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

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

Abstract. Cognitive science has shown that humans perceive videos in terms of events separated by the state changes of dominant subjects. State changes trigger new events and are one of the most useful among the large amount of redundant information perceived. However, previous research focuses on the overall understanding of segments without evaluating the fine-grained status changes inside. In this paper, we introduce a new dataset called **Kinetic-GEB+**. The dataset consists of over 170k boundaries associated with captions describing status changes in the generic events in 12K videos. Upon this new dataset, we propose three tasks supporting the development of a more fine-grained, robust, and human-like understanding of videos through status changes. We evaluate many representative baselines in our dataset, where we also design a new **TPD (Temporal-based Pairwise Difference) Modeling** method for visual difference and achieve significant performance improvements. Besides, the results show there are still formidable challenges for current methods in the utilization of different granularities, representation of visual difference, and the accurate localization of status changes. Further analysis shows that our dataset can drive developing more powerful methods to understand status changes and thus improve video level comprehension. The dataset is available at https://github.com/Yuxuan-W/GEB-Plus

Keywords: Video Captioning, Generic Event Understanding, Status Changes, Difference Modelling

1 Introduction

According to cognitive science [24], humans perceive videos in terms of different events, which are separated by the status changes of dominant subjects in the video. For example, in Fig. 1, humans perceive the process of "javelin sport" by the action events such as "walking", "running" and "throwing". These events are triggered by the athlete's status changes, like the instantaneous change from "walking" to "running". The moment that instantly triggers status changes of persons, objects, or scenes often conveys useful and interesting information
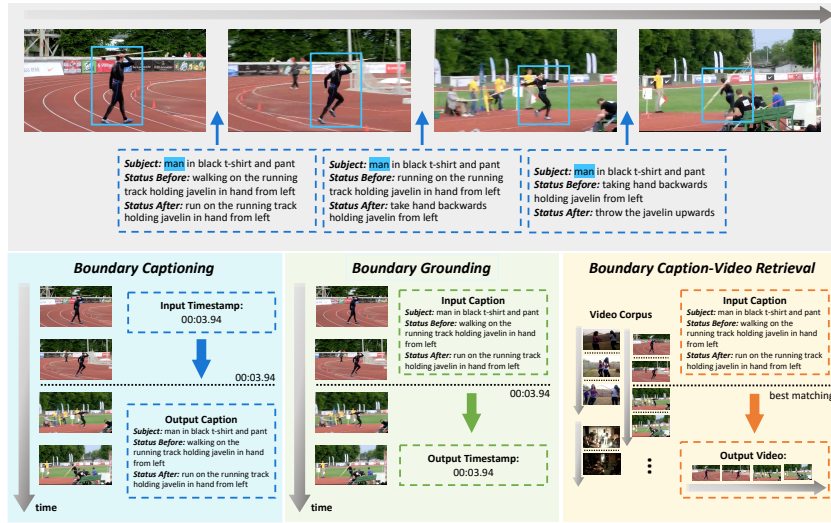
Fig. 1. An example of generic event boundaries with captions in Kinetic-GEB+, as well as three downstream tasks designed upon the boundaries.

among a large amount of repeated, static, or regular events. Therefore, developing the understanding of the salient, instantaneous status changes is another step towards a more fine-grained and robust video understanding. Previous works, like Dense Video Captioning [15,38,32,17,12] and Video Grounding [25,9,5,10,21,35,36] attempt to develop the understanding of events in video or video segments. However, these works only focus on developing an overall understanding of events rather than delving into the fine-grained status changes in the video. Other researches focusing on image level changes [22,13] employ the visual difference modeling to capture the status changes in image pairs. However, since the image contains only static information, the state changes exhibited by the two images involve only a few simple patterns, e.g., appear, move. These tasks are hard to evaluate the ability on understanding generic status changes.

More recently, *Shou et al.* [27] proposes Kinetic-GEBD dataset with annotated boundary timestamps for detection in Kinetic-400 videos [6], where a boundary is defined as the splitter between two status of the subject. Though the videos in Kinetic-400 [6] are categorized, the events selected inside are generic and mostly independent from the whole video's category. However, in addition to letting the model predict where is the boundary, it is more important to understand why this is the boundary, which associates the visual information of boundaries with natural human languages.

Motivated by this idea, we build a new dataset called **Kinetic-GEB+ (Generic Event Boundary Captioning, Grounding and Retrieval)** which includes the video boundaries indicating status changes happening in generic events. For every boundary, our Kinetic-GEB+ provides the temporal location

Table 1. Comparison with most relevant Video Captioning datasets. Our *Kinetic-GEB+* has comparable scale and is the only one targeting the generic boundaries, while conventional datasets focus on entire videos or video segments

	#Videos	Video Domain	#Captions	Caption Target	Target Type	Annotation in Segments
MSR-VTT	7,180	20 categories	200K	video	generic event	caption
VATEX	41,250	in-the-wild	825K	video	action	caption
Charades	67,000	household	20K	segment	action	time range + caption
MSVD	2,089	in-the-wild	85K	segment	generic event	time range + caption
YouCook2	2,000	kitchen	15K	segment	action	time range + caption
ActivityNet Captions	20,000	in-the-wild	100K	segment	action	time range + caption
Kinetics-GEB+	12,434	in-the-wild	177K	boundary	generic event	timestamp/range + caption

and a natural language description, which consists of the dominant *Subject*, *Status Before* and *Status After* the boundary. In total, our dataset includes 176,681 boundaries in 12,434 videos selected from all categories in Kinetic-400 [6]. The detailed definition of our boundary is described in Sec. 3.1. For future applications like AI assistant robots, with the comprehension developed from the visual status changes and natural language captions, they could understand the real time, instantaneous occurrences without hints to assist the users.

In order to comprehensively evaluate the machine's understanding of our boundaries, we further propose three downstream tasks shown in Fig. 1: *(1) Boundary Captioning.* Provided with the timestamp of a boundary, the machine is required to generate sentences describing the status change at the boundary. *(2) Boundary Grounding.* Provided with a description of a boundary, the machine is required to locate that boundary in the video. *(3) Boundary Caption-Video Retrieval.* Provided with the description of a boundary, the machine is required to retrieve the video containing that boundary from video corpus.

In the experiment, we compare several state-of-the-art methods [16,20,39,37,3] along with many variants on our datasets to analyze the limitation of current methods and show the challenges of the proposed tasks. Due to the need of visual difference for understanding the status changes, we further propose a **Temporal-based Pairwise Difference (TPD) Modeling** method representing a fine-grained visual difference before and after the boundary. This method brings a significant performance improvement. On the other hand, the results show that there are still formidable challenges for current SOTA methods in developing the comprehension of status changes.

2   Related Work

Video Captioning is a conventional task with many benchmarks [34,7,33,15,38] established which aim to caption trimmed videos with natural language descriptions. More recently, several works [15,38,32,17,12], e.g., Dense Video Captioning [15], focus on captioning the self-proposed event segments in videos. All tasks above are evaluating the overall understanding of an event, whether the event is presented in the form of a trimmed video or a video segment. In contrast,

our Boundary Captioning task is to develop the comprehension of instantaneous status changes happening at boundaries, i.e., describing the important moment that caused a dramatic change in the state of persons, objects or scenes. As a result, there is a more urgent need for models to understand the changes in various granularity of visual concepts, e.g., action, attributes, scene status, etc. In Tab. 1, we compare the most relevant video captioning datasets with ours.

Image Change Captioning is a task evaluating the ability on capturing and describing the difference between two images. There are many existing benchmarks targeting at this task. Early works [28,19] focus on changes in aerial imagery for monitoring disaster. Some other datasets [1,13] are about captioning the changes in street scenes, e.g., Spot-the-diff [13]. Recently, [22] proposes a more challenging change caption dataset, CLEVR-change, which utilizes the CLEVR engine to construct complicated synthetic scenes to evaluate models on finding more subtle change. One crucial limitation of previous works is that images can only present static information, thus status changes presented by two images can only involve a limited number of patterns, e.g., "appear", "disappear", "add" and "move". Towards a generic understanding of change, we extend the setting from images to videos which supports a open set of change pattern, including human action change, scene state change, etc.

Video Retrieval and Grounding are both language-to-vision tasks. Given a text description of a video or event, Video Retrieval requires models to select the target video from the corpus [3,8,23], and Video Grounding requires models to locate the target event segment (i.e. start and end boundaries) from an untrimmed video [5,10,21,35,36]. These tasks are based on the event level understanding to find the best matching video or time span. Compared with previous works, our Boundary Caption-Video Retrieval and Boundary Grounding tasks requires locating the two states of the subject, while traditional grounding only localizes one event. Besides, our captions are more fine-grained (describing detailed status changes) than those in traditional tasks (describing a general event).

Generic Understanding is a popular topic aiming to drive models from understanding predefined classes to open world vocabulary. Many pioneer works [4] propose open-set recognition tasks, which extend image classification tasks to generic understanding versions. Some works [15,38] introduce datasets for the generic event understanding requiring models to describe videos with natural language. More recently, a new dataset called Kinetic-GEBD (Generic Event Boundary Detection) [27] is proposed, which focuses on detecting the status changes between generic events. Our work is an extension to Kinetic-GEBD. We also study the boundary between events. However, we believe a sophisticated model should not only know where is the boundary but also understand why it is a boundary. Thus, this paper constructs a dataset with a large scale of boundary captions and introduces new boundary language-related tasks.

Example 1

Subject: man in black t-shirt and pant
Status Before: walking on the running track holding javelin in hand from left
Status After: run on the running track holding javelin in hand from left
Change of Action

Subject: man in black t-shirt and pant
Status Before: running on the running track holding javelin in hand from left
Status After: take hand backwards holding javelin from left
Change of Action

Subject: man in black t-shirt and pant
Status Before: taking hand backwards holding javelin from left
Status After: throw the javelin upwards
Change of Action

Example 2

Subject: boy in blue t-shirt and blue jeans pant
Status Before: /0 (*not in the scene)
Status After: /1 (*in the scene)
Change of Subject

Example 3

Subject: whole screen
Status Before: brightening light in center of the scene
Status After: switch off the light
Change of Color

Fig. 2. Three samples from Kinetic-GEB+. Each boundary consists of a temporal position and an associated caption, with the boundary type noted at the bottom

3   Benchmark Construction: Kinetic-GEB+

To build the Kinetic-GEB+ dataset, we select 12,434 videos from the Kinetic-400 dataset [14] and annotate 176,681 boundaries following a designed guideline and format. In total, our selected videos cover all the 400 categories of the Kinetic-400. It is then split into 70% train, 15% val and 15% test non-overlapping sets. Several samples of boundaries are shown in Fig. 2.

3.1   Boundary Collection

When annotating Kinetic-GEB+, one simple way would be directly captioning the boundaries in Kinetic-GEBD [27]. However, our annotators did their jobs quicker when being asked to re-annotate the boundary positions than to interpret GEBD's boundaries. Yet, the boundaries from GEBD and GEB+ are highly consistent: when following Supp., nearly 90%/70% boundary positions in GEB+ reaching f1 scores higher than 0.5/0.7 with the boundaries in GEBD.

Format and Guideline. Following GEBD [27], a boundary is defined as the splitter between two status of the subject in the video. Generally, we categorized our boundaries into five types: Change of Action, Change of Subject, Change of Object, Change of Color and Multiple Changes. When annotating, we accept both single timestamps and time ranges as in [27], and each video is allocated to at least five annotators. Each annotator could independently decide whether to accept or reject the video following the criteria. The statistical results of annotation numbers and formats is shown in Tab. 2 and Tab. 3. Following [27],

Sentence: The man in black t-shirt and pant used to walk on the running track holding the javelin in hand from left, then at the boundary, he starts to running with the javelin in his hand.	Valid Percentage (%)	93.87
	Avg. Annotation Time (s)	74.3
	Avg. Evaluation Time (s)	20.3
Subject: man in black t-shirt and pant Change: first walked on the running track holding the javelin in hand from left, then at the boundary, starts to running with the javelin in his hand.	Valid Percentage (%)	97.60
	Avg. Annotation Time (s)	58.0
	Avg. Evaluation Time (s)	17.9
Subject: man in black t-shirt and pant Status Before: walking on the running track holding javelin in hand from left Status After: run on the running track holding javelin in hand from left	Valid Percentage (%)	99.53
	Avg. Annotation Time (s)	50.2
	Avg. Evaluation Time (s)	12.5

Fig. 3. Three candidate formats of Boundary Captions and their evaluation results, respectively *One-Sentence format*, *Two-Item format* and *Our Finalized format*

we set a minimum threshold for both temporal and spatial details' level to ensure the consistency among different annotators. Further details are shown in Supp.

Table 2. 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 3. Timestamp v.s. Time Range

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

3.2 Caption Collection

In our Kinetic-GEB+, annotators are supposed to add a language description to each boundary they annotated in Sec. 3.1. To clearly and comprehensively represent humans' understanding of the status changes, we randomly sampled 300 videos for pilot annotation to design the formats and guidelines of captioning.

Format. Our final format of caption consists of three compulsory items: *(1)* Dominant *Subject* that performs the status changes. *(2)* Subject's *Status Before* the boundary. *(3)* Subject's *Status After* the boundary. In the pilot stage, we compare different versions of annotation formats as shown in Fig. 3:

One-Sentence format: Annotators use a single sentence to describe the status change happening at each boundary. In order to obtain an open-vocab description close to daily language, we do not restrict or request anything to the expression and annotators have full autonomy in narrating. Though this format enables fluent and natural descriptions, there are significant problems in the annotations: *(1) Ambiguity of subject*: Annotators tend to describe the subject shortly without further restriction, causing ambiguity, e.g., in a scene full of people, a short description like "a man" might indicate multiple persons. *(2) Dual changes*: Without restriction, annotators could wrongly combine two state changes of different subjects together in caption, like "Musician stops playing and an auditor starts clapping". *(3) Low efficiency*: Long sentences costs annotators more time to construct and takes our raters more time to understand.

Two-Items format: To address the problems in the one-sentence format, we separate the sentence into a *Subject* item and its *Change* item as shown in Fig. 3.
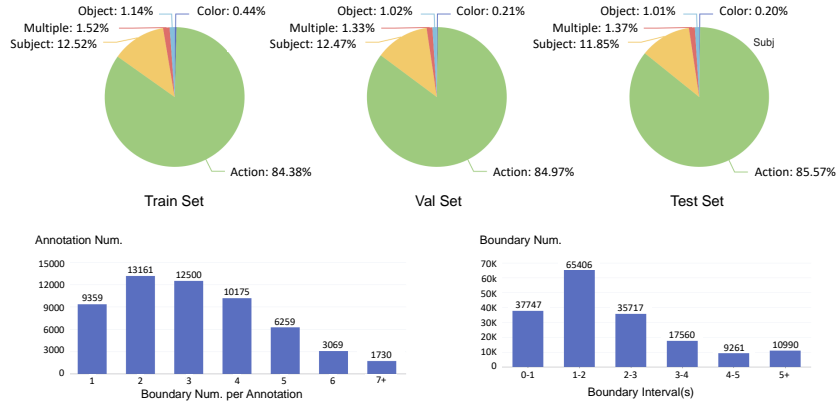
Fig. 4. *Top.* Distribution of boundary types in Train/Val/Test split. *Bottom Left.* Annotation numbers versus the numbers of boundary in the annotation. *Bottom Right.* Boundary numbers versus the duration of the interval before the boundary

For *Subject*, annotators should fill in a noun phrase. We notice that this separation makes it easier for annotators to check the singularity and specification of *Subject*. Although we see that the efficiency of both annotation and evaluation are improved, this scheme still have some shortcomings: *(1) Incomplete status*: Annotators sometimes forgot to describe the status before the boundary. For example, when describing an athlete's changing from walking to running, an annotator only filled "starts to run on the track" in *Change* and forgot to mention the "walking" status before the boundary. *(2) Low efficiency*: Even though this separation improves the efficiency, the *Change* item could still be too long for auditors to evaluate. Therefore, we further separate *Change* into *Status Before* and *Status After* to ensure the completeness of the status change description. Finally, we found this fully separated format the most efficient and robust for annotation, as shown in Fig. 3.

Guideline. In our Kinetic-GEB+, the caption is defined as the reason why the annotator separates the preceding and succeeding segment of the boundary. Following the format of annotation, we brought up some specific guidance for annotating the items. Specifically, when annotating the *Subject* item, annotators are required to provide distinguishable attributes of the dominant subject. However, in complex cases where the subject is difficult to describe without ambiguity (e.g. many people dressing similarly in the scene), the annotator could just describe some attributes to avoid verbose descriptions.

When annotating *Status Before* and *Status After*, annotators are required to limit their attention to the time range between the proceeding boundary and succeeding boundary, thus to ensure all the status changes in the same video are at the same temporal level. To further improve the consistency of expressions, we employ the symbol */1* and */0* to represent the appearance and disappearance
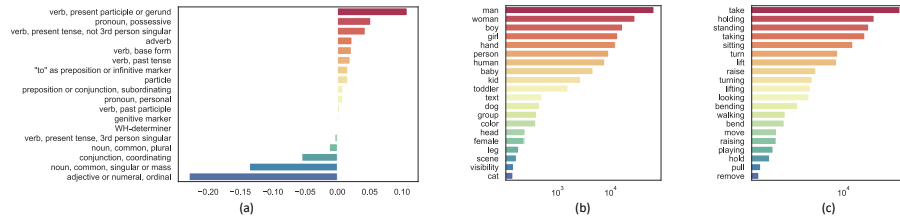
Fig. 5. *(a)* The parts of speech distribution of the *Status Before* and *Status After* compared with that of the Subject part. The two status parts contains more verbs and focus more on motions. *(b)* The 20 most frequent nouns in *Subject*. *(c)* The 20 most frequent verbs in *Status Before* and *Status After*

of a subject in the scene, as shown in *Example 2* of Fig. 2. Finally, we embrace all the tenses only if the annotators feel natural. In this way, we ensure the specification of descriptions while keeping their naturalness.

3.3   Statistics

Splitting. When splitting our Kinetic-GEB+ into train, validation and test sets, the boundary type is the most important characteristic of consistency, since it determines which granularities the model should depend on to understand the state change. We allocate videos containing different types of boundaries by proportion to ensure the distribution is the same in all splits. The final distribution is shown in Fig. 4, where we see the distribution is consistent in three splits. More details of splitting criteria is discussed in Supp.

Boundary number. To quantify the density of annotated boundaries, we make a statistics of the boundary number in each piece of annotation. Notably, due to the variant understanding of annotator, annotations of the same video could have different numbers of boundaries. The bottom left side of Fig. 4 shows the counts of annotations versus their boundary numbers, from which we could see that most of annotations have 1 to 4 boundaries.

Boundary interval. Furthermore, to investigate the duration of events located between two boundaries, we conduct the statistics on the length of intervals. For the first boundary in the video, we take the distance to the start of the video as its interval duration. The result is shown in the bottom right side of Fig. 4 which is similar to the statistic of boundary numbers.

Part of speech comparison in caption. For the captions in our dataset, we first analyze and compare the part of speech distributions in the subject and two status parts. In Fig. 5(a), the comparison result indicates that the status parts contain more verbs and focus more on actions than the subject part. On the other hand, the subject part includes more nouns and adjectives than the two status parts, suggesting it focuses more on appearance information.

Frequent subjects and actions in caption. To further analyze the different aspects of information in the three parts. In Fig. 5(b)(c), we extract the first
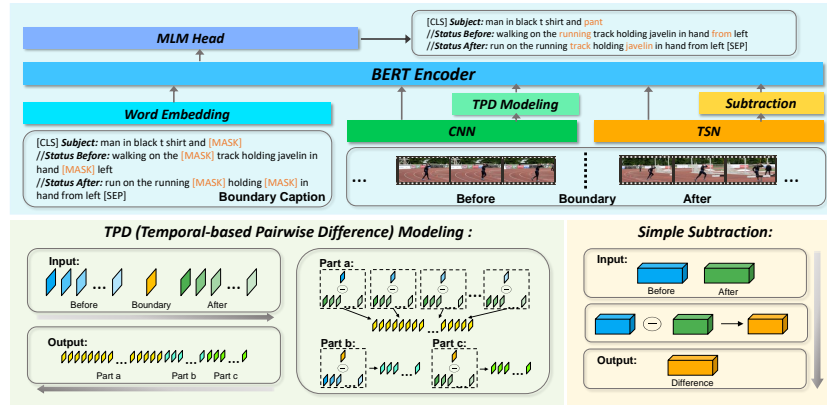
Fig. 6. *Top.* A general modification for BERT model showing on **ActBERT-revised**. *Bottom.* Our difference modeling methods designed for BERT model

noun in every *Subject* as well as the first verb in all *Status Before* and *Status After*, and then illustrate the 20 most frequent words. Same with Kinetic-400, we see that both the nouns and verbs in our datasets are mainly correlated with the appearance and motions of humans. This conforms to the scenarios of practical application, since humans are also the dominant subject in most of the scenes.

3.4 Adjustment for downstream tasks

For downstream tasks, we select one annotator whose labeled boundaries are mostly consistent with others to reduce noise and duplication. 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, this selection includes 40k anchors from all videos. Furthermore, we find two different boundaries in the same video could be occasionally too similar in semantics for even humans to tell. For *Boundary Grounding*, we mark these pair of boundaries as equal in the ground truth. More details are discussed in Supp.

4 Experiments

Kinetic-GEB+ dataset enables us to benchmark how well current mainstream methods could comprehend the instantaneous status changes in videos. For each task, we implement and compare among SOTA models with our modifications, as well as further explorations on ablation and visual difference modeling methods.

4.1 Methods

Granularities of Input Features. We extract multiple granularities of features and utilize different combinations of them in experiments. Given each

boundary, we sampled multiple frames before and after the timestamp and one frame at the timestamp for further extraction.

Our features include: *(1) ResNet*: Firstly, we extract a 1024 dimensional ResNet-roi feature using ResNet [11] followed by Region of Interest (RoI) pooling. Then we extract another ResNet-conv feature to fit [22]: We sample one frame before and another frame after the boundary, then extract the Conv features from the two frames. *(2) TSN*: For frames before and after the timestamp, we extract a 2048 dimensional TSN feature for the before and after snippets using pre-trained TSN [31] network. *(3) Faster R-CNN*: For every sampled frame, we employ Faster R-CNN [26] to extract the 1024 dimensional R-CNN feature by selecting 20 objects with highest confidence. *(4) C3D*: Similar to the TSN feature, we extract 4096 dimensional C3D features with pre-trained C3D [29] network for the before and after snippets to fit [37].

These features are categorized into two granularities: *Instant-granularity* features extracted from the instantaneous appearance in a single frame, such as the R-CNN and ResNet features, are to provide fine-grained visual information of instants. *Event-granularity* features, like the TSN and C3D feature, could provide an overall representation of appearance and motion information in event snippets. We assume that developing a fine-grained understanding of status changes requires both the granularities.

Backbones. We implement the following backbones with various adoption and modification according to the tasks: *(1) CNN+LSTM*: A rudimentary backbone that simply uses a vanilla LSTM which takes the CNNs extracted features as input. The output of LSTM is mapped to caption tokens in Boundary Captioning, or is max-pooled to be the matching score in other two tasks. *(2) Dual Dynamic Attention Model (DUDA)*: The baseline method in [22] which consists of a CNN-based Change Detector and a LSTM-based Dynamic Speaker. Besides, it utilizes a simple visual difference modeling by subtraction. *(3) ActionBERT-revised*: A one-stream BERT architecture using early fusion from [39]. We modify the structure by applying difference modeling after the embedding and employing different feature combinations. *(4) UniVL-revised*: A two-stream BERT architecture from [20], which includes a caption encoder, a context encoder and a cross-encoder for late fusion. We apply difference modeling to the context encoder with different feature combinations. *(5) FROZEN-revised*: A two-stream BERT architecture from [3], which includes a caption encoder and a context encoder with no fusion. The revision is the same as *UniVL-revised. (6) TVQA*: The baseline method in [16], where we remove all the "answer" substreams and process each visual granularity with one stream. *(7) 2D-TAN*: The baseline method in [37], where we only keep the diagonal elements in the 2D map.

Visual Difference Modeling. Developing a fine-grained understanding of status changes at the boundary requires visual difference information. Most existing methods are focused on image-pair differences [22], where the difference is obtained by simply subtracting the "before" image from the "after" image. A simple inference of this method on video tasks is by pooling the sampled frames then doing subtraction. However, this method only provides an event-granularity

Table 4. Performance of Different Methods in Boundary Captioning. For *UniVL-revised* and *ActBERT-revised*, we apply the *TPD Modeling* and take the "ResNet-roi+TSN" combination as input feature

Method	CIDEr				SPICE				ROUGE_L			
	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.
CNN+LSTM	49.73	80.11	34.39	34.69	13.62	18.84	9.92	12.10	26.46	39.77	20.77	18.83
DUDA	58.56	**104.41**	47.12	24.14	16.34	**21.72**	14.63	12.68	27.57	**42.76**	21.76	18.18
UniVL-revised (two-stream)	65.74	91.51	56.58	49.13	18.06	21.08	17.06	16.05	26.12	40.67	19.42	18.28
ActBERT-revised (one-stream)	**74.71**	85.33	**75.98**	**62.82**	**19.52**	20.10	**20.66**	**17.81**	**28.15**	39.16	**23.70**	**21.60**

representation of the visual difference between the before and after snippets, and thus loses the instant-granularity visual differences.

To address this problem, we design a new method of **Temporal-based Pairwise Difference (TPD) Modeling** for BERT models. As shown in Fig. 6, we first compute the pairwise subtraction between the embedding of frames in "before" and "after" as *Part a*, where the embeddings of the frames are sampled in Sec. 4.1. This provides us a fine-grained and wide-viewing visual comparison between the status before and after. To represent the visual difference between the boundary and other sampled timestamps, we further compute *Part b* and *Part c*, which includes the pairwise subtraction between the frame embeddings at the boundary and that before or after the boundary. Finally, we concatenate all these differences together as the output of *TPD Modeling*.

The advantage of our *TPD Modeling* is that, compared with previous methods designed for image tasks, it provides multiple granularities of information and ensures the fine-grained representation of visual differences. In the ablation study of Boundary Captioning, we design an experiment to explore the difference modeling methods and verify our perceptions.

4.2   Boundary Captioning

For Boundary Captioning, we first implement and compare the performance of *CNN+LSTM*, *DUDA*, *UniVL-revised* and *ActBERT-revised*. To further explore how different input granularities support the understanding, we design a series of ablation studies using *ActBERT-revised* for all combinations of input features. In these two experiments, we apply our *TPD Modeling* as shown in Fig. 6.

To find the best schemes to represent visual difference, we further compare the performances of three schemes on *ActBERT-revised*: *(1)* Embedding with no difference modeling. *(2)* Max-pooling the frames before and after the boundary and simply subtracting one from another, which is inferred from the current method in [22]. *(3)* Using *TPD Modeling* to represent the visual differences. In Supp., we conduct an ablation study of different parts of *TPD Modeling* and explore on several other methods for visual difference representation.

Implementation. For *CNN+LSTM* and *DUDA*, we utilize the ResNet-conv feature following [22]. For *UniVL-revised* and *ActBERT-revised*, we utilize the ResNet-roi feature and TSN feature described in Sec. 4.1, where the sampling

Table 5. *Upper.* Ablation study results of the Boundary Captioning utilizing *ActBERT-revised* with *TPD Modeling* employed to all rows with "ResNet-roi". *Lower.* The performance comparison of visual difference modeling methods, where the *TPD Modeling* is employed to the last row

Input Granularity	CIDEr				SPICE				ROUGE_L			
	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.	Avg.	Sub.	Bef.	Aft.
ResNet-roi	51.93	67.79	46.59	41.42	14.30	16.01	13.54	13.34	24.20	35.42	19.04	18.13
ResNet-conv	66.18	**96.86**	54.77	46.91	17.07	20.58	15.82	14.8	26.30	40.38	19.71	18.82
TSN	70.80	92.54	65.64	54.21	19.00	**20.97**	18.98	17.04	26.89	**40.53**	20.82	19.32
ResNet-roi + ResNet-conv	56.64	83.82	45.64	40.45	15.68	19.17	13.77	14.1	25.46	38.64	19.26	18.47
ResNet-conv + TSN	69.58	83.56	68.88	56.3	18.95	20.15	19.51	17.2	27.14	38.52	22.36	20.53
ResNet-roi + TSN	**74.71**	85.33	**75.98**	**62.82**	**19.52**	20.10	**20.66**	**17.81**	**28.15**	39.16	**23.70**	**21.60**
ResNet-roi + ResNet-conv + TSN	65.83	80.9	63.22	53.38	18.69	19.37	19.25	17.46	26.84	37.82	22.11	20.59
ResNet-roi + TSN (w/o Diff.)	67.38	**85.59**	63.06	53.49	18.47	19.84	18.69	16.87	24.23	31.65	21.14	19.90
ResNet-roi + TSN (simple)	67.75	85.31	64.28	53.65	18.96	**20.35**	19.13	17.39	26.78	39.14	21.20	20.00
ResNet-roi + TSN (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**

range is from the preceding boundary to the succeeding boundary. In evaluation, we separate the prediction into the three items, and then compute the similarity score of each item with the ground truth. After that, we employ CIDEr [30], SPICE [2] and ROUGE_L [18] as evaluation metrics, which are widely utilized in image and video captioning benchmarks. Further details are discussed in Supp.

Result. From Tab. 4, we see that the *ActBERT-revised* backbone performs the best. However, the results in are still far from satisfactory, thus we further analyze the challenges of our task through the result in Tab. 5:

Accurate captioning of the status changes requires both the instant and event granularities. First, the event-granularity features perform as the base of the understanding. In Tab. 5, the "ResNet-roi+TSN" combination outperforms all the groups using only the instant-granularity features (e.g. the combinations of ResNet features). Second, a proper usage of the instant-granularity features could help to enrich the understanding. As in Tab. 5, the "ResNet-roi+TSN" combination outperforms the single TSN feature.

Our task requires adaptive usage of different granularities. Machines need to know when to look at which granularity. Simply assembling different features together could sometimes disturb the attention resulting in worse performance. In Tab. 5, when only utilizing the TSN feature, the performance is better than using either "ResNet-roi+TSN" or "ResNet-roi+ResNet-conv+TSN" combination.

Understanding the status changes requires effective modeling of visual differences In the comparison of difference modeling schemes in Tab. 5, the plain embedding without difference modeling performs the worst, while the utilization of simple-subtraction difference modeling brings little improvement to the performance. At the same time, the group with our *TPD Modeling* method significantly outperforms others. This gap in performance conforms to our perspective that learning a fine-grained understanding of status changes requires not only an overall but also a fine-grained representation of visual differences.

Table 6. Performance comparison among different methods in Boundary Grounding. For *UniVL-revised* and *ActBERT-revised*, we apply *TPD Modeling* and take the "ResNet-roi+TSN" combination as input feature

Method	Threshold (s)								
	0.1	0.2	0.5	1	1.5	2	2.5	3	Avg.
Random Guess	2.14	4.56	11.46	22.81	31.63	40.43	48.06	54.37	26.93
TVQA	2.60	5.30	12.90	23.73	32.94	41.33	48.56	55.17	27.82
2D-TAN	2.91	6.32	15.04	26.95	36.94	45.34	51.87	58.22	30.45
ActBERT-revised	3.12	6.14	14.79	26.78	36.61	45.45	52.99	59.41	30.66
FROZEN-revised	**4.28**	**8.54**	18.33	**31.04**	**40.48**	47.86	54.81	61.45	**33.35**
FROZEN-revised-GEBD	4.20	8.48	**18.49**	29.91	39.54	**48.37**	**55.29**	**61.55**	33.23

Table 7. Performance comparison of different methods in Boundary Caption-Video Retrieval. For *FROZEN-revised*, we add another group without difference modeling

Method	mAP	R@1	R@5	R@10	R@50
Random	0.39	0.05	0.23	0.44	2.52
CNN+LSTM	9.25	4.08	12.49	19.53	42.26
ActBERT-revised (one-stream)	19.14	9.52	28.89	40.14	64.50
FROZEN-revised (two-stream)	**23.39**	**12.80**	**34.81**	**45.66**	**68.10**
FROZEN-revised (two-stream) w/o diff	22.44	12.12	33.42	43.89	65.61

4.3   Boundary Grounding

In Boundary Grounding, we compare the performance of four backbones: *TVQA*, *2D-TAN*, *FROZEN-revised* and *ActionBERT-revised*. Given a video and a caption query, the model computes the matching scores of each candidate sampled from the video, followed with post-processing to finalize the prediction.

Implementation. In the training period, we use the ground truth boundaries processed in Sec. 3.4 and their timestamps. In testing, we employ two strategies to sample the timestamp candidates for groups as specified in their suffix. More details are discussed in Supp. For *2D-TAN*, we utilize the C3D feature as in [37]. For *TVQA*, we utilize the R-CNN and ResNet-roi features as context. Besides, we build the triplets consisting of one positive and two negative pairs, and then compute the cross-entropy loss for each triplet in training. In *ActBERT-revised* and *FROZEN-revised*, we apply the contrastive loss in [3] as objective and implemented a batch-randomed sequential sampler in training. The batch-randomed sampler allocates the boundaries in the same video to the same batch, encouraging the model to learn the visual differences within videos.

After the models generate the matching scores of all candidate timestamps, we apply the Laplace-of-Gaussian filter in [27] to derive local maximas of the scores. Then we select the top-k maximas as final prediction, where k is subject to the statistical number of ground truth timestamps marked in Sec. 3.4. To evaluate the accuracy of the prediction, we compute F1 scores based on the absolute distance between ground truth timestamps and predicted timestamps, with the threshold varying from 0.1s to 3s. Further details are discussed in Supp.

Result. We see that *FROZEN-revised* performs the best in the comparison of SOTA methods in Tab. 6. However, all the SOTA methods struggle when the

threshold is less than 1s, indicating that *improving the temporal resolution of understanding is still a main challenge of our task*. Future improvements still need to focus on how to delve deeper into the temporal details and prevent the models from taking a glance and learning a rough impression of status changes.

4.4   Boundary Caption-Video Retrieval

We implement and compare the performance of the *CNN+LSTM*, *FROZEN-revised* and *ActionBERT-revised* backbones. Same as in Boundary Grounding, the backbones is to compute the matching score between the query and context.

Implementation. In order to find the target video from the corpus, each query is to be tried to match with every boundary candidate from all videos. Considering the corpus size, we only apply the baseline in [27] to generate the boundary candidates. When implementing *CNN+LSTM*, we take the R-CNN and ResNet-roi features as visual contexts. For *FROZEN-revised* and *ActBERT-revised*, we utilize the same configuration with Boundary Grounding. To evaluate the retrieval accuracy, for each query, we sort all the videos by the highest scores of their boundary candidates and then compute the mAP and recall metrics.

Result. In Tab. 7, *FROZEN-revised* with difference modeling performs the best, but the performance gap is significantly smaller than in Boundary Grounding, suggesting that this video-level retrieval task relies less on the fine-grained visual differences. It is natural since the overall video-level understanding is already enough to distinguish the target among different videos.

5   Conclusion

In this paper, we have introduced our new dataset *Kinetic-GEB+* with the methods of benchmark construction and proposed three tasks that aim to develop a more fine-grained, robust and human-like understanding of videos based on status changes. We further explore the challenges with designed experiments, where we design a new *Temporal-based Pairwise Difference (TPD)* modeling method to represent visual differences and obtain significant improvement in performance. Concluding the results from the experiments, we summarize the challenges of our benchmarks as three parts: *(1)* How to adaptively utilize multiple granularities of features and exclude the disturbance. *(2)* How to effectively represent the visual differences around the boundary. *(3)* How to improve the temporal resolution of understanding. We believe our work could be a stepping stone for the following works to develop more powerful methods to understand status changes and thus improve video-level comprehension.

References

1. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Autonomous Robots **42**(7), 1301–1322 (2018)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European conference on computer vision. pp. 382–398. Springer (2016)
3. 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)
4. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1893–1902 (2015)
5. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multi-modal transformer for video grounding. In: EMNLP. pp. 9810–9823 (2021)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
7. Chen, D., Dolan, W.: Collecting highly parallel data for paraphrase evaluation. In: ACL. pp. 190–200 (Jun 2011), https://aclanthology.org/P11-1020
8. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)
9. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: ICCV. pp. 5267–5275 (2017)
10. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: WACV. pp. 245–253. IEEE (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Iashin, V., Rahtu, E.: Multi-modal dense video captioning. In: CVPR. pp. 958–959 (2020)
13. Jhamtani, H., Berg-Kirkpatrick, T.: Learning to describe differences between pairs of similar images. In: EMNLP (2018)
14. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
15. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV. pp. 706–715 (2017)
16. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018)
17. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: CVPR. pp. 7492–7500. IEEE Computer Society (2018)
18. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Liu, Z., Li, G., Mercier, G., He, Y., Pan, Q.: Change detection in heterogenous remote sensing images via homogeneous pixel transformation. IEEE Transactions on Image Processing **27**(4), 1822–1834 (2017)
20. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)

21. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: CVPR. pp. 10810–10819 (2020)
22. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. In: ICCV. pp. 4624–4633 (2019)
23. Portillo-Quintero, J.A., Ortiz-Bayliss, J.C., Terashima-Marín, H.: A straightforward framework for video retrieval using clip. In: Mexican Conference on Pattern Recognition. pp. 3–12. Springer (2021)
24. Radvansky, G.A., Zacks, J.M.: Event perception. Wiley Interdisciplinary Reviews: Cognitive Science **2**(6), 608–620 (2011)
25. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. TACL **1**, 25–36 (2013)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS **28** (2015)
27. 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)
28. Tian, J., Cui, S., Reinartz, P.: Building change detection based on satellite stereo imagery and digital surface models. IEEE Transactions on Geoscience and Remote Sensing **52**(1), 406–417 (2013)
29. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
30. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
31. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. PAMI **41**(11), 2740–2755 (2018)
32. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. In: ICCV. pp. 6847–6857 (2021)
33. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: ICCV. pp. 4581–4591 (2019)
34. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR. pp. 5288–5296 (2016)
35. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI. vol. 33, pp. 9159–9166 (2019)
36. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: CVPR. pp. 10287–10296 (2020)
37. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks formoment localization with natural language. In: AAAI (2020)
38. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI (2018)
39. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: CVPR. pp. 8746–8755 (2020)