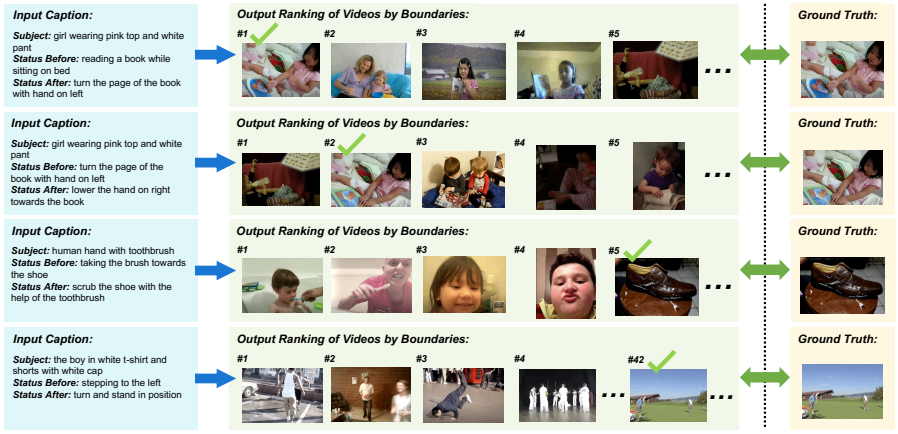

Fig. 8. Samples of Prediction in Boundary Caption-Video Retrieval

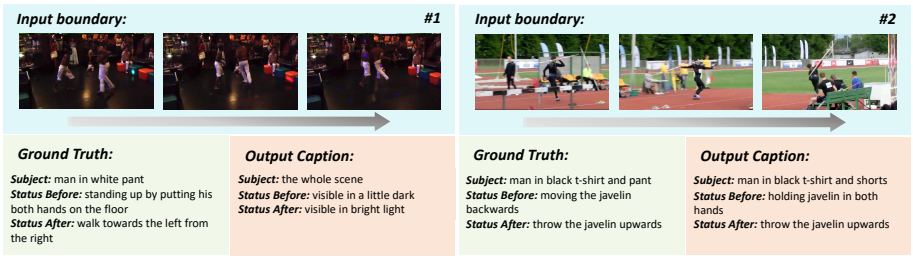

Fig. 9. Two Common Failure Cases in the Task of Boundary Captioning

action in status before or after is subtle and the model mistakenly considers there is nothing happening.

4.2 "Can we replace Kinetic-GEB+ with existing video captioning datasets simply by concatenating two segments together?"

It may not work well as (1) previous and next captions from existing datasets could correspond to different subjects while our boundary caption targets one subject; (2) Event caption usually summarises the whole time span, while boundary caption focuses on detailed, fine-grained status change of the subject.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV. pp. 1728–1738 (2021)
2. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: ICCV. pp. 2612–2620 (2017)
3. Lindeberg, T.: Feature detection with automatic scale selection. IJCV **30**(2), 79–116 (1998)
4. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (Jul 2002). https://doi.org/10.3115/1073083.1073135
5. Reimers, N., Gurevych, I., Thakur, N., Daxenberger, J., et al.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP. pp. 671–688 (2019)
6. Shou, M.Z., Lei, S.W., Wang, W., Ghadiyaram, D., Feiszli, M.: Generic event boundary detection: A benchmark for event segmentation. In: ICCV. pp. 8075–8084 (2021)